

Folien zur Statistik

Einführung in die Statistik
für Pädagogen und Psychologen

WS 2002/2003
Fr 9.45 – 11.15
Fr 11.30 – 12.15 (Vertiefung)
Uni Bonn, Römerstraße 164, HS D

E.Jo. Zimmermann

Inhalt

1 Deskriptive Statistik	3
1.1 Skalenniveaus	3
1.2 Häufigkeitsverteilungen.....	9
1.2.1 Univariate Häufigkeitsverteilungen.....	10
1.2.2 Bivariate Häufigkeitsverteilungen	16
1.3 Statistische Kennwerte.....	19
1.3.1 Maße der zentralen Tendenz.....	20
1.3.2 Dispersionsmaße	25
1.4 Die z-Transformation.....	28
1.5 Zusammenhangsmaße.....	30
1.5.1 Kovarianz und Korrelation	31
1.5.2 Weitere bivariate Zusammenhangsmaße.....	34
1.6 Lineare Regression	35
1.6.1 Bestimmung der Geradengleichung	37
1.6.2 Beobachtete, vorhergesagte und Residualwerte	40
2 Wahrscheinlichkeit und Verteilungen.....	43
2.1 Wahrscheinlichkeit	43
2.2 Diskrete Verteilungen.....	49
2.3 Stetige Verteilungen	53
Literatur	58
Griechisches Alphabet.....	59

1 Deskriptive Statistik

In der deskriptiven Statistik geht es um die graphische und tabellarische Darstellung von Datenmengen sowie deren Beschreibung durch Kennwerte.

1.1 Skalenniveaus

- *Messen* ist eine *strukturerohaltende Zuordnung von Zahlen zu Objekten und Ereignissen*.
- Diese Zuordnung kann Aussagen unterschiedlicher „Güte“ erlauben, wobei die jeweiligen Gütestufen als Skalenniveaus bezeichnet werden.

Grundsätzlich wird unterschieden zwischen

- Nominalskalen
- Ordinalskalen
- Intervallskalen
- Ratio- oder Verhältnisskalen

Diese Skalenniveaus unterscheiden sich in ihrem Informationsgehalt. Dieser determiniert die erlaubten (mathematischen) Operationen und damit die statistischen Verfahren, die auf die jeweiligen Daten angewendet werden dürfen.

Nominalskala

Objekte mit gleicher Merkmalsausprägung erhalten gleiche Zahlen, Objekte mit verschiedener Merkmalsausprägung erhalten verschiedene Zahlen. Die einzelnen Ausprägungen des Merkmals und damit die Stufen der Skala sind ungeordnet (gleichberechtigt).

- Mögliche Aussagen: Gleichheit, Verschiedenheit.
- Erlaubte Operationen: $=, \neq$
- Zulässige Transformation: Eindeutige Transformationen, die zahlenmäßige Gleichheit und Verschiedenheit nicht verändern.
- Beispiele: Augenfarbe, Geschlecht

Ordinalskala

Eine Ordinalskala ordnet Objekten Zahlen zu, die so geartet sind, dass von jeweils zwei Objekten das Objekt mit der größeren Merkmalsausprägung die größere Zahl erhält. Über das Ausmass des Unterschiedes zwischen den einzelnen Ausprägungen ist nichts bekannt.

- Mögliche Aussagen: Größer-kleiner-Relationen.
- Erlaubte Operationen: $=, \neq, >, <$
- Zulässige Transformation: Monotone Transformationen, die Ordnung der Daten muss erhalten bleiben.
- Beispiele: Militärische Ränge, Windstärken.

Intervallskala

Wie bei der Ordinalskala, so sind auch bei der Intervallskala die Stufen geordnet. Zusätzlich sind hier die Abstände zwischen den einzelnen Stufen alle gleich gross. Hat z.B. ein Objekt den Skalenwert 1, so ist es von dem Objekt mit dem Skalenwert 2 genauso weit entfernt, wie dieses von einem Objekt mit dem Skalenwert 3. Der Nullpunkt einer Intervallskala ist relativ.

- Mögliche Aussagen: Abstandsvergleiche.
- Erlaubte Operationen: $=, \neq, >, <, +, -$
- Zulässige Transformation: Lineare Transformationen der Form $X' = a \cdot X + b$, die Verhältnisse zwischen Differenzen bleiben erhalten.
- Beispiele: Temperatur in Celsius oder Fahrenheit.

Verhältnisskala

Eine Verhältnisskala ist eine Intervallskala mit absolutem Nullpunkt, der das „Nichtvorhandensein“ des Merkmals ausdrückt.

- Mögliche Aussagen: Vergleich von Verhältnissen.
- Erlaubte Operationen: $=, \neq, >, <, +, -, \cdot, /$
- Zulässige Transformation: Ähnlichkeitstransformation der Form $X' = a \cdot X$, der Nullpunkt wird nicht verändert und die Verhältnisse von Werten bleiben erhalten.
- Beispiele: Größe in *m*, Gewicht in *kg*, Zeit in *sec*, Temperatur in Kelvin.

Weitere Bezeichnungen

Intervall- und Verhältnisskalen werden zusammenfassend auch als *Kardinalskalen* bezeichnet. Häufigkeiten besitzen Verhältnisskalenniveau und werden auch als *Absolutskala* bezeichnet.

Skalenniveaus, mögliche Operationen und statistische Kennwerte

	Nominal	Ordinal	Intervall	Verhältnis
Aussagen	Gleich- Ungleich	Gleich- Ungleich	Gleich- Ungleich	Gleich- Ungleich
		Größer- Kleiner	Größer- Kleiner	Größer- Kleiner
			Vergleich von Diffe- renzen	Vergleich von Diffe- renzen
				Vergleich von Verhält- nissen
Transfor- mationen	Eindeutige Transfor- mation	Monotone Transfor- mationen	Lineare Transforma- tion	Ähnlich- keitstrans- formation
Zentrale Tendenz	Modus	Modus, Median	Modus, Median, Mittelwert	Modus, Median, Mittelwert
Dispersion		Variations- breite	Variations- breite, AD- Streuung, Varianz	Variations- breite, AD- Streuung, Varianz

1.2 Häufigkeitsverteilungen

Urliste mit *Daten* (*Messwerte*) von $n = 92$ Teilnehmern der Lehrveranstaltung Statistik im WS 98/99. Die Liste enthält neben den Probanden-Nummern (nr) die *Variablen* „Geschlecht“ (sex), „Ängstlichkeit“ (angst) und „Körperlänge“ (lang).

nr	sex	angst	lang	nr	sex	angst	lang	nr	sex	angst	lang	nr	sex	angst	lang
1	2	4	173	24	1	3	160	47	1	2	163	70	1	5	169
2	1	4	165	25	1	4	174	48	1	2	175	71	1	5	175
3	1	4	169	26	1	5	174	49	2	5	189	72	1	6	168
4	1	5	174	27	1	4	158	50	1	4	170	73	1	3	169
5	1	3	172	28	1	2	180	51	2	3	170	74	1	4	164
6	1	4	173	29	1	5	180	52	1	4	167	75	1	5	169
7	1	2	176	30	1	3	167	53	1	4	169	76	2	1	176
8	1	2	160	31	1	1	177	54	1	2	170	77	2	5	185
9	1	3	171	32	1	2	167	55	1	2	170	78	1	3	169
10	1	3	170	33	1	2	165	56	2	3	181	79	1	4	175
11	1	2	176	34	1	5	184	57	1	3	162	80	1	4	166
12	1	5	174	35	1	3	168	58	1	2	162	81	1	2	159
13	1	3	170	36	1	2	171	59	1	5	171	82	1	3	173
14	1	2	170	37	1	3	165	60	1	4	164	83	2	2	184
15	2	2	175	38	1	3	170	61	1	4	168	84	1	2	168
16	2	5	195	39	1	4	170	62	1	4	168	85	1	4	175
17	1	5	165	40	1	5	168	63	2	6	180	86	1	3	176
18	1	2	165	41	1	6	178	64	1	7	167	87	1	2	167
19	1	1	166	42	2	2	175	65	2	1	175	88	1	3	165
20	1	2	160	43	1	4	173	66	1	4	180	89	1	4	176
21	1	4	164	44	1	4	170	67	1	5	169	90	1	4	171
22	1	4	177	45	1	2	176	68	1	5	165	91	1	5	173
23	1	5	165	46	2	2	172	69	1	5	152	92	1	4	171

1.2.1 Univariate Häufigkeitsverteilungen

- *Ordnung* der Daten nach der jeweiligen Variablen in aufsteigender Reihenfolge.
- Auszählung des Auftretens einzelner *Punktwerte* bzw. *Kategorien*. Folgende Häufigkeiten lassen sich bestimmen:

f	<i>Absolute Häufigkeit</i> mit der ein Wert auftritt.
h	<i>Relative Häufigkeit:</i> $\frac{f}{n}$.
$f_{\%}$	<i>Prozentuale Häufigkeit:</i> $\frac{f}{n} \cdot 100$.
f_{kum}	<i>Kumulierte absolute Häufigkeit:</i> Die bis zu dem betrachteten Wert einschließlich aufsummierten absoluten Häufigkeiten.
h_{kum}	<i>Kumulierte relative Häufigkeit:</i> Die bis zu dem betrachteten Wert einschließlich aufsummierten relativen Häufigkeiten.
$f_{\% \text{kum}}$	<i>Kumulierte prozentuale Häufigkeit:</i> Die bis zu dem betrachteten Wert einschließlich aufsummierten prozentualen Häufigkeiten. $\frac{f_{\text{kum}}}{n} \cdot 100$ wird auch als <i>Prozentrang</i> bezeichnet.

Aufsteigende Reihenfolge der „Ängstlichkeit“

nr	angst	nr	angst	nr	angst	nr	angst	nr	angst
19	1	48	2	57	3	53	4	34	5
31	1	54	2	73	3	60	4	40	5
65	1	55	2	78	3	61	4	49	5
76	1	58	2	82	3	62	4	59	5
7	2	81	2	86	3	66	4	67	5
8	2	83	2	88	3	74	4	68	5
11	2	84	2	1	4	79	4	69	5
14	2	87	2	2	4	80	4	70	5
15	2	5	3	3	4	85	4	71	5
18	2	9	3	6	4	89	4	75	5
20	2	10	3	21	4	90	4	77	5
28	2	13	3	22	4	92	4	91	5
32	2	24	3	25	4	4	5	41	6
33	2	30	3	27	4	12	5	63	6
36	2	35	3	39	4	16	5	72	6
42	2	37	3	43	4	17	5	64	7
45	2	38	3	44	4	23	5		
46	2	51	3	50	4	26	5		
47	2	56	3	52	4	29	5		

Primäre Häufigkeitsverteilung der „Ängstlichkeit“

Wert	f	h	$f_{\%}$	f_{kum}	h_{kum}	$f_{\% \text{kum}}$
1	4	0,043	4,3	4	0,043	4,3
2	23	0,250	25,0	27	0,293	29,3
3	17	0,185	18,5	44	0,478	47,8
4	25	0,272	27,2	69	0,750	75,0
5	19	0,207	20,7	88	0,957	95,7
6	3	0,033	3,3	91	0,989	98,9
7	1	0,011	1,1	92	1,000	100,0
Σ	$n = 92$	1	100,0			

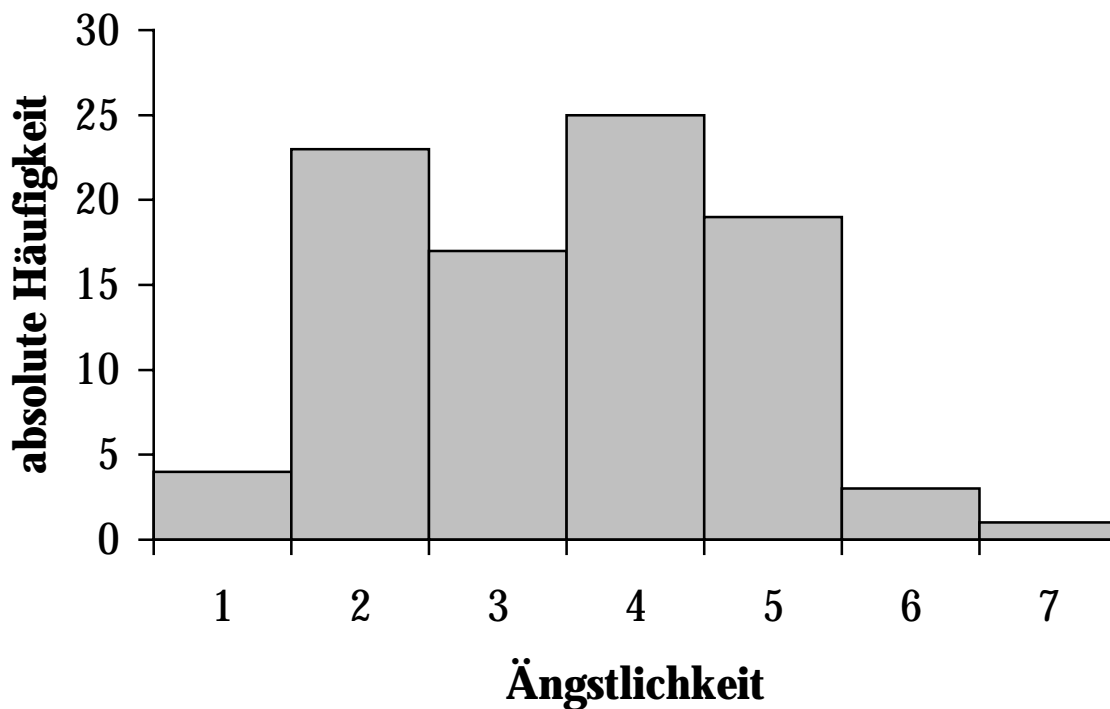


Abb. 1. *Histogramm* der Verteilung von „Ängstlichkeit“.

Primäre Häufigkeitsverteilung der „Körperlänge“

Wert	f	h	$f_{\%}$	f_{kum}	h_{kum}	$f_{\% \text{kum}}$
152	1	0,011	1,1	1	0,011	1,1
158	1	0,011	1,1	2	0,022	2,2
159	1	0,011	1,1	3	0,033	3,3
160	3	0,033	3,3	6	0,065	6,5
162	2	0,022	2,2	8	0,087	8,7
163	1	0,011	1,1	9	0,098	9,8
164	3	0,033	3,3	12	0,130	13,0
165	8	0,087	8,7	20	0,217	21,7
166	2	0,022	2,2	22	0,239	23,9
167	5	0,054	5,4	27	0,293	29,3
168	6	0,065	6,5	33	0,359	35,9
169	7	0,076	7,6	40	0,435	43,5
170	10	0,109	10,9	50	0,543	54,3
171	5	0,054	5,4	55	0,598	59,8
172	2	0,022	2,2	57	0,620	62,0
173	5	0,054	5,4	62	0,674	67,4
174	4	0,043	4,3	66	0,717	71,7
175	7	0,076	7,6	73	0,793	79,3
176	6	0,065	6,5	79	0,859	85,9
177	2	0,022	2,2	81	0,880	88,0
178	1	0,011	1,1	82	0,891	89,1
180	4	0,043	4,3	86	0,935	93,5
181	1	0,011	1,1	87	0,946	94,6
184	2	0,022	2,2	89	0,967	96,7
185	1	0,011	1,1	90	0,978	97,8
189	1	0,011	1,1	91	0,989	98,9
195	1	0,011	1,1	92	1,000	100,0
Σ	$n = 92$	1	100,0			

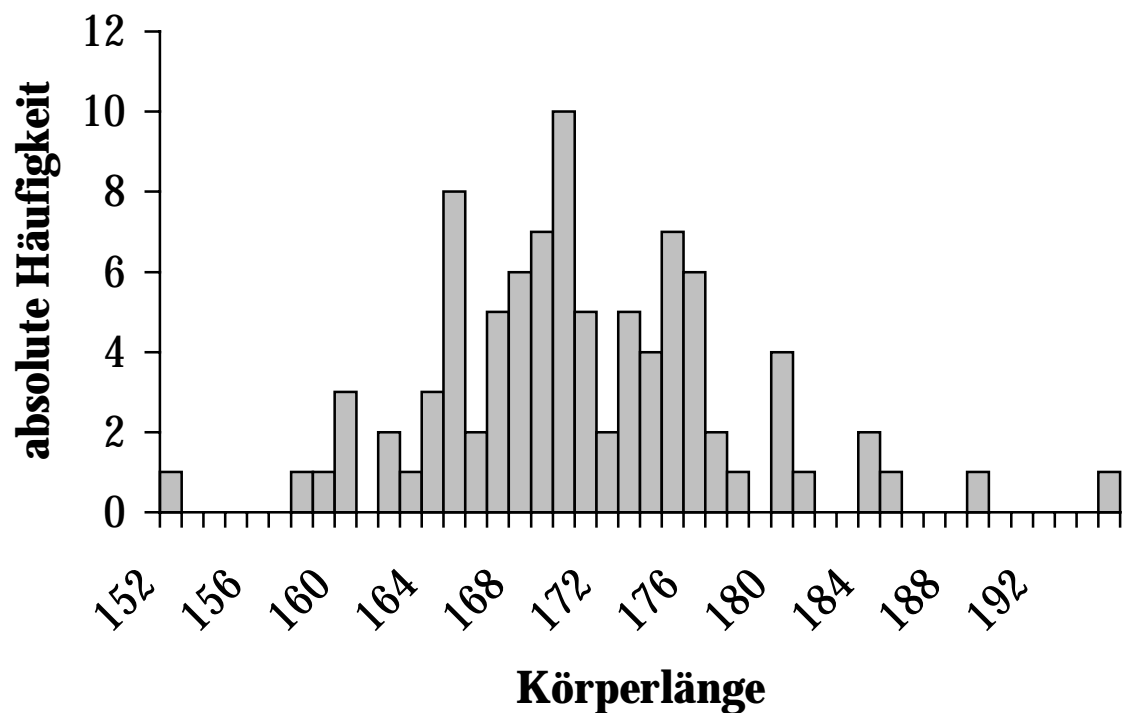


Abb. 2. Histogramm der „Körperlänge“.

Sekundäre Häufigkeitsverteilung der „Körperlänge“

Kategorie	f	h	$f_{\%}$	f_{kum}	h_{kum}	$f_{\% \text{kum}}$
150 – 154	1	0,011	1,1	1	0,011	1,1
155 – 159	2	0,022	2,2	3	0,033	3,3
160 – 164	9	0,098	9,8	12	0,130	13,0
165 – 169	28	0,304	30,4	40	0,435	43,5
170 – 174	26	0,283	28,3	66	0,717	71,7
175 – 179	16	0,174	17,4	82	0,891	89,1
180 – 184	7	0,076	7,6	89	0,967	96,7
185 – 189	2	0,022	2,2	91	0,989	98,9
190 – 194	0	0,000	0,0	91	0,989	98,9
195 – 199	1	0,011	1,1	92	1,000	100,0
Σ	$n = 92$	1	100,0			

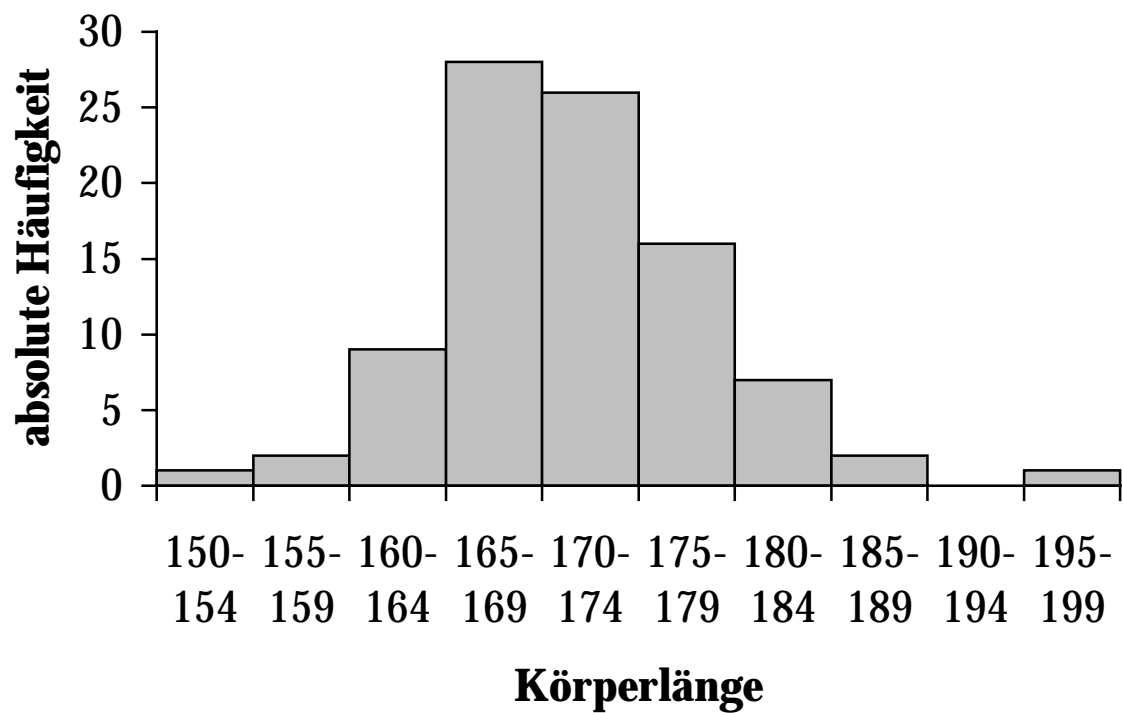


Abb. 3. Histogramm der kategorisierten „Körperlänge“.

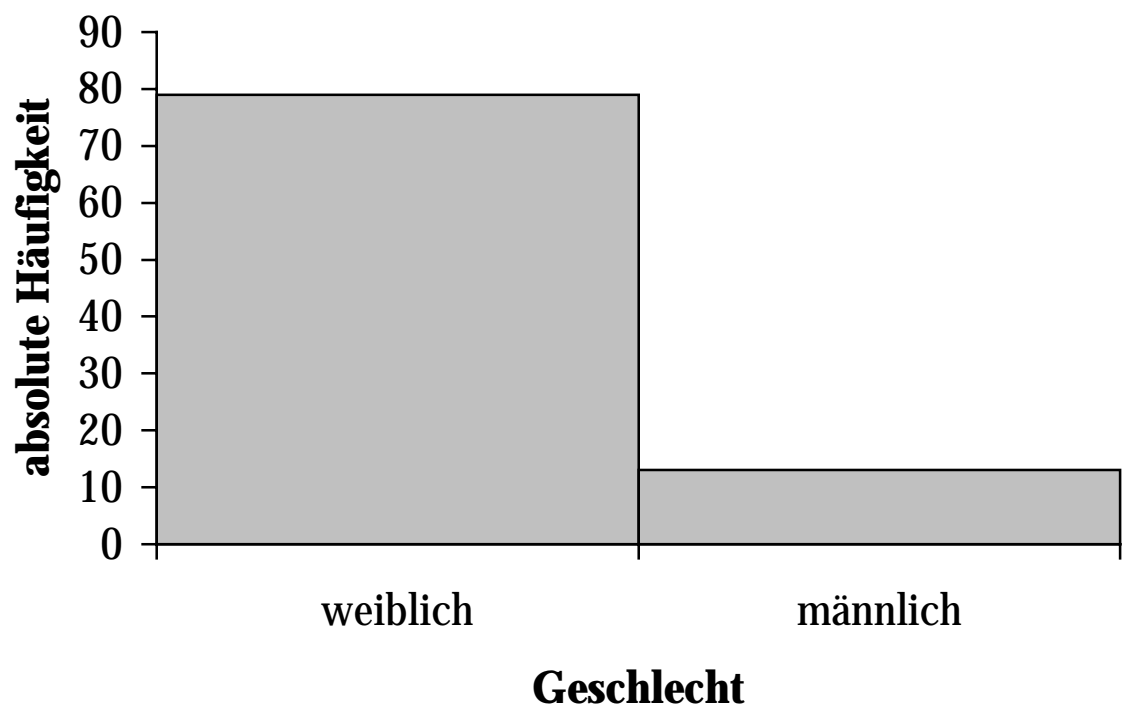


Abb. 4. Histogramm von „Geschlecht“.

1.2.2 Bivariate Häufigkeitsverteilungen

- *Ordnung* der Daten nach den jeweiligen Variablenpaaren in aufsteigender Reihenfolge.
- Auszählung des Auftretens einzelner Kombinationen von *Punktwerten* bzw. *Kategorien*.

Aufsteigende Reihenfolge von „Geschlecht“ und „Ängstlichkeit“

nr	sex	angst	nr	sex	angst	nr	sex	angst	nr	sex	angst	nr	sex	angst
19	1	1	58	1	2	78	1	3	60	1	4	40	1	5
31	1	1	81	1	2	82	1	3	61	1	4	59	1	5
65	2	1	84	1	2	86	1	3	62	1	4	67	1	5
76	2	1	87	1	2	88	1	3	66	1	4	68	1	5
7	1	2	15	2	2	51	2	3	74	1	4	69	1	5
8	1	2	42	2	2	56	2	3	79	1	4	70	1	5
11	1	2	46	2	2	2	1	4	80	1	4	71	1	5
14	1	2	83	2	2	3	1	4	85	1	4	75	1	5
18	1	2	5	1	3	6	1	4	89	1	4	91	1	5
20	1	2	9	1	3	21	1	4	90	1	4	16	2	5
28	1	2	10	1	3	22	1	4	92	1	4	49	2	5
32	1	2	13	1	3	25	1	4	1	2	4	77	2	5
33	1	2	24	1	3	27	1	4	4	1	5	41	1	6
36	1	2	30	1	3	39	1	4	12	1	5	72	1	6
45	1	2	35	1	3	43	1	4	17	1	5	63	2	6
47	1	2	37	1	3	44	1	4	23	1	5	64	1	7
48	1	2	38	1	3	50	1	4	26	1	5			
54	1	2	57	1	3	52	1	4	29	1	5			
55	1	2	73	1	3	53	1	4	34	1	5			

Primäre Häufigkeitsverteilung von „Geschlecht“
und „Ängstlichkeit“

sex	angst	f	h	$f_{\%}$
1	1	2	0,022	2,2
1	2	19	0,207	20,7
1	3	15	0,163	16,3
1	4	24	0,261	26,1
1	5	16	0,174	17,4
1	6	2	0,022	2,2
1	7	1	0,011	1,1
2	1	2	0,022	2,2
2	2	4	0,043	4,3
2	3	2	0,022	2,2
2	4	1	0,011	1,1
2	5	3	0,033	3,3
2	6	1	0,011	1,1
2	7	0	0,000	0,0
Σ		$n = 92$	1	100,0

Darstellung in einer Kreuztabelle oder Kontingenztafel:

angst	sex		Σ
	1	2	
1	2 2,2	2 2,2	4 4,3
2	19 20,7	4 4,3	23 25,0
3	15 16,3	2 2,2	17 18,5
4	24 26,1	1 1,1	25 27,2
5	16 17,4	3 3,3	19 20,7
6	2 2,2	1 1,1	3 3,3
7	1 1,1	0 0,0	1 1,1
Σ	79 85,8	13 14,2	92 100,0

1.3 Statistische Kennwerte

Statistische Kennwerte (etwas verwirrend häufig auch *Statistiken* genannt) haben die Funktion, in zusammengefaßter (aggregierter) Form Auskunft über Eigenschaften von Verteilungen zu geben.

Grundlegende statistische Kennwerte sind

- *Maße der zentralen Tendenz*, die die Lage (Lokation) einer Reihe von Meßwerten zusammenfassend repräsentieren und
- *Dispersionsmaße* (Streuungsmaße), die Auskunft über die Variabilität (Breite der Streuung um einen Zentralwert) einer Reihe von Meßwerten geben.

1.3.1 Maße der zentralen Tendenz

Durch welchen Wert wird eine Vielzahl von Werten am besten repräsentiert? – Hierzu geben Maße der zentralen Tendenz Auskunft über die *Lage* bzw. *Lokation* einer Verteilung in Termini der zugrundeliegenden Skala.

Modalwert (Modus)

Der *Modalwert* oder *Modus* ist der Wert, der in einer Verteilung am häufigsten vorkommt. Es kann auch mehr als einen Modus geben.

Beispiele:

$$Mo_{\text{sex}} = 1, Mo_{\text{angst}} = 4, Mo_{\text{lang}} = 170.$$

Median

Der *Median* ist der Wert, der eine der Größe nach geordnete Merkmalsverteilung in zwei gleichgroße Hälften teilt. Bei gerader Anzahl von Werten wird das arithmetische Mittel aus dem größten Wert der unteren Hälfte und dem kleinsten Wert der oberen Hälfte bestimmt.

Statistische Eigenschaften:

Die Summe der absoluten Abweichungen vom Median ist minimal.

Beispiele:

$Md_{\text{angst}} = \frac{x_{(46)} + x_{(47)}}{2} = \frac{4 + 4}{2} = 4$, wobei $x_{(a)}$ der a-te Wert in der aufsteigenden Reihenfolge ist;

$Md_{\text{lang}} = 170$.

Arithmetisches Mittel

Das *arithmetische Mittel* bzw. der *Mittelwert* ist die Summe aller Messwerte dividiert durch die Anzahl der Messwerte n :

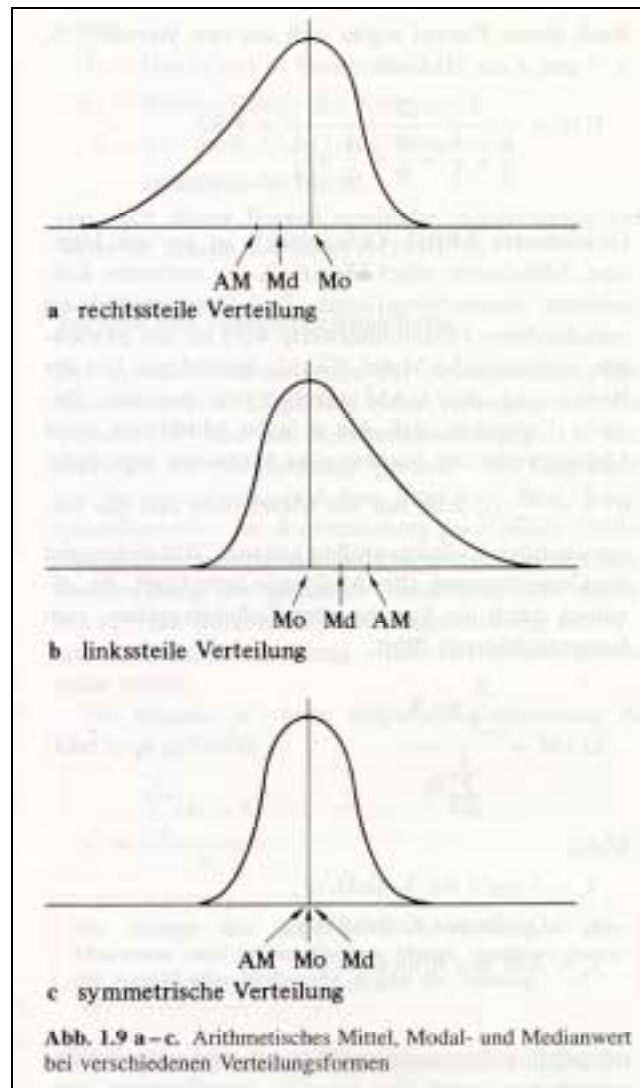
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

Statistische Eigenschaften:

- Die Summe der Abweichungen aller Messwerte vom Mittelwert ist Null.
- Die Summe der quadrierten Abweichungen der Messwerte vom Mittelwert ist ein Minimum.

Beispiel: $\bar{x}_{\text{lang}} = \frac{15717}{92} = 170,837.$

Zusammenhang zwischen Modus, Median und Mittelwert



Statistische Kennwerte der zentralen Tendenz und Skalenniveaus

	Nominal	Ordinal	Intervall	Verhältnis
Zentrale Tendenz	Modus	Modus, Median	Modus, Median, Mittelwert	Modus, Median, Mittelwert

Weitere Lokationsmaße

Geometrisches Mittel, harmonisches Mittel, gewichtetes Mittel, getrimmtes Mittel u.v.a.m..

Perzentile

Das *Perzentil* $P_{X\%}$ ist der Punkt auf einer Skala, unterhalb dessen $X\%$ der aufsteigend angeordneten Messwerte einer Stichprobe liegen. Der Median ist das 50%-Perzentil einer Verteilung.

1.3.2 Dispersionsmaße

Zwei Verteilungen können gleiche Maße der Zentralen Tendenz haben, dabei aber durchaus unterschiedlich breit streuen. *Dispersionsmaße* geben Auskunft über die Streuung bzw. Variabilität der Werte.

Variationsbreite

Die *Variationsbreite* (engl. *Range*) ist die Differenz zwischen dem größten (*Maximum*) und dem kleinsten Wert (*Minimum*).

Varianz

Die Varianz ist die Summe der quadrierten Abweichungen vom Mittelwert, dividiert durch die Anzahl der Messwerte n :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}.$$

Dieser Kennwert ist auf den ersten Blick nur schwer interpretierbar, hat aber in der Statistik einen hohen Stellenwert.

Standardabweichung

Die *Standardabweichung* oder *Streuung* ist die Wurzel aus der Varianz:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}.$$

Statistische Kennwerte der Dispersion und Skalenniveaus

	Nominal	Ordinal	Intervall	Verhältnis
Dispersion		Variationsbreite	Variationsbreite, Varianz	Variationsbreite, Varianz

Beispiel

Deskription von „Geschlecht“, „Ängstlichkeit“ und „Körperlänge“

	sex	Angst	lang
<i>n</i> gültig	92	92	92
fehlende Werte	0	0	0
Mittelwert	--	--	170,837
Median	--	4,000	170,000
Modus	1,000	4,000	170,000
Varianz	--	--	45,832
Standardabweichung	--	--	6,770
Minimum	--	1,000	152,000
Maximum	--	7,000	195,000
Variationsbreite	--	6,000	43,000
Perzentile 5%	--	2,000	160,000
25%	--	2,000	167,000
50%	--	4,000	170,000
75%	--	4,000	175,000
95%	--	6,000	184,000

1.4 Die z-Transformation

Die *z-Transformation* überführt jede Verteilung in eine Verteilung mit Mittelwert 0 und Streuung bzw. Varianz 1:

$$Z_i = \frac{x_i - \bar{x}}{s_x}.$$

Die *z-Transformation* ermöglicht den *relativen* Vergleich von Variablenausprägungen, da sie Mittelwerts- und Streuungsunterschiede „wegrelativiert“.

Beispiel

Ist eine Psychologiestudentin mit 184 cm Körperlänge *relativ* länger als ein Psychologiestudent derselben Länge? Relativ bedeutet in Bezug zur jeweiligen Geschlechtsgruppe.

	Körperlänge	
	Frauen	Männer
N	79	13
Mittelwert	169,456	179,231
Streuung	5,632	7,040

$$z_{\text{Frauen}} = \frac{184 - 169,456}{5,632} = 2,582$$

$$z_{\text{Männer}} = \frac{184 - 179,231}{7,040} = 0,677$$

Während eine Frau von 184 cm Körperlänge um 2,582 Einheiten (Standardabweichungen) über dem Mittelwert liegt, liegt ein Mann derselben Länge um nur 0,677 Einheiten (Standardabweichungen) über dem Mittelwert, jeweils in Bezug zur Geschlechtsgruppe. Die Frau besitzt also eine „überdurchschnittlichere“ Länge.

1.5 Zusammenhangsmaße

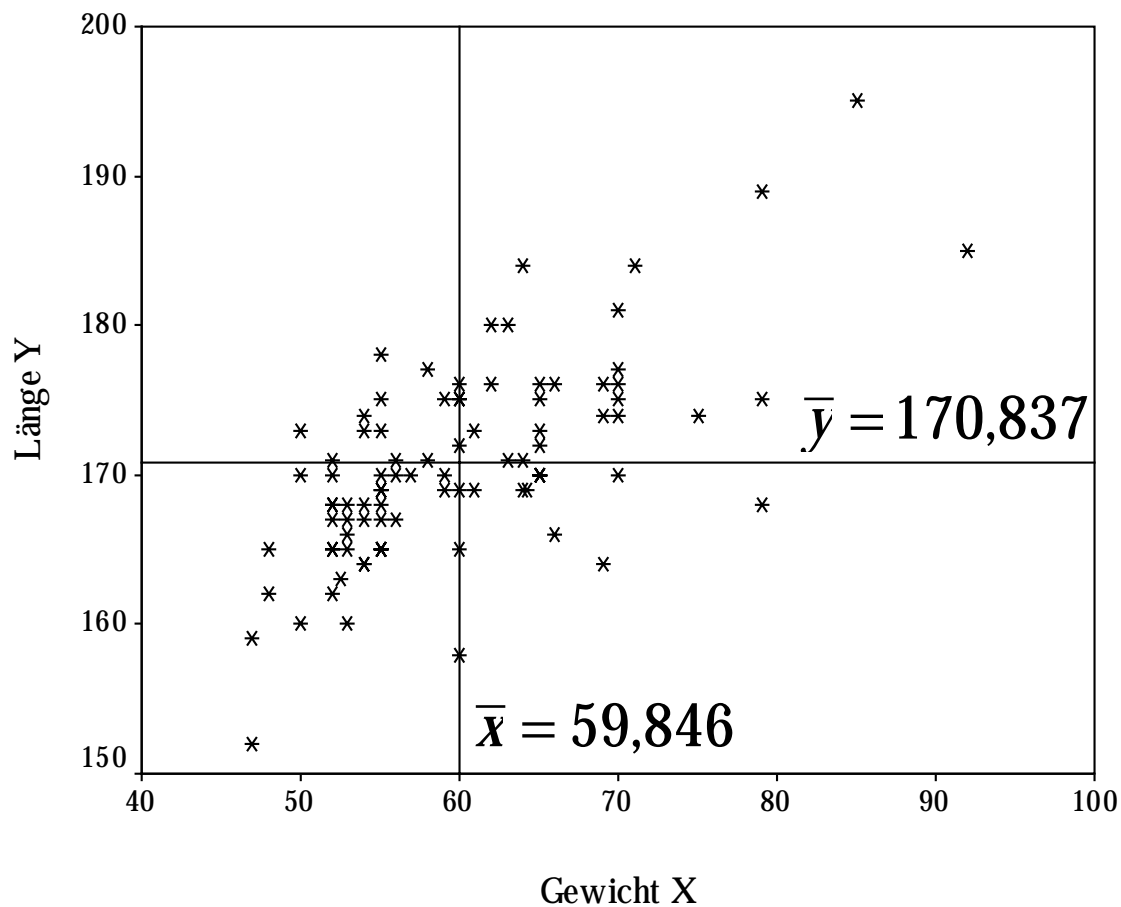


Abb. 5. *Streudiagramm (Scatterplot) der Variablen „Körperlänge und -gewicht“ ($n = 92$).*

1.5.1 Kovarianz und Korrelation

Mit der *Kovarianz* läßt sich bestimmen, wie sich die relativen Positionen von gepaarten Messwerten aus zwei Variablen zueinander verhalten:

$$\text{COV}_{(x,y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Die Kovarianz kann nur hinsichtlich ihres Vorzeichens, nicht aber hinsichtlich ihrer Größe interpretiert werden.

- Ein positives Vorzeichen der Kovarianz zeigt an, daß sowohl überdurchschnittliche Werte von X stärker mit überdurchschnittlichen Werten von Y gepaart sind als auch, daß unterdurchschnittliche Werte von X stärker mit unterdurchschnittlichen Werten von Y zusammen auftreten.
- Ein negatives Vorzeichen zeigt entsprechend an, daß überdurchschnittliche Werte von X stärker mit unterdurchschnittlichen Werten von Y gepaart sind und umgekehrt.

Die Kovarianz kann maximal den Wert $s_x \cdot s_y$ annehmen. Dies ist damit ein sehr schöner Term zur Relativierung bzw. Standardisierung von Kovarianzen.

Diese Standardisierung nennt man *Produkt-Moment-Korrelation* $r_{(x,y)}$.

$$r_{(x,y)} = \frac{COV_{(x,y)}}{s_x \cdot s_y}.$$

Die Korrelation $r_{x,y}$ kann Werte zwischen -1 und $+1$ annehmen, wobei

- -1 einen perfekt negativen linearen Zusammenhang anzeigt,
- $+1$ einen perfekt positiven linearen Zusammenhang anzeigt und
- 0 keinen *linearen* Zusammenhang anzeigt.

Perfekter linearer Zusammenhang: Die Werte der einen Variablen lassen sich durch eine Lineartransformation in die der anderen überführen:

$$y_i = a + b \cdot x_i$$

Beispiel:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
62	176	-2,2	2,0	-4,4
53	166	-11,2	-8,0	89,6
70	177	5,8	3,0	17,4
70	175	5,8	1,0	5,8
66	176	1,8	2,0	3,6
Σ 321	870	0	0	112

$$\bar{x} = 64,2, s_x = \sqrt{40,16} = 6,337$$

$$\bar{y} = 174,0, s_y = \sqrt{16,4} = 4,050$$

$$\text{COV}_{(x,y)} = \frac{112}{5} = 22,4$$

$$r_{(x,y)} = \frac{\text{COV}_{(x,y)}}{s_x \cdot s_y} = \frac{22,4}{6,337 \cdot 4,050} = 0,873$$

Korrelation und Kausalität

Die Kovariation zweier Variablen kann auf eine kausale Beziehung der beiden Variablen oder auf die Beeinflussung durch eine oder mehrere *Drittvariablen* zurückgehen. Auf Basis einer Korrelation kann nicht auf die Art der Verursachung geschlossen werden. (Z.B. korrelieren bei Kindern Intelligenz und Schuhgröße.)

1.5.2 Weitere bivariate Zusammenhangsmaße

	dichotom	ordinal	dichotom- isiert	intervall
dichotom	Phi	Rang- biserial	—	Punkt- biserial
ordinal		Rangkor- relation (Tau, Rho)	—	—
dichotom- isiert			Tetracho- risch	Biserial
intervall				Produkt- Moment

1.6 Lineare Regression

Hängen zwei Variablen (wie z.B. Körperlänge und Gewicht) zusammen, so ist es möglich, eine Variable jeweils auf Basis der anderen vorherzusagen.

Die Vorhersagevariable wird als *unabhängige Variable* oder *Prädiktorvariable* (X) bezeichnet, die Variable, die vorhergesagt werden soll, als *abhängige Variable* oder *Kriteriumsvariable* (Y).

Im Rahmen der bivariaten Regression wird die Vorhersage über folgende lineare Beziehung vorgenommen:

$$\hat{y}_i = a + b \cdot x_i, \text{ wobei}$$

\hat{y}_i = vorhergesagter Wert, a = Schnittpunkt der Geraden mit der y -Achse (Ordinatenabschnitt, Intercept) und b = Steigung der Geraden.

Ziel ist die Bestimmung einer Geraden, die den Gesamttrend aller Punkte am besten wiedergibt.

Beispiel:

Für die Körper der 92 Erstsemester des WS 1998 ermittelt man folgende Gleichung:

$$\text{Länge} = 139,74 + 0,52 \cdot \text{Gewicht}.$$

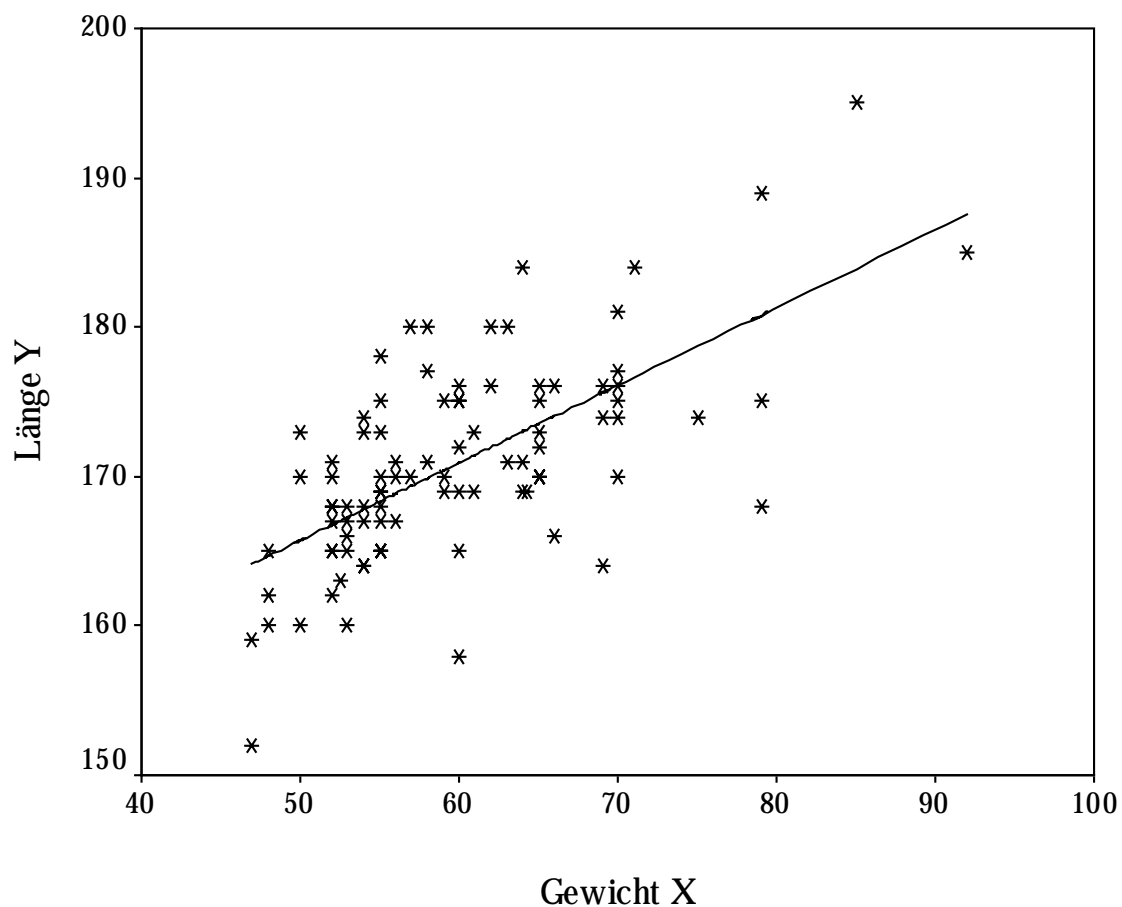


Abb. 6. Lineare Regression von Länge auf Gewicht ($n = 92$).

1.6.1 Bestimmung der Geradengleichung

Bei der linearen Regression wird die Gleichung

$$\hat{y}_i = a + b \cdot x_i$$

gesucht, für die die Summe der quadrierten Abweichungen zwischen vorhergesagten und tatsächlichen Werten minimal ist (*Kriterium der kleinsten Quadrate*):

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 =$$
$$\sum_{i=1}^n [y_i - (a + b \cdot x_i)]^2 = \min.$$

Durch partielle Ableitung dieser Funktion nach a bzw. b lassen sich allgemeine Lösungen sowohl für a als auch für b finden, die obiges Kriterium erfüllen:

$$b_{yx} = \frac{\text{COV}_{(x,y)}}{s_x^2} \quad \text{und} \quad a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}.$$

Bei b_{yx} und a_{yx} bezeichnet der erste Index (hier y) die Variable, die vorhergesagt wird, und der zweite Index die Variable, die vorhersagt (hier x).

Beispiel:

Körpergewicht X (Prädiktor)	Körperlänge Y (Kriterium)
62	176
53	166
70	177
70	175
66	176

$$\bar{x} = 64,2, s_x = \sqrt{40,16} = 6,337, \bar{y} = 174,0$$

$$\text{COV}_{(x,y)} = \frac{112}{5} = 22,4$$

$$b_{yx} = \frac{\text{COV}_{(x,y)}}{s_x^2} = \frac{22,4}{40,16} = 0,558$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x} = 174 - 0,558 \cdot 64,2 = 138,176$$

Die Regressionsgleichung lautet also:

$$\hat{y}_i = 138,176 + 0,558 \cdot x_i$$

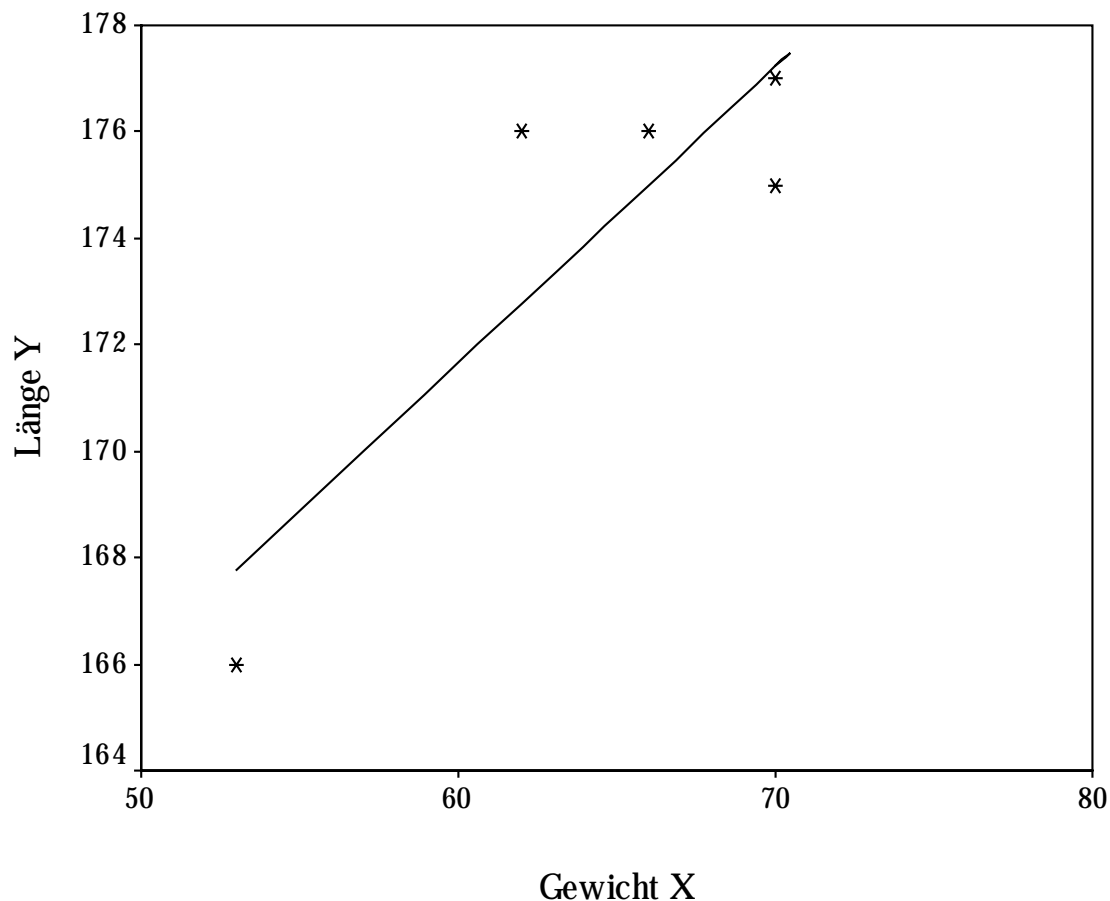


Abb. 7. Lineare Regression von Länge auf Gewicht.

Für eine Person, die 70 kg wiegt, sagen wir gemäß der linearen Regression folgende Körperlänge voraus:

$$\begin{aligned}\hat{y} &= 138,176 + 0,558 \cdot 70 \\ &= 177,236.\end{aligned}$$

1.6.2 Beobachtete, vorhergesagte und Residualwerte

Die Abweichungen der beobachteten Werte von den vorhergesagten Werten heißen *Regressionsresiduen*:

$$y_i^* = y_i - \hat{y}_i.$$

Die Residuen enthalten die Anteile der Kriteriumsvariablen, die durch die Prädiktorvariable nicht erfaßt werden.

Mittelwerte

- Die Mittelwerte der Kriteriumsvariablen und der vorhergesagten Werte sind gleich:

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \hat{y}_i.$$

- Der Mittelwert der Residuen ist 0:

$$\frac{1}{n} \sum_{i=1}^n y_i^* = 0.$$

Im Beispiel:

	x_i	y_i	\hat{y}_i	y_i^*
	62	176	172,772	3,228
	53	166	167,750	-1,750
	70	177	177,236	-0,236
	70	175	177,236	-2,236
	66	176	175,004	0,996
Σ	321	870	870	0
AM	64,2	174	174	0

Zerlegung der Kriteriumsvarianz

Die Varianz der y -Werte setzt sich additiv aus der Varianz der vorhergesagten \hat{y} -Werte und der Varianz der Residuen y^* zusammen:

$$s_y^2 = s_{\hat{y}}^2 + s_{y^*}^2 .$$

Im Beispiel: $16,4 = 12,5 + 3,9$.

Determinationskoeffizient

Varianzanteil der abhängigen Variablen, der mit der unabhängigen vorhergesagt bzw. erklärt werden kann:

$$r_{(y, \hat{y})}^2 = \frac{s_{\hat{y}}^2}{s_y^2} .$$

Im Beispiel: $12,5/16,4 = 0,762$.

$0,762 \cdot 100\% = 76,2\%$ der Varianz der „Körperlänge“ kann durch das „Körpergewicht“ erklärt werden (und umgekehrt).

Korrelationen

- Die Korrelation zwischen Kriterium und Vorhersage ist gleich der Korrelation zwischen Kriterium und Prädiktor:

$$I_{(y, \hat{y})} = I_{(y, x)}.$$

- Die Korrelation zwischen Prädiktor und Residuen ist 0:

$$r_{(x, y^*)} = 0.$$

- Für die Korrelation zwischen Kriterium und Residuen gilt:

$$r_{(y, y^*)}^2 = 1 - r_{(x, y)}^2.$$

Im Beispiel:

	x	y	\hat{y}	y^*
x	1	0,873	1,000	0,000
y		1	0,873	0,488
\hat{y}			1	0,000
y^*				1

$$0,488^2 = 0,238 = 1 - 0,762 = 1 - 0,873^2$$

2 Wahrscheinlichkeit und Verteilungen

2.1 Wahrscheinlichkeit

Ein *Zufallsexperiment* ist ein beliebig oft wiederholbarer wohldefinierter Vorgang, dessen Ausgang nicht vorhersagbar ist.

Ein Ausgang wird als *Ergebnis* oder *Elementarereignis* ω , die Menge aller *Elementarereignisse* wird als *Ergebnisraum* $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ bezeichnet. Die Ergebnisse eines Raumes sind vollständig und schließen sich gegenseitig aus.

Bei der *theoretischen (kombinatorischen) Bestimmung* einer *Wahrscheinlichkeit* geht man oft von der Gleichwahrscheinlichkeit aller Ergebnisse aus:

$$p(A) = \frac{\text{Anzahl günstiger Ergebnisse}}{\text{Anzahl möglicher Ergebnisse}}.$$

Günstige Ergebnisse sind solche, bei denen sich A ereignet, mögliche Ergebnisse sind alle Ergebnisse des endlichen Ergebnisraumes (A kann durchaus mehr als ein einziges ω_i umfassen).

Beispiel 1:

Zufallsexperiment „Münzwurf“: Die möglichen Elementarereignisse sind „Kopf“ (K) und „Zahl“ (Z): $\Omega = \{K, Z\}$. Unter Annahme einer „fairen“ Münze (Gleichwahrscheinlichkeit) ist $p(K) = p(Z) = \frac{1}{2}$.

Beispiel 2:

Zufallsexperiment „Reißzweckenwurf“: Der Ergebnisraum ist auch hier $\Omega = \{\omega_1, \omega_2\} = \{\text{„Spitze oben“}, \text{„Spitze schräg unten“}\}$, aber die Annahme der Gleichwahrscheinlichkeit ist unplausibel. In solchen Fällen können die $p(\omega_i)$ meist nur über viele Versuche geschätzt werden.

Ein *statistischer Schätzer* der Wahrscheinlichkeit eines Ereignisses A ist die relative Häufigkeit eben dieses Ereignisses:

$$h(A) = \frac{f(A)}{n}.$$

Begriffe:

- Der *Durchschnitt* $A \cap B$ bedeutet, daß sich A und B ereignen.
- Die *Vereinigung* $A \cup B$ bedeutet, daß sich A und/oder B ereignen.
- *Addition*: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.
- *Disjunkt* bzw. *unvereinbar* sind Ereignisse, die keinen Durchschnitt besitzen; dann gilt: $p(A \cup B) = p(A) + p(B)$.
- *Multiplikation*: $p(A \cap B) = p(B) \cdot p(A|B)$.

Bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit* $p(A|B)$ ist die Wahrscheinlichkeit von A , sofern B eingetreten ist:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}.$$

Ein Schätzer für $p(A|B)$ ist die *bedingte, relative Häufigkeit*:

$$h(A|B) = \frac{h(A \cap B)}{h(B)} = \frac{f(A \cap B)/n}{f(B)/n} = \frac{f(A \cap B)}{f(B)}.$$

Beispiel:

In einer Zufallsstichprobe mit $n = 1000$ sind:

- Frauen: 500.
- Raucherinnen: 100.
- Raucher: 150.

	R („raucht“)	\bar{R}	f
F (Frau)	$f(F \cap R) = 100$	$f(F \cap \bar{R}) = 400$	500
\bar{F} (Mann)	$f(\bar{F} \cap R) = 150$	$f(\bar{F} \cap \bar{R}) = 350$	500
f	250	750	1000

Der Anteil der Frauen an der gesamten Stichprobe beträgt: $h(F) = 500/1000 = 0,5$.

Wie groß ist der Anteil der Frauen unter den Rauchern und Raucherinnen?

$$h(F/R) = \frac{f(F \cap R)}{f(R)} = \frac{100}{250} = 0,4.$$

Oder: Wie groß ist der Anteil der Raucher und Raucherinnen unter den Frauen?

$$h(R/F) = \frac{f(F \cap R)}{f(F)} = \frac{100}{500} = 0,2.$$

$h(F) \neq h(F | R)$ deutet daraufhin, daß die Ereignisse F und R *nicht unabhängig* voneinander auftreten.

Stochastische Unabhängigkeit

Zwei Ereignisse A und B sind *stochastisch unabhängig*, wenn die Wahrscheinlichkeit für das gemeinsame Auftreten der Ereignisse A und B dem Produkt ihrer Einzelwahrscheinlichkeiten entspricht: $p(A \cap B) = p(A) \cdot p(B)$.

Äquivalent hierzu ist $p(A) = p(A|B)$, sofern $p(B) > 0$, und $p(B) = p(B|A)$, sofern $p(A) > 0$.

Wenn A unabhängig von B , dann ändert das Eintreten von B nichts an der Eintretenswahrscheinlichkeit von A , und umgekehrt.

Beispiel für Unabhängigkeit:

	R	\bar{R}	p
F	0,1	0,4	0,5
\bar{F}	0,1	0,4	0,5
p	0,2	0,8	1

Bei Unabhängigkeit müßte gelten:

$$p(R) \cdot p(F) = p(R \cap F).$$

Das ist hier der Fall:

$$0,2 \cdot 0,5 = 0,1.$$

Dann gilt auch:

$$p(R|F) = p(R) \Leftrightarrow 0,1 / 0,5 = 0,2.$$

Beispiel für Abhängigkeit:

	R	\bar{R}	p
F	0,05	0,45	0,5
\bar{F}	0,15	0,35	0,5
p	0,2	0,8	1

Bei Unabhängigkeit müßte gelten:

$$p(R) \cdot p(F) = p(R \cap F),$$

aber:

$$0,2 \cdot 0,5 \neq 0,05.$$

Dann ist auch:

$$p(R|F) \neq p(R), \text{ hier: } 0,05 / 0,5 \neq 0,2.$$

2.2 Diskrete Verteilungen

Diskrete Wahrscheinlichkeitsfunktionen ordnen einer endlichen Anzahl von Ereignissen Auftretenswahrscheinlichkeiten zu.

Zufallsexperiment „Wurf von 2 Münzen“:

Mögliche Ergebnisse: $\Omega = \{ZZ, ZK, KZ, KK\}$. Wenn die Münzen fair und unabhängig voneinander sind, tritt jedes Ergebnis mit $p = \frac{1}{4} = 0,5 \cdot 0,5 = 0,25$ auf. Bei den möglichen Ergebnissen wird die Abfolge der Ereignisse (mit welcher Münze was geworfen wird) berücksichtigt. Man kann aber auch fragen, wie wahrscheinlich ist es, $X \times$ Kopf zu werfen:

$$p(\text{„}0 \times \text{Kopf“}) = p(\{ZZ\}) = 0,25.$$

$$p(\text{„}1 \times \text{Kopf“}) = p(\{ZK, KZ\}) = 0,5.$$

$$p(\text{„}2 \times \text{Kopf“}) = p(\{KK\}) = 0,25.$$

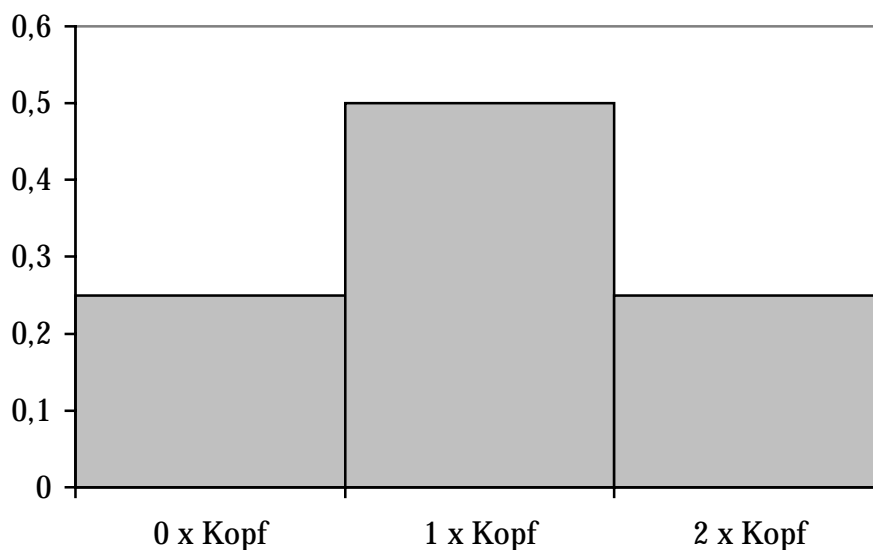


Abb. 8. Diskrete Wahrscheinlichkeitsfunktion.

Zufallsexperiment „Wurf von 3 Münzen“:

$\Omega = \{ZZZ, ZZK, ZKZ, KZZ, ZKK, KZK, KKZ, KKK\}$,
 $p(\omega_i) = 1/8 = 0,5^3 = 0,125$.

$p(\text{„}0 \times \text{Kopf}\text{“}) = p(\{ZZZ\}) = 0,125$.

$p(\text{„}1 \times \text{Kopf}\text{“}) = p(\{ZZK, ZKZ, KZZ\}) = 0,375$.

$p(\text{„}2 \times \text{Kopf}\text{“}) = p(\{ZKK, KZK, KKZ\}) = 0,375$.

$p(\text{„}3 \times \text{Kopf}\text{“}) = p(\{KKK\}) = 0,125$.

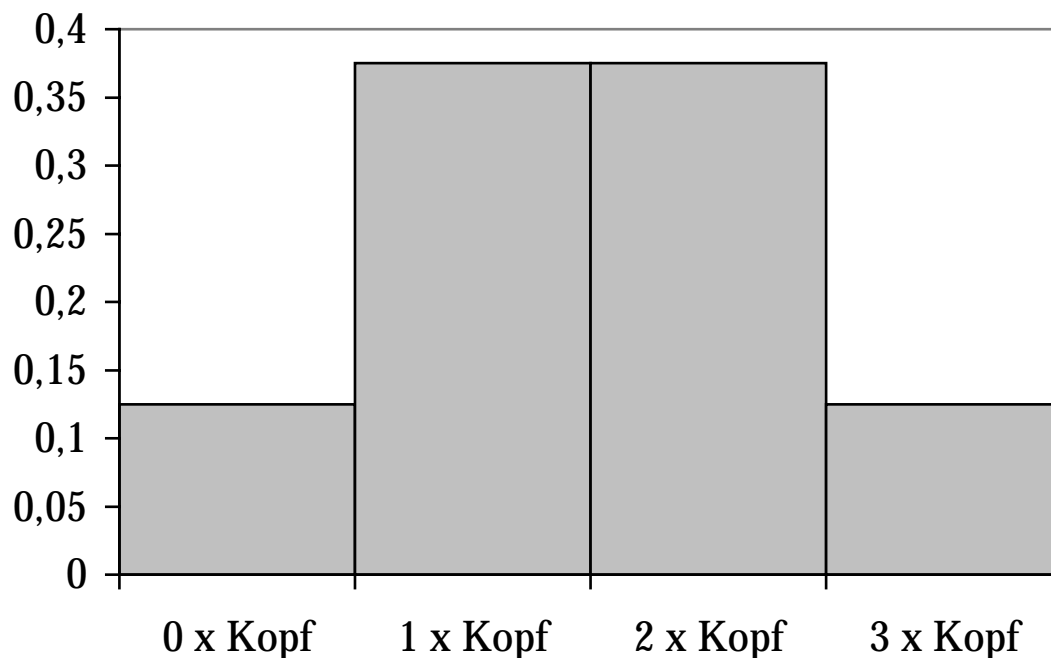


Abb. 9. Diskrete Wahrscheinlichkeitsfunktion.

Binomialverteilung

Ein Experiment, welches in der n -fachen Wiederholung von unabhängigen Einzelexperimenten mit nur zwei möglichen Ausgängen besteht, heißt *Bernoulli-Experiment*. Wenn ein Ereignis X in einem Einzelexperiment mit einer Wahrscheinlichkeit p auftritt, dann kann die Wahrscheinlichkeit P , daß X in n Wiederholungen k -mal auftritt, allgemein bestimmt werden:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}.$$

Diese *Wahrscheinlichkeitsfunktion* heißt *Binomialverteilung* und gibt an, wie wahrscheinlich das Eintreten eines Ereignisses $X = k$ ist.

Die *Verteilungsfunktion* hingegen gibt an, mit welcher Wahrscheinlichkeit ein Ereignis höchstens k -mal auftritt:

$$P(X \leq k) = \sum_{j=0}^k \binom{n}{j} \cdot p^j \cdot (1 - p)^{n-j}.$$

Die Verteilungsfunktion ist die kumulierte Wahrscheinlichkeitsfunktion.

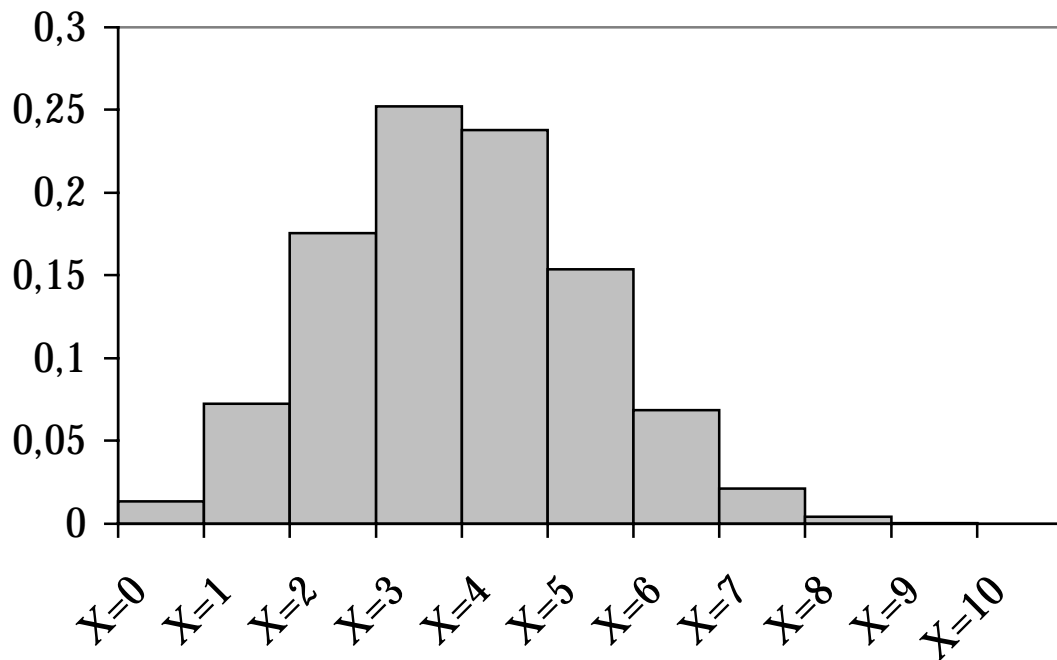


Abb. 10. Wahrscheinlichkeitsverteilung mit $p = 0,35$, $n = 10$.

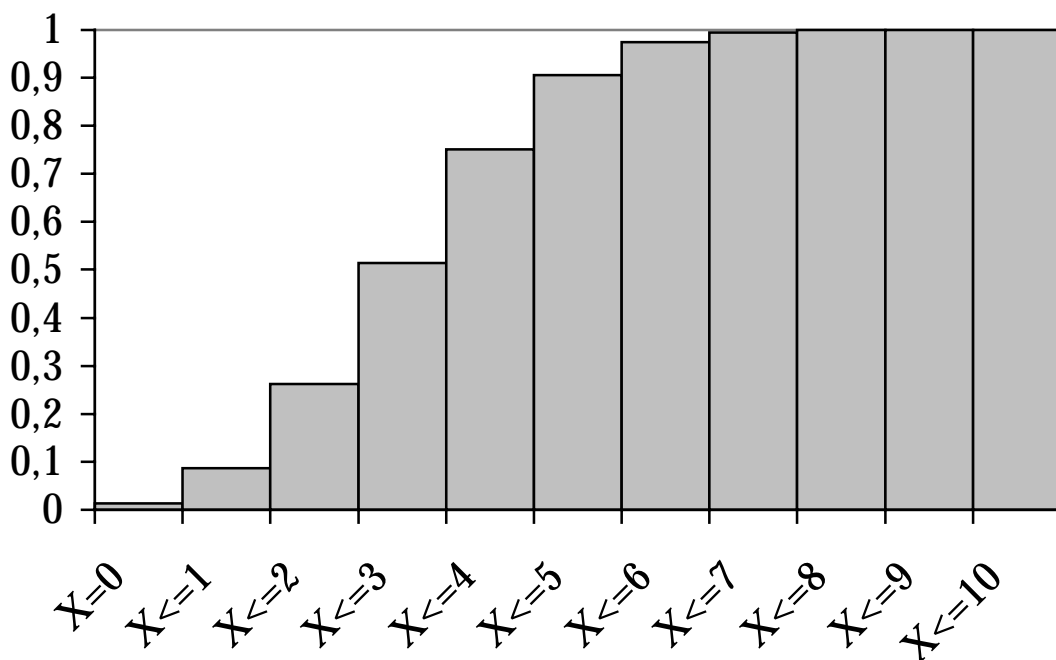


Abb. 11. Verteilungsfunktion für $p = 0,35$, $n = 10$.

Andere diskrete Verteilungen sind die *hypergeometrische*, die *Poisson*- und die *multinomiale Verteilung*.

2.3 Stetige Verteilungen

In einem Zufallsexperiment mit einer stetigen Variable besteht die Ergebnismenge aus unendlich vielen möglichen Ergebnissen. Die Wahrscheinlichkeit einzelner Ergebnisse läßt sich daher nicht bestimmen.

Stetige Verteilungen werden nicht mit Balken, sondern mit Kurven symbolisiert.

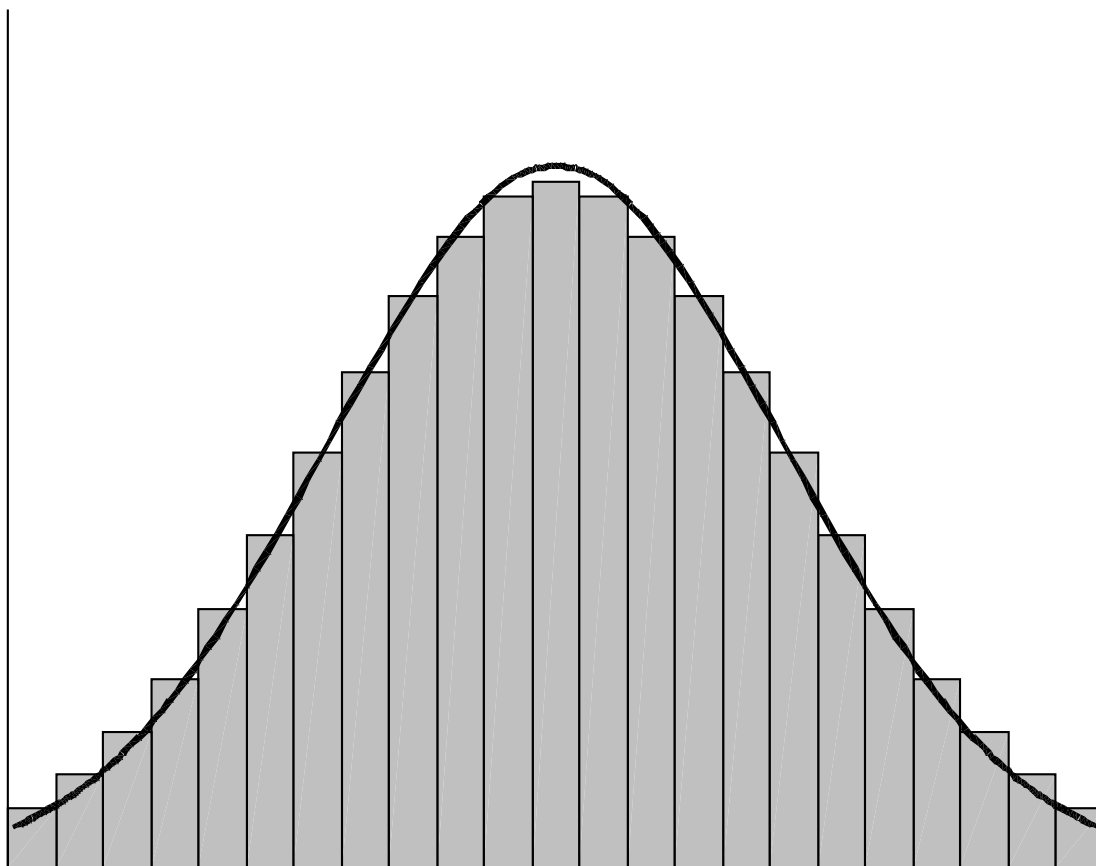


Abb. 12. Binomial- vs. Normalverteilung.

Die Gesamtfläche unter der Kurve ist auf 1 normiert, was bedeutet, daß mit einer Wahrscheinlichkeit 1 irgendein Ergebnis eintritt. Die Fläche über einem Intervall von zwei Werten gibt an, mit welcher Wahrscheinlichkeit Werte innerhalb dieses Intervalls eintreten.

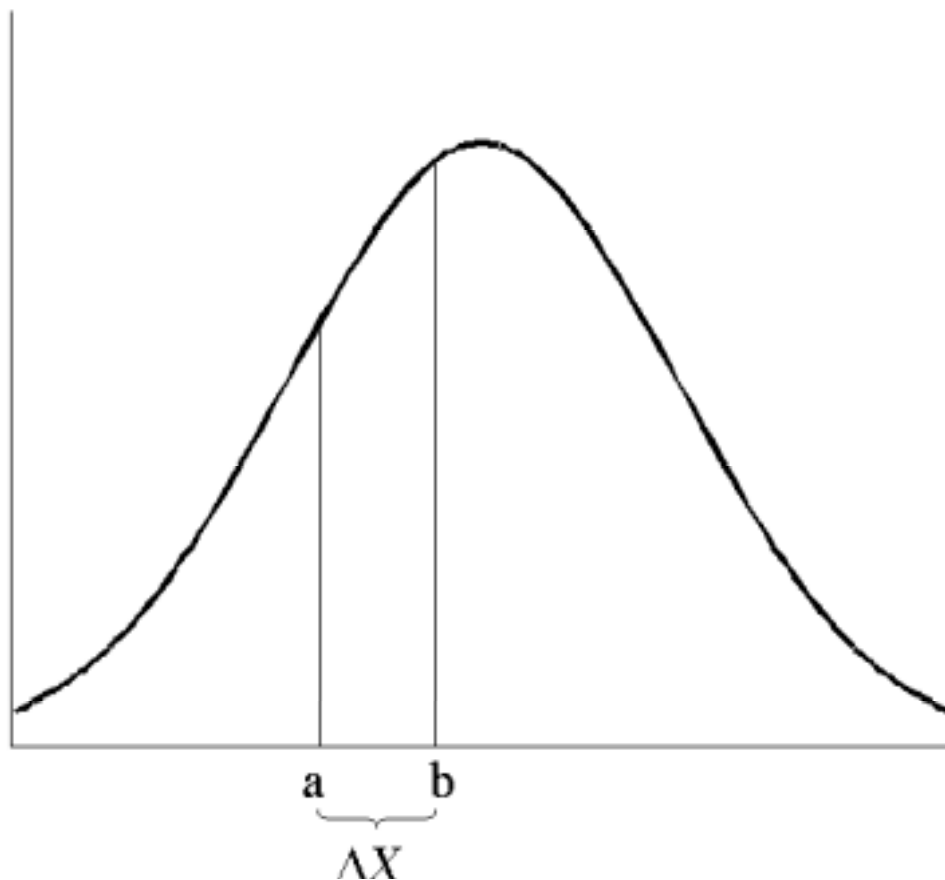


Abb. 13. Wahrscheinlichkeit eines Intervalls ΔX .

Die Wahrscheinlichkeitsfunktion einer stetigen Verteilung wird als *Wahrscheinlichkeitsdichte* oder *Dichtefunktion* bezeichnet. Erst über Integrale dieser Dichtefunktion können Wertebereichen Wahrscheinlichkeiten zugeordnet werden.

Die Dichtefunktion der **Normalverteilung** lautet:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-(x-\mu)^2 / 2\sigma^2}.$$

Diese Funktion beschreibt eine Schar von Verteilungen, die sich hinsichtlich μ und σ unterscheiden.

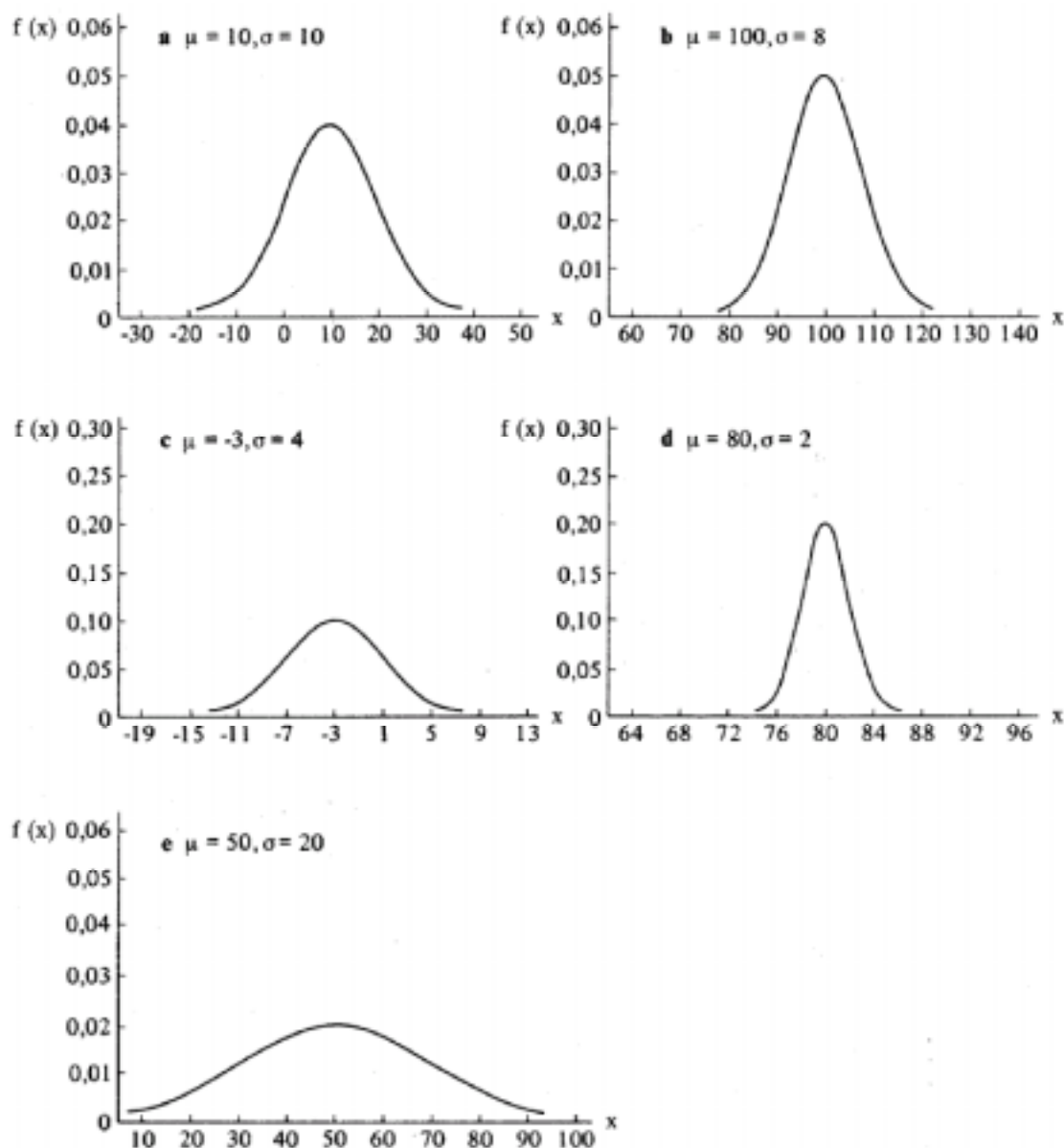


Abb. 14. Verschiedene Normalverteilungen.

Eigenschaften der Normalverteilung:

- Glockenförmiger Verlauf.
- Symmetrie, d.h. Modalwert, Median und Mittelwert fallen zusammen.
- Die Verteilung nähert sich asymptotisch der x -Achse.
- Zwischen den zu den Wendepunkten gehörenden x -Werten befinden sich ungefähr $2/3$ der Gesamtfläche.

Standardnormalverteilung

Durch eine z -Transformation kann jede Normalverteilung in die Standardnormalverteilung mit $\mu = 0$ und $\sigma = 1$ [Abk.: $N(0, 1)$] überführt werden:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2}.$$

Diese Dichte (siehe „Ordinate“) und deren Verteilungsfunktion (siehe „Fläche“) ist bei Bortz (1999) in Tabelle B (S. 768-772) wiedergegeben.

Beispiele

- Wie groß ist die Fläche von $-\infty$ bis $-1,96$?
 $p(z \leq -1,96) = F(z = -1,96) = 0,025$.
- Wie groß ist die Fläche zwischen -1 und 1 ?
 $p(-1 \leq z \leq 1) = F(z = 1) - F(z = -1) = 0,8413 - 0,1587 = 0,6826$.
D.h. bei Normalverteilungen liegen zwischen $-\sigma$ und $+\sigma$ immer 68,26% aller Werte.
- Wie groß ist die Fläche von $1,65$ bis $+\infty$?
 $p(z \geq 1,65) = 1 - F(z = 1,65) = F(z = -1,65) = 0,0495$.

Bedeutsamkeit der Normalverteilung

Die Normalverteilung nimmt eine zentrale Stellung unter den theoretischen Verteilungen ein, da sich verschiedene Verteilungen von ihr herleiten (χ^2 , t , F) und andere gegen sie konvergieren (Binomial, Poisson).

Die Normalverteilung dient als theoretisches Modell für die Verteilung von ...

- Stichprobenkennwerten,
- Zufallsfehlern,
- (manchen) empirischen Merkmalen.

Literatur

Grundlage dieser Veranstaltung:

Bortz, J. (1999). Statistik für Sozialwissenschaftler. Berlin: Springer.

Weitere Einführungen:

Diehl, J.M. & Kohr, H.U. (1991). Deskriptive Statistik. Frankfurt a.M.: Klotz.

Diehl, J.M. & Arbinger, R. (1989). Einführung in die Inferenzstatistik. Frankfurt a.M.: Klotz.

Hays, W.L. (1988). Statistics. Orlando: Holt, Rinehart & Winston.

Krengel, U. (1988). Einführung in die Wahrscheinlichkeitstheorie und Statistik. Wiesbaden: Vieweg.

Griechisches Alphabet

A	α	Alpha	N	ν	Ny
B	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	Ο	ο	Omikron
Δ	δ	Delta	Π	π	Pi
E	ε	Epsilon	Ρ	ρ	Rho
Z	ζ	Zeta	Σ	σ	Sigma
H	η	Eta	Τ	τ	Tau
Θ	θ	Theta	Υ	υ	Ypsilon
I	ι	Jota	Φ	φ	Phi
K	κ	Kappa	Χ	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	My	Ω	ω	Omega

Dieses Skript richtet sich nach den Anforderungen, die die Abteilung Methodenlehre & Diagnostik für das Vordiplom stellt. Das Skript erhebt keinen Anspruch auf Vollständigkeit und Richtigkeit, ist aber sicher eine gute Hilfe zum Verständnis der Thematiken sowie zur Prüfungsvorbereitung.

Bortz: Statistik

SKRIPTLISTE:			
Grundstudium:		erstellt	Hauptstudium:
		erstellt	
B – Methodenlehre (+ Bortz)	2002		
C – Allgemeine Psychologie 1	2001		
D – Wahrnehmung & Denken	2002		
KEINE weiteren Skripte geplant		weitere Skripte in Vorbereitung	

Viel Spaß!

Manuel Ulrich

Zur Überprüfung: 151 Seiten

Erstellt wurde das Skript Januar bis April 2002.

Inhaltsverzeichnis:

Diverses: S. 5

(Konzeption des Skriptes; Literaturliste; griechisches Alphabet; Warnhinweis)

Statistik (B2):

Deskriptive Statistik: S. 8

(Skalenniveaus; Maße der zentralen Tendenz: Modus, Median & Arithmetisches Mittel; Dispersionsmaße: Variationsbreite und Perzentile, Quadratsummen, Varianz und Standardabweichung; z-Werte)

Merkmalszusammenhänge & -vorhersagen: S. 12

(Kovarianz und Korrelation; Korrelation und Kausalität; Lineare Regression; Beobachtete, vorhergesagte und Residualwerte; Determinationskoeffizient)

Wahrscheinlichkeitstheorie und -verteilungen: S. 16

(Zufallsexperimente; Grundbegriffe der Wahrscheinlichkeitsrechnung; Axiome der Wahrscheinlichkeitstheorie; Das Theorem von Bayes; Variationen, Permutationen, Kombinationen; Verteilungsfunktionen; Diskrete Verteilungen; Stetige Verteilungen 1: Normalverteilung & Standardnormalverteilung; Stetige Verteilungen 2: χ^2 -Verteilung, t-Verteilung sowie F-Verteilung; Zusammenhänge von NV, STNV, χ^2 -, t- und F-Verteilung)

Stichprobe und Grundgesamtheit: S. 27

(Begriffe der Inferenzstatistik; Stichprobenarten; Verteilung von Stichprobenkennwerten; Kriterien der Parameterschätzung: Erwartungstreue, Konsistenz, Effizienz, Suffizienz; Methoden der Parameterschätzung: Methode der kleinsten Quadrate, Maximum-likelihood-Methode; Intervallschätzung: Wertebereich, Konfidenzintervalle; Bedeutung des Stichprobenumfangs)

Formulierung und Überprüfung von Hypothesen: S. 35

(Alternativhypothesen; Varianten von Alternativhypothesen; Die Nullhypothese; Fehlerarten bei statistischen Entscheidungen: Alpha- & Beta-Fehler; Signifikanzaussagen; Bestimmung der Irrtumswahrscheinlichkeit; Einseitige und zweiseitige Tests; Statistische Signifikanz und praktische Bedeutsamkeit; Effektgröße; Der Beta-Fehler; Indifferenzbereiche; Teststärke („power“); Bedeutung der Stichprobengröße; Monte-Carlo-Studien, Bootstrap-Technik)

Verfahren zur Überprüfung von Unterschiedshypothesen: S.44

(Verfahren für Intervalldaten; z-Test, t-Test; t-Test für unabhängige Stichproben; t-Test für abhängige Stichproben; χ^2 -Test für eine Varianz; F-Test; Verfahren für Ordinaldaten; Randomisationstest nach Fisher; U-Test; Wilcoxon-Test; Verfahren für Nominaldaten; Vergleich der Häufigkeit eines zweifach gestuften Merkmals; Vergleich der Häufigkeit eines k-fach gestuften Merkmals; 4-Felder- χ^2 -Test; k x l- χ^2 -Test; Konfigurationsanalyse; Allgemeine Bemerkungen zu den χ^2 -Techniken)

Verfahren zur Überprüfung von Zusammenhangshypothesen: S.65

(Die Statistische Absicherung von linearen Regressionen; Determinanten der Vorhersagegenauigkeit; Nonlineare Regression; Interpretationshilfen für r; k-fach gestufte Merkmale; Korrelation für nonlineare Zusammenhänge; Fishers Z-Transformation; Spezielle Korrelationstechniken)

Varianzanalyse (B3):

Einfaktorielle Versuchspläne: S. 75

(Begriffe der Varianzanalyse; Grundprinzip der einfaktoriellen Varianzanalyse; Quadratsummenzerlegung; Varianzschätzungen; Entscheidung; Effektgrößen; Varianzaufklärung; Voraussetzungen; Einzelvergleiche; Alpha-Fehler-Korrektur; Scheffé -Test; Trendtests; Zusammenhänge; Rangvarianzanalyse)

Mehrfaktorielle Versuchspläne: S. 88

(Zweifaktorielle Varianzanalyse; Quadratsummen; Freiheitsgrade und Varianzschätzungen; Hypothesenprüfung; Varianzaufklärung; Interaktionsdiagramme; Feste und zufällige Faktoren; Effektgrößen und Einzelvergleiche; Drei- und mehrfaktorielle Varianzanalysen; Ungleiche Stichprobengrößen; Voraussetzungen)

Versuchspläne mit Messwiederholungen: S. 97

(Einfaktorielle Varianzanalyse mit Messwiederholung; Datenschema; Hypothesen; Quadratsummen; Varianzschätzungen; Prüfung der Nullhypothese; Trendtests und Einzelvergleiche; Mehrfaktorielle Varianzanalyse mit Messwiederholung; Mathematik total; Voraussetzungen)

Kovarianzanalyse: S. 103

(Einfaktorielle Kovarianzanalyse; Datenschema; Hypothesen; Quadratsummen; Freiheitsgrade und Varianzschätzungen; Hypothesenprüfung; Unterschiedliche Stichprobengrößen, A-priori Einzelvergleiche; Effektgrößen; Zusammenfassung; Voraussetzungen; Mehrfaktorielle Kovarianzanalyse; Kovarianzanalyse mit Messwiederholung)

Unvollständige, mehrfaktorielle Versuchspläne: S. 109

(Hierarchische und teilhierarchische Versuchspläne; Lateinische Quadrate; Griechisch-lateinische Quadrate; Quadratische Anordnungen mit Messwiederholungen)

Theoretische Grundlagen der Varianzanalyse: S. 111

(Einfaktorielle Varianzanalyse; Mehrfaktorielle Varianzanalyse; Varianzanalyse mit Messwiederholung; Einfaktorielle Kovarianzanalyse; Weitere Verfahren; Allgemeine Regeln)

Multivariate Verfahren (B4):

Partialkorrelation und Multiple Korrelation: S. 114

(Partialkorrelation; Berechnung der Partialkorrelation; Semipartialkorrelation; Berechnung der Semipartialkorrelation; Partialkorrelationen höherer Ordnung; Signifikanztests; Multiple Korrelation und Regression; Voraussetzungen und Signifikanztests; Schrumpfungskorrektur; Beta-Koeffizienten; Strukturkoeffizienten; Effektgröße; Lineare Strukturgleichungsmodelle; Zeit & Korrelation)

Das Allgemeine Lineare Modell (ALM): S. 127

(Grundprinzip des Allgemeinen Linearen Modells; Dummycodierung; Effektcodierung; Kontrastcodierung; Zusammenfassung)

Faktorenanalyse: S. 130

(Erste Schritte zur Faktorenanalyse; Grundprinzip der Faktorenanalyse; Zusammenfassung; Die „richtige“ Faktorenanalyse; Kennwerte der Faktorenanalyse: Eigenwert, Faktorladung & Kommunalität; Die Kennwerte Eigenwert, Faktorladung & Kommunalität am empirischen Beispiel; Kriterien für relevante Faktoren; Interpretation der Faktoren; Rotationsverfahren: Orthogonal vs. Oblique; Weiteres)

Clusteranalyse: S. 141

(Grundprinzip der Clusteranalyse; Grundprinzip hierarchischer clusteranalytischer Methoden; Fusionskriterien; Nichthierarchische clusteranalytische Verfahren; Die k-means-Methode; Probleme bei der Clusteranalyse; Zusammenfassung)

Multivariate Mittelwertvergleiche: S. 146

(Mehrfach univariate Analysen vs. eine multivariate Analyse)

Diskriminanzanalyse: S. 148

(Grundprinzip der Diskriminanzanalyse)

Kanonische Korrelationsanalyse: S. 153

(Grundprinzip der kanonischen Korrelationsanalyse; Schlussbemerkung)

© Manuel Ulrich

Gegen eine Vervielfältigung dieses Skriptes für private Zwecke bestehen keine Einwände. Die kommerzielle Nutzung ist untersagt.

Zusammenfassung der gebrauchten Abkürzungen:

NV	=	Normalverteilung
STNV	=	Standardnormalverteilung
SKV	=	Stichprobenkennwerteverteilung
H_0	=	Nullhypothese
H_1	=	Alternativhypothese
KI	=	Konfidenzintervall
df	=	Freiheitsgrade
r	=	Korrelation
Abszisse	=	x-Achse
AV	=	abhängige Variable
UV	=	unabhängige Variable
KV	=	Kontrollvariable
QS	=	Quadratsumme
ANOVA	=	Varianzanalyse
ANCOVA	=	Kovarianzanalyse
MANOVA	=	Multivariate Varianzanalyse
FA	=	Faktoranalyse
Vp(n)	=	Versuchsperson(en)
VT	=	Vortest
NT	=	Nachtest
QUEX	=	Quasi-Experiment
Reliabilität	=	Zuverlässigkeit der Daten
Validität	=	Gültigkeit inhaltlicher Schlussfolgerungen
Axiom	=	nicht beweisbare Grundannahme
DK	=	Determinationskoeffizient (der %-Anteil der Varianz, der vorhergesagt wird)
EG	=	Experimentalgruppe
KG	=	Kontrollgruppe
SP	=	Stichprobe
IA	=	Interaktion

Konzeption des Skriptes

Das Skript folgt beim Mathematikteil (B2, B3 & B4) den Kapiteln des Bortz (sofern nicht anders angegeben), ergänzend dazu habe ich die Folien von B2, B3 & B4 nachgearbeitet.

Ein Eingang auf SPSS – Ausdrücke (wäre wohl eher B5) erfolgt im B4 Teil.

Da ich im Mathematikteil auf den Bortz als Grundlage, sowie zur Absicherung auf die offiziellen Folien und ein Skript aus der Fachschaft zurückgreifen konnte, meine ich, dass ich alle relevanten Informationen (eher zuviel) für B2, B3 & B4 in diesen Skript stehen habe.

Den Methodenlehreteil (B1) habe ich in einem gesonderten Skript stehen, da die Anforderungen & Themen pro Uni unterschiedlich sind. In der Psychologiefachschaft der Uni Bonn ist aber der B1-Teil zu finden (oder schreibt mir eine eMail mit der Bitte um Sendung des B1-Teils).

Das Erstellen eines eigenen Lernskriptes – was den eigentlichen Lernerfolg bringt, da ihr die Sachen nicht nur verstehen, sondern auch erklären müsst – ersetzt das Lesen dieses Skriptes mit Sicherheit nicht.

Benutzte Literatur für das Skript:

Statistik (B2), Varianzanalyse (B3) & Multivariate Verfahren (B4):

- Bortz (1999): Statistik für Sozialwissenschaftler (5. Auflage) [Unsere Bibel]
- Die Folien der Lehrveranstaltungen B2 (WS 01/02), B3 (SS 01) & B4 (WS 01/02)
- Das Statistik-Skript von Benjamin Zeller (B2 & B3)

Griechisches Alphabet:

A	a	Alpha	N	n	Ny
B	b	Beta	X	x	Xi
G	g	Gamma	O	o	Omikron
D	d	Delta	P	p	Pi
E	e	Epsilon	R	r	Rho
Z	z	Zeta	S	s	Sigma
H	h	Eta	T	t	Tau
Q	ϑ	Theta	U	u	Ypsilon
I	i	Jota	F	j	Phi
K	k	Kappa	C	c	Chi
L	l	Lambda	Y	y	Psi
M	m	My	W	w	Omega

ACHTUNG!

Bevor es losgeht:

In diesem Skript stehen viele mit dem Formeleditor von Windows erstellte Formeln sowie einige Formeln aus dem B2 & B3-Skript. Es kann sein, dass die Formeln sich verändern, da der Formeleditor von Windows nicht bei allen PCs die Formeln genau gleich darstellt (dies kann beim „Anklicken“ der Formeln geschehen).

Deswegen ist das Skript schreibgeschützt. Bitte also nur Änderungen speichern, wenn Ihr – zu eurem eigenen Interesse – eine anders benannte Sicherheitskopie erstellt habt.

Wenn gar nichts klappt, kopiert Euch mein Skript ausgedruckt aus der Fachschaft.

Zur Kontrolle: es folgen eine **fehlerfreie** Kovarianzformel und ein z-Test (für eine Population).

$$\text{COV}_{(x,y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$z_{\text{emp}} = \frac{\bar{x} - m}{s_{\bar{x}}}$$

Statistik

Deskriptive Statistik:

Deskriptive Statistik = beschreibende Statistik, d.h. sie dient zur Zusammenfassung und Darstellung von Daten (z.B. durch schöne Grafiken, Kennwerte wie Mittelwert, Streuung, etc.).

Skalenniveaus:

Terminologie: um Menschen empirisch untersuchen zu können, müssen wir ihnen Zahlen zuordnen können (nach bestimmten, vorgegebenen Regeln), so dass wir auf dieser Basis Berechnungen anstellen können:

Die z.B. Größe von Herrn Schmitz (ein **empirisches Relativ**) lässt sich durch Zuordnung von Zahlen in einem **numerischen Relativ** (z.B. 1,85 m) ausdrücken.

Wenn ein empirisches Relativ nur durch bestimmte Zahlen ausgedrückt werden kann und nicht auch durch andere, dann ist es eindeutig bzw., anders ausgedrückt: **homomorph**.

D.h. auf das Beispiel bezogen: die Größe von Herrn Schmitz beträgt 1,85 m; oder mathematisch formuliert:

$$\text{Herr Schmitz} = \text{Größe in Metern (1,85)}$$

Die oben zu sehende Gleichung ist eine homomorphe Abbildungsfunktion zusammen mit einem empirischen und einem numerischen Relativ. Dies bezeichnet man auch als **SKALA**.

Die Funktionswerte (nur der Teil der Gleichung rechts vom „=“) werden auch als **Skalenwerte** oder **Messwerte** bezeichnet.

**Messen ist also eine strukturerhaltende Zuordnung von Zahlen
zu Objekten und Ereignissen.**

Diese Zuordnung kann Aussagen unterschiedlicher „Güte“ erlauben, wobei die jeweiligen Gütestufen als Skalenniveaus bezeichnet werden:

Grundsätzlich wird unterschieden zwischen

- **Nominalskalen**
- **Ordinalskalen**
- **Intervallskalen**
- **Ratio- oder Verhältnisskalen**

Eselsbrücke: Wenn Ihr bei der Reihenfolge schwarz (französisch: **„noir“**) seht.

Die Skalenniveaus unterscheiden sich in ihrem Informationsgehalt. Dieser determiniert die erlaubten (mathematischen) Operationen und damit die statistischen Verfahren, die auf die jeweiligen Daten angewendet werden dürfen.

Nominalskala:	Objekte mit gleicher Merkmalsausprägung erhalten gleiche Zahlen, Objekte mit verschiedenen Merkmalsausprägung erhalten verschiedene Zahlen. Die einzelnen Ausprägungen des Merkmals und damit die Stufen der Skala sind ungeordnet (gleichberechtigt). Beispiele: Augenfarbe, Geschlecht (solange es ohne Wertung bleibt)
Ordinalskala:	Eine Ordinalskala ordnet Objekten Zahlen zu, die so geartet sind, dass von jeweils zwei Objekten das Objekt mit der größeren Merkmalsausprägung die größere Zahl erhält. Über das Ausmaß des Unterschiedes zwischen den einzelnen Ausprägungen ist nichts bekannt. Beispiele: Windstärken, Militärische Ränge, Ranglisten beim Sport
Intervallskala:	Wie bei der Ordinalskala, so sind auch bei der Intervallskala die Stufen geordnet. Zusätzlich sind hier die Abstände zwischen den einzelnen Stufen alle gleich groß. Hat z.B. ein Objekt den Skalenwert 1, so ist es von dem Objekt mit dem Skalenwert 2 genauso weit entfernt, wie dieses von einem Objekt mit dem Skalenwert 3. Der Nullpunkt einer Intervallskala ist relativ. Beispiele: Temperatur in Celsius oder Fahrenheit
Verhältnisskala:	Eine Verhältnisskala ist eine Intervallskala mit absolutem Nullpunkt, der das „Nichtvorhandensein“ des Merkmals ausdrückt. Beispiele: Temperatur in Kelvin, Größe in m; Gewicht in kg, Zeit in sec. Null Kelvin (ca. -273 °C): das Fehlen jedweder atomarer Bewegung

	Mögliche Aussagen:	Mögliche Transformationen:
Nominalskala:	gleich / ungleich	Alle, die die Unterscheidung der Werte nicht verhindern.
Ordinalskala:	gleich / ungleich; größer/ kleiner	Alle, die die Ordnung der Werte nicht verändern.
Intervallskala:	gleich / ungleich; größer/ kleiner; Abstandsvergleiche	Multiplikation/ Division, Addition/ Subtraktion
Verhältnisskala:	gleich / ungleich; größer/ kleiner; Abstands-; Verhältnisvergleiche	nur Multiplikation/ Division (Erhaltung des Nullpunktes)

	Mögliche zentrale Tendenzen:	Mögliche Dispersionsmaße:
Nominalskala:	Modus	---
Ordinalskala:	Modus, Median	Variationsbreite
Intervallskala:	Modus, Median, Arithmetisches Mittel	Variationsbreite, AD-Streuung, Varianz
Verhältnisskala:	Modus, Median, Arithmetisches Mittel	Variationsbreite, AD-Streuung, Varianz

Um noch vernünftige Berechnungen anstellen zu können, geht man u.a. in der psychologischen Forschung meist von Intervallskalenniveau aus, auch wenn das streng mathematisch gesehen nicht vorhanden ist.

[Die Kultusministerkonferenz definiert Schulnoten als intervallskaliert, um so einen Durchschnitt u.a. beim Abitur berechnen zu können, auch wenn z.B. zum Erreichen einer „Vier gut“ 50 % oder zwei Drittel der Punkte zu erreichen sind.]

Maße der zentralen Tendenz: Modus, Median & Arithmetisches Mittel:

Durch welchen Wert wird eine Vielzahl von Werten am besten repräsentiert?

Hierzu geben Maße der zentralen Tendenz Auskunft über die *Lage* bzw. *Lokation* einer Verteilung in Termini der zugrundeliegenden Skala.

Modalwert (Modus):

Der Modus ist der Wert, der in einer Verteilung am häufigsten vorkommt. Es kann auch mehr als einen Modus geben (durch mehrere Maxima; Ausdruck bei Kurvendiskussion: Hochpunkte).

Statistische Abkürzung: **Mo**

Median:

Der Median ist der Wert, der eine der Größe nach geordnete Merkmalsverteilung (also mind. Ordinalskalenniveau) in zwei gleichgroße Hälften teilt. Bei gerader Anzahl von Werten wird das arithmetische Mittel aus dem größten Wert der unteren Hälfte und dem kleinsten Wert der oberen Hälfte bestimmt. Bei ungerader Anzahl wird der Wert in der Mitte genommen.

Statistische Abkürzung: **Md**

Statistische Eigenschaften:

- Die Summe der absoluten Abweichungen vom Median ist minimal.

Arithmetisches Mittel

Das arithmetische Mittel bzw. der Mittelwert ist die Summe aller Messwerte dividiert durch die Anzahl der Messwerte n :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

Statistische Eigenschaften:

- Die Summe der Abweichungen aller Messwerte vom Mittelwert ist Null.
- Die Summe der quadrierten Abweichungen der Messwerte vom Mittelwert ist ein Minimum.

Modus, Median und Arithmetisches Mittel sind im wissenschaftlichen, populärwissenschaftlichen und alltäglichen Gebrauch beliebte Werkzeuge zum Beweisen eigener Thesen:

„Die meisten Abiturienten wählen Jura.“ (auch wenn es nur 8,6 % sind) [Zahl fiktiv!]

„Die besten 50% haben einen Schnitt von 3,4 oder besser.“ (Aha! Und wie verteilt sich das?)

„Im Gegensatz zu den 50er können die Hauptschüler heute im Durchschnitt nicht sauber schreiben.“ (In den 50ern gingen auch fast 90% aller Schüler auf die Hauptschule; heute 30%.)

Weitere Maße der zentralen Tendenz:

Geometrisches Mittel, Harmonisches Mittel, Gewichtetes Mittel (siehe Bortz S. 40)

Dispersionsmaße: u.a. Perzentile, Quadratsummen & Varianzen:

Zwei Verteilungen mit ähnlichen oder sogar gleichen Maßen der zentralen Tendenz können trotzdem sehr unterschiedlich sein, und zwar aufgrund ungleicher Streuungen (Dispersionen) ihrer einzelnen Werte.

Dispersionsmaße geben also Auskunft über die Streuung bzw. Variabilität der Werte.

„Für die empirische Forschung sind Dispersionswerte den Maßen der zentralen Tendenz zumindest ebenbürtig.“ (Bortz)

Perzentile und Variationsbreite:

Das Perzentil $P_{X\%}$ ist der Punkt auf einer Skala, unterhalb dessen X% der aufsteigend angeordneten Messwerte einer Stichprobe liegen. Der Median ist das 50. Perzentil einer Verteilung.

Die **Variationsbreite** (engl. Range) ist die Differenz zwischen dem größten (Maximum) und dem kleinsten Wert (Minimum).

Wenn dieser Wert stark von Extremwerten abhängt, betrachtet man nur einen „eingeschränkten“ Streubereich, z.B. nur die mittleren 90% aller Werte (also den Bereich zwischen dem 5. und 95. Perzentil.). [Bei einer Standardnormalverteilung liegen die Werte zwischen $-\infty$ und $+\infty$, betrachtet man aber „nur“ 99, 74% aller Werte, liegen die Werte zwischen $-3,00$ und $+3,00$!]

Die **Quadratsumme (QS)** ist die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert.

Die **Varianz** ist die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert, dividiert durch die Anzahl der Messwerte n :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}.$$

Die **Standardabweichung** oder **Streuung** ist die Wurzel aus der Varianz:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}.$$

z-Werte:

In der psychologischen Diagnostik ergibt sich oftmals die Aufgabe, Testwerte zweier Personen aus verschiedenen Kollektiven (oder Gruppen, etc.) zu vergleichen.

Deswegen werden die individuellen Leistungen an der Gesamtleistung des (eigenen) Kollektivs relativiert, d.h. standardisiert.

Dies geschieht über eine z-Transformation:

Die z-Transformation ermöglicht den *relativen* Vergleich von Variablenausprägungen, da sie Mittelwerts- und Streuungsunterschiede „wegrelativiert“:

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

Die z-Transformation überführt jede Verteilung in eine Verteilung mit Mittelwert 0 und Streuung bzw. Varianz 1.

Merkmalszusammenhänge & -vorhersagen:

Ich ziehe hier die Bereiche „Kovarianz und Korrelation“, „Korrelation und Kausalität“, „Lineare Regression“, „Beobachtete, vorhergesagte und Residualwerte“ und „Determinationskoeffizient“ aus Kapitel 6 im Bortz vor, da mir eine Thematisierung in Kapitel 6 als zu spät erscheint.

Kovarianz und Korrelation:

Mit der **Kovarianz** lässt sich bestimmen, wie sich die relativen Positionen von gepaarten Messwerten aus zwei Variablen zueinander verhalten:

$$\text{COV}_{(x,y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Die Kovarianz kann nur hinsichtlich ihres Vorzeichens, nicht aber hinsichtlich ihrer Größe interpretiert werden. [Da man auch mit Inches statt cm oder m statt cm arbeiten kann, kann die Größe der Kovarianz stark schwanken.] [Die Kovarianz ist bei Monte-Carlo-Studien sehr robust.]

- Ein positives Vorzeichen der Kovarianz zeigt an, dass sowohl überdurchschnittliche Werte von X stärker mit überdurchschnittlichen Werten von Y gepaart sind als auch, dass unterdurchschnittliche Werte von X stärker mit unterdurchschnittlichen Werten von Y zusammen auftreten.
- Ein negatives Vorzeichen zeigt entsprechend an, dass überdurchschnittliche Werte von X stärker mit unterdurchschnittlichen Werten von Y gepaart sind und umgekehrt.

- Keine Kovarianz entsteht dann, wenn sich die Werte von X und Y ausgleichen. Dies macht eine Interpretation unmöglich.
- Nur zur Erinnerung: selbst der kleinste positive lineare Zusammenhang ergibt eine positive Kovarianz! Damit man das auch erkennen kann, nutzt man Korrelationen.

Man standardisiert eine Kovarianz und erhält eine Korrelation:

Die Kovarianz kann maximal den Wert $s_x \cdot s_y$ annehmen. Dies ist damit ein sehr schöner Term zur Relativierung bzw. Standardisierung von Kovarianzen. (Man macht so was wie eine z-Transformation, nur für beide Variablen (X, Y) gleichzeitig.)

Diese **Standardisierung** nennt man *Produkt-Moment-Korrelation* $r_{(x,y)}$.

$$r_{(x,y)} = \frac{COV_{(x,y)}}{s_x \cdot s_y}.$$

Die Korrelation $r_{x,y}$ kann Werte zwischen -1 und $+1$ annehmen, wobei

- -1 einen perfekt negativen linearen Zusammenhang anzeigt,
- $+1$ einen perfekt positiven linearen Zusammenhang anzeigt und
- 0 keinen linearen Zusammenhang anzeigt.

Perfekter linearer Zusammenhang: Die Werte der einen Variablen lassen sich durch eine Lineartransformation in die der anderen überführen:

$$y_i = a + b \cdot x_i$$

(Und schon schließt sich der Kreis zu Linearer Regression.)

[Die Abkürzung „r“ bei der Korrelation ist auf das Wort „Regression“ zurückzuführen, da Korrelations- wie Regressionsrechnung eng miteinander verknüpft sind.]

Korrelation und Kausalität:

Die Kovariation zweier Variablen kann auf eine kausale Beziehung der beiden Variablen oder auf die Beeinflussung durch eine oder mehrere *Drittvariablen* zurückgehen. Auf Basis einer Korrelation kann **nicht** auf die Art der Verursachung geschlossen werden. (Z.B. korrelieren bei Kindern Intelligenz und Schuhgröße.) **Kausalität lässt sich nur widerlegen, nicht beweisen.**

Lineare Regression:

Hängen zwei Variablen (wie z.B. Körperlänge und Gewicht) zusammen, so ist es möglich, eine Variable jeweils auf Basis der anderen vorherzusagen.

Die Vorhersagevariable wird als *unabhängige Variable* oder *Prädiktorvariable* (X) bezeichnet, die Variable, die vorhergesagt werden soll, als *abhängige Variable* oder *Kriteriumsvariable* (Y):

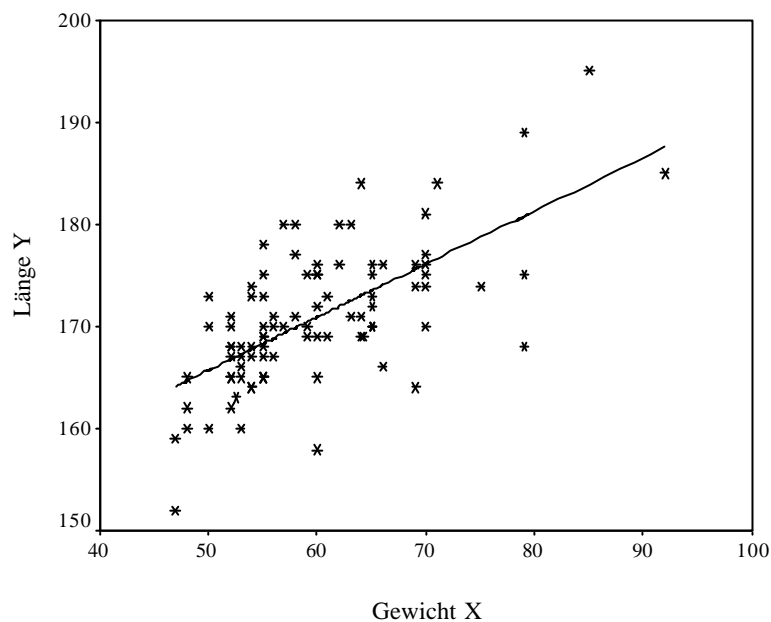
(Prädiktorvariablen sind im allgemeinen einfacher und billiger zu messen; z.B. IQ-Tests)

Im Rahmen der bivariaten Regression wird die Vorhersage über folgende lineare Beziehung vorgenommen:

$$\hat{y}_i = a + b \cdot x_i, \text{ wobei}$$

\hat{y}_i = vorhergesagter Wert, a = Schnittpunkt der Geraden mit der y-Achse (Ordinatenabschnitt, Interzept) und b = Steigung der Geraden.

Ziel ist die Bestimmung einer Geraden, die den Gesamttrend aller Punkte am besten wiedergibt.



Gleichungen:

$$b_{yx} = \frac{\text{COV}_{(x,y)}}{s_x^2}; \quad a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

(Grafiken im Bortz S. 181)

Bei der linearen Regression wird die Gleichung gesucht, für die die Summe der quadrierten Abweichungen zwischen vorhergesagten und tatsächlichen Werten minimal ist (= Kriterium

der kleinsten Quadrate). $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min.$, wobei \hat{y}_i = Wert auf der Geraden

Beobachtete, vorhergesagte und Residualwerte:

Die Abweichungen der beobachteten Werte von den vorhergesagten Werten heißen *Regressionsresiduen*:

$$y_i^* = y_i - \hat{y}_i.$$

Die Residuen enthalten die Anteile der Kriteriumsvariablen (Y), die durch die Prädiktorvariable nicht erfasst werden:

- Die Mittelwerte der Kriteriumsvariablen und der vorhergesagten Werte sind gleich.
- Der Mittelwert der Residuen ist 0.

Diese Anteile der Kriteriumsvariablen sind auf Messfehler, vor allem aber auch durch Bestandteile des Kriteriums, die durch andere, mit der Prädiktorvariablen nicht zusammenhängende Merkmale, erklärt werden können (z.B. Störvariablen).

Zerlegung der Kriteriumsvarianz:

Die Varianz der y-Werte setzt sich additiv aus der Varianz der vorhergesagten \hat{y} -Werte und der Varianz der Residuen y^* zusammen:

$$s_y^2 = s_{\hat{y}}^2 + s_{y^*}^2.$$

Determinationskoeffizient:

= Varianzanteil der abhängigen Variablen, der mit der unabhängigen vorhergesagt bzw. erklärt werden kann:

$$r_{(y, \hat{y})}^2 = \frac{s_{\hat{y}}^2}{s_y^2}.$$

Dabei kann man die „Power“ einer Regressionsgeraden erfassen, d.h. wie viel Prozent sagt sie vorher und wie viel bleibt im Dunkeln.

Korrelationen:

- Die Korrelation zwischen Kriterium und Vorhersage ist gleich der Korrelation zwischen Kriterium und Prädiktor:

$$r_{(y, \hat{y})} = r_{(y, x)}.$$

- Die Korrelation zwischen Prädiktor (X) und Residuen (von Y) ist 0: (muss ja, die Residuen sind ja der Teil, den der Prädiktor X nicht aufklärt)

$$r_{(x, y^*)} = 0.$$

- Für die Korrelation zwischen Kriterium und Residuen gilt:

$$r_{(y, y^*)}^2 = 1 - r_{(x, y)}^2.$$

Wahrscheinlichkeitstheorie und -verteilungen:

Eine der wichtigsten Eigenschaften des Menschen ist es, Redundanzen (unnötige Ausführlichkeit einer Information) zu erkennen und zu erlernen. Dies verhindert, dass der Mensch in dem Chaos von Zufälligkeiten und in der Fülle von Informationen zugrunde geht.

Der Mensch schafft sich ein Ordnungssystem (z.B. Zahlen, Buchstaben, Gesetze), an dem er sein Verhalten orientiert – in (teils blindem) Vertrauen darauf, dass das System auch funktioniert.

Jedoch gibt es in der Welt keine absoluten Wahrheiten und Wahrscheinlichkeiten. Wir regulieren unser Verhalten nach einem System unterschiedlich wahrscheinlicher Hypothesen, die teils völlig blöde sind (Geld in Glücksspiele investieren), teils sehr vernünftig sind (wir steigen in ein Flugzeug, weil es sehr unwahrscheinlich ist, dass es abstürzt.).

Der Mensch schafft sich Grundsätze (in der Mathematik: Axiome), die er beweislos voraussetzt. Aus mehreren Axiomen (einem Axiomssystem) können Lehrsätze nach den Regeln der Logik hergeleitet werden. Wichtig dabei:

- 1.) **Widerspruchsfreiheit**
- 2.) **Unabhängigkeit:** kein Axiom lässt sich aus einem anderen herleiten.
- 3.) **Vollständigkeit:** alle Axiome reichen aus, um alle inhaltlich richtigen Sätze zu beweisen.

Zufallsexperimente:

Für die Definition objektiver Wahrscheinlichkeiten (das will ja der Methodiker) ist der Begriff des „Zufallsexperimentes“ zentral:

Ein Zufallsexperiment ist ein beliebig oft wiederholbarer wohldefinierter Vorgang, dessen Ausgang nicht vorhersagbar ist.

Ein Ausgang wird als *Ergebnis* oder *Elementarereignis* ω , die Menge aller *Elementarereignisse* wird als *Ergebnisraum* $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ bezeichnet.

Bei der *theoretischen (kombinatorischen) Bestimmung* einer *Wahrscheinlichkeit* geht man oft von der Gleichwahrscheinlichkeit aller Ergebnisse aus:

$$p(A) = \frac{\text{Anzahl günstiger Ergebnisse}}{\text{Anzahl möglicher Ergebnisse}} .$$

Günstige Ergebnisse sind solche, bei denen sich A ereignet, mögliche Ergebnisse sind alle Ergebnisse des endlichen Ergebnisraumes (A kann durchaus mehr als ein einziges ω_i umfassen).

Ein Beispiel:

Ich werfe einen perfekten sechsseitigen Würfel: Die Chance, das ich eine „sechs“ erhalte, beträgt also ein Sechstel, oder, mathematisch formuliert:

$$p(\text{Wurf}) = \frac{1}{6}$$

Grundbegriffe der Wahrscheinlichkeitsrechnung:

disjunkt = unvereinbar; die Ereignisse haben keine gemeinsame Menge, keinen **Durchschnitt**

p = Possibility (Wahrscheinlichkeit)

$p(A)$ = Wahrscheinlichkeit von A

$p(\overline{A})$ = Wahrscheinlichkeit des Fehlens von A

$p(A \cap B)$ = die Wahrscheinlichkeit von A und B (gleichzeitig)

$p(A \cup B)$ = die Wahrscheinlichkeit von A und/oder B (also entweder A oder B oder A&B gleichzeitig!)

- Das Ereignis, dass sich A und B ereignen, wird als **Durchschnitt** bezeichnet.
Symbol: \cap (= logisches Produkt)
- Das Ereignis, dass sich A und/oder B ereignen, wird als **Vereinigung** bezeichnet.

Bei unvorhersehbaren Wahrscheinlichkeiten muss man über viele Versuche die Wahrscheinlichkeiten schätzen (z.B. Reiszwecken werfen: zu wie viel Prozent landet es mit der Spitze nach oben, zu wie viel mit der Spitze schräg nach unten?) (je größer n , desto genauer):

Ein *statistischer Schätzer* der Wahrscheinlichkeit eines Ereignisses A ist die relative Häufigkeit eben dieses Ereignisses:

$$h(A) = \frac{f(A)}{n}.$$

Axiome der Wahrscheinlichkeitstheorie:

- 1.) Für die Wahrscheinlichkeit eines zufälligen Ereignisses A gilt $p(A) \geq 0$ (Nicht- Negativität).
- 2.) Die Wahrscheinlichkeit eines sicheren Ereignisses ist gleich 1 (Normierung).
- 3.) Die Wahrscheinlichkeit, dass eines der disjunkten Ereignisse A_1 oder A_2 oder ... A_k auftritt, ist gleich der Summe der Einzelwahrscheinlichkeiten $p(A_1)$, $p(A_2)$, ... $p(A_k)$.

Additionstheorem für vereinbare bzw. nicht disjunkte Ereignisse:

- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

Bei disjunkten bzw. nicht vereinbaren Ereignissen:

- $p(A \cup B) = p(A) + p(B)$.

Die bedingte Wahrscheinlichkeit $p(A|B)$ ist die Wahrscheinlichkeit von A, sofern B bereits eingetreten ist:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}.$$

Multiplikationstheorem für **abhängige** Ereignisse:

$$p(A \cap B) = p(A) \cdot p(B|A) \quad \text{oder}$$

$$p(A \cap B) = p(B) \cdot p(A|B)$$

Multiplikationstheorem für **unabhängige** Ereignisse:

$$p(A \cap B) = p(A) \cdot p(B)$$

Zwei Ereignisse A und B sind *stochastisch unabhängig*, wenn A unabhängig von B eintreten kann, d.h. das Eintreten von B ändert nichts an der Eintretenswahrscheinlichkeit von A, und umgekehrt.

Das Theorem von Bayes:

Das Theorem von Bayes verknüpft die bedingten Wahrscheinlichkeiten $p(A|B)$ und $p(B|A)$ unter Verwendung des Satzes der totalen Wahrscheinlichkeit (Bortz S. 57). Eine der beiden bedingten Wahrscheinlichkeiten kann damit von der anderen aus erschlossen werden.

In der statistischen Entscheidungstheorie hat somit das Theorem von Bayes eine besondere Bedeutung; statistischen Entscheidungen werden immer aufgrund bedingter Wahrscheinlichkeiten getroffen.

[Die Wahrscheinlichkeit für das Auftreten bestimmter Daten (D) unter der Bedingung, dass eine bestimmte Hypothese (H) richtig ist: $p(D|H)$]

Variationen, Permutationen, Kombinationen:

Insbesondere durch Glücksspiele wurde eine Reihe von Rechenregeln erarbeitet, mit denen die Wahrscheinlichkeit bestimmter Ereigniskombinationen von gleichwahrscheinlichen Elementarereignissen ermittelt wird.

1. Variationsregel: Wenn jedes von k sich gegenseitig ausschließenden Ereignissen bei jedem Versuch auftreten kann, ergeben sich bei n Versuchen k^n verschiedene Ereignisabfolgen.

Beispiel: die Chance, fünf mal hintereinander beim Münzwurf „Zahl“ zu erhalten (2 mögliche Ereignisse), beträgt $k^n = 2^5 = 32$ verschiedene Ereignisabfolgen. Da es sich um ein günstiges Ereignis handelt, beträgt die Wahrscheinlichkeit $p = 1/32 = 0,031$

2. Variationsregel: Werden n voneinander unabhängige Zufallsexperimente durchgeführt und besteht die Ereignismenge des 1. Zufallsexperiments aus k_1 , die Ereignismenge des 2. Zufallsex-

periments aus k_2, \dots und die Ereignismenge des n -ten Zufallsexperiments aus k_n verschiedenen Elementarereignissen, sind $k_1 \times k_2 \times \dots \times k_n$ Ereignisabfolgen möglich (\times bedeutet „mal“).

Beispiel: Ein Münzwurf und ein Würfelwurf, man möchte Zahl und eine 6 erhalten. Es ergeben sich $2 \times 6 = 12$ Ereignismöglichkeiten. Die Wahrscheinlichkeit beträgt also $p = 1/12 = 0.08$.

Permutationsregel: n verschiedene Objekte können in $n! = 1 \times 2 \times 3 \times \dots \times (n-1) \times n$ verschiedenen Abfolgen angeordnet werden. ($n!$ bedeutet n Fakultät)

Beispiel: In einer Urne befinden sich 6 unterschiedlich schwere Kugeln. Wie groß ist die Chance, die Kugeln in der Reihenfolge ihres Gewichtes herauszunehmen? (Es ist also ein unterschiedliches Ereignis ohne Zurücklegen; ergo ein abhängiges Ereignis)
 $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ mögliche Abfolgen. Nur eine ist richtig, also $p = 1/720 = 0,0014$

1. Kombinationsregel: Wählt man aus n verschiedenen Objekten r zufällig aus, ergeben sich $n! / (n-r)!$ verschiedene Reihenfolgen der r Objekte.

Beispiel: Ich will aus einem Kartenspiel (32 Karten) in vier Zügen die Asse herausziehen. Die Reihenfolge der Asse ist **festgelegt**:

$$\frac{32!}{(32-4)!} = 32 \times 31 \times 30 \times 29 = 863040$$

Die Wahrscheinlichkeit beträgt also $1/863040 = 1,16 \times 10^{-6}$ Möglichkeiten

2. Kombinationsregel: Wählt man aus n verschiedenen Objekten r zufällig aus und lässt hierbei die Reihenfolge außer acht, ergeben sich für die r Objekte $\binom{n}{r}$ verschiedene Kombinationen.

Beispiel: Wie groß ist die Wahrscheinlichkeit, 6 Richtige aus 49 zu haben? (Reihenfolge ist im Gegensatz zur 1. Kombinationsregel ja egal)

Da wir noch alle wissen, dass

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \text{ ist also eingesetzt: } \binom{49}{6} = \frac{49!}{6!(49-6)!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} =$$

$$= 13.983.816 \quad (\text{Noch genauere Rechenschritte: Bortz S. 61})$$

Die Wahrscheinlichkeit, sechs Richtige im Lotto zu haben beträgt also $1/13.983.816 = 7,15 \times 10^{-8}$ Möglichkeiten.

(Und die Lottobranche boomt. Sollten wir Psychologen da nicht mal eingreifen?)

3. Kombinationsregel: Sollen n Objekte in k Gruppen der Größen n_1, n_2, \dots, n_k eingeteilt werden (wobei $n_1 + n_2 + \dots + n_k = n$), ergeben sich $n! / (n_1! \dots n_k!)$ Möglichkeiten.

Beispiel: In einer Urne sind 4 rote, 3 blaue und 3 grüne Kugeln. Wie groß ist die Wahrscheinlichkeit, erst die 4 roten Kugeln zusammen, denn die 3 blauen und zuletzt die 3 grünen Kugeln zu entnehmen?

$$\frac{10!}{4!3!3!} = 4200$$

Die Wahrscheinlichkeit, erst die 4 roten Kugeln zusammen, denn die 3 blauen und zuletzt die 3 grünen Kugeln zu entnehmen, beträgt also $1/4200 = 2,38 \times 10^{-4}$ Möglichkeiten.

Verteilungsfunktionen:

Wenn man Experimente durchführt, erhält man jede Menge Daten (z.B. Studiengang von allen Studenten der Uni Bonn). Diesen Daten kann man bestimmte Kennungen (z.B. 1, 2, 3, etc.) hinzufügen, und so erhält man eine **Tabelle**. Das ist eine **diskrete** Wahrscheinlichkeitsfunktion.*

Diskrete Wahrscheinlichkeitsfunktionen bestehen aus einer endlichen Anzahl von Ereignissen. Auftretenswahrscheinlichkeiten einzelner Ereignisse lassen sich damit bestimmen.

Beispiele: Würfelergebnisse, Noten (manche haben gleiche Ergebnisse)

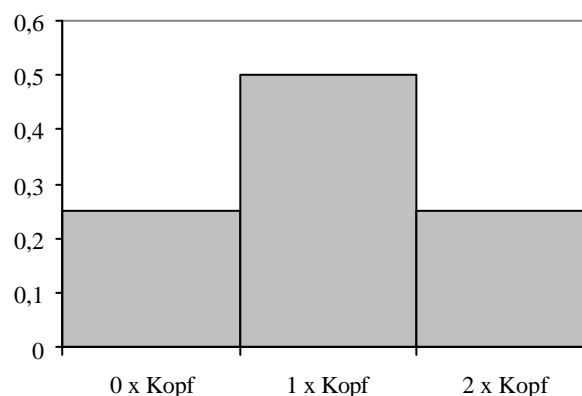
Man kann aber Daten auf bestimmte Kennungen zufügen, d.h. z.B. ich ordne alle Studenten der Uni Bonn der Größe nach. So erhält man eine **Kurve**. Das ist eine **stetige** Wahrscheinlichkeitsfunktion.*

Stetige Wahrscheinlichkeitsfunktionen bestehen aus einer unendlichen Anzahl von Ereignissen. Auftretenswahrscheinlichkeiten einzelner Ereignisse lassen sich damit nicht bestimmen.

Beispiele: Körpergröße, Geschwindigkeit (Niemand ist gleich groß, wenn man es genau nimmt.)

* Diese Beispiele sind von mir sind konstruiert. Ich kann z.B. den Zahlen 1 bis X die Studiengänge der Studenten der Uni Bonn zuordnen, bin deswegen aber nicht bei einer stetigen Wahrscheinlichkeitsfunktion.

Der eigentliche Unterschied zwischen STETIG und DISKRET ist in dem **fett** Geschrieben festgehalten. (Bei Stetigkeit Kurven und bei Diskretheit Tabellen zu benutzen stimmt aber.)

Diskrete Verteilungen:

\bar{x} und s^2 einer empirischen Verteilung / Stichprobe werden durch μ (Erwartungswert) und σ^2 bei einer theoretischen Verteilung einer Zufallsvariablen / Population ersetzt, bei diskreter Verteilung

ergibt sich $m = \sum_{i=1}^N x_i \cdot p_i$ und $s^2 = \sum_{i=1}^N p_i \cdot (x_i - m)^2$.

Binomialverteilung:

Ein Experiment, welches in der n -fachen Wiederholung von unabhängigen Einzelexperimenten mit nur zwei möglichen Ausgängen besteht, heißt **Bernoulli-Experiment**. Wenn ein Ereignis X in einem Einzelexperiment mit einer Wahrscheinlichkeit p auftritt, dann kann die Wahrscheinlichkeit P , dass X in n Wiederholungen k -mal auftritt, allgemein bestimmt werden:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \quad [1-p = q]$$

Diese **Wahrscheinlichkeitsfunktion** heißt *Binomialverteilung* und gibt an, wie wahrscheinlich das Eintreten eines Ereignisses $X = k$ ist.

Beispiel: Ich möchte wissen, wie groß die Chance ist in 10 Münzwürfen genau 7x „Zahl“ zu erhalten. Die Wahrscheinlichkeit einer Zahl beträgt 50%, mathematisch formuliert:

$$p = 0,5 = 1/2$$

Wenn ich jetzt einsetzte: [$n = 10$, da 10 Würfe; $k = 7$, da 7x „Zahl“ gewünscht]

$$p(X = 7) = \binom{10}{7} \cdot \left(\frac{1}{2}\right)^7 \cdot \left(\frac{1}{2}\right)^3 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \cdot \left(\frac{1}{2}\right)^{10} = 0,117$$

Die Wahrscheinlichkeit, mit 10 Münzen genau 7 mal Zahl zu treffen, beträgt 11,7%.

Die **Verteilungsfunktion** hingegen gibt an, mit welcher Wahrscheinlichkeit ein Ereignis höchstens k -mal auftritt:

$$P(X \leq k) = \sum_{j=0}^k \binom{n}{j} \cdot p^j \cdot (1-p)^{n-j}$$

Die Verteilungsfunktion ist die kumulierte Wahrscheinlichkeitsfunktion.

Andere diskrete Verteilungen:

- Hypergeometrische Verteilung: wird benutzt bei **schwankenden Wahrscheinlichkeiten** (z.B. durch Ziehen ohne Zurücklegen)
- Poisson-Verteilung: **Verteilung seltener Ereignisse**; wird benutzt bei großem n & kleinem k .
- Multinominale Verteilung: auch Polynominalverteilung genannt; ermöglicht **mehrere unterschiedlich große k_x** .
- Negative Binomialverteilung: ermittelt die Wahrscheinlichkeit, dass ein Ereignis nach x Versuchen erst dann eintritt, wird häufig benutzt zur **Analyse von Wartezeiten**.

Stetige Verteilungen 1: Normalverteilung & Standardnormalverteilung:**Für alle stetigen Verteilungen geltend:**

- Die Gesamtfläche unter der Kurve ist auf 1 normiert, was bedeutet, dass mit einer Wahrscheinlichkeit 1 irgendein Ergebnis eintritt. Die Fläche über einem Intervall von zwei Werten gibt an, mit welcher Wahrscheinlichkeit Werte innerhalb dieses Intervalls eintreten.

Normalverteilung(NV):

Die Normalverteilung ist die für die Statistik wichtigste Verteilung; aus ihr werden weitere stetige Verteilungen wie u.a. die χ^2 -Verteilung, die t-Verteilung sowie die F-Verteilung abgeleitet; diskrete Verteilungen wie die Binomial- und Poisson-Verteilung konvergieren gegen sie.

Die Zusammenhänge zwischen χ^2 -Verteilung, t-Verteilung sowie F-Verteilung folgen.

Alle Normalverteilungen haben typische Eigenschaften:

- Die Verteilung hat einen glockenförmigen Verlauf.
- Die Verteilung ist symmetrisch.
- Modus, Median und Mittelwert fallen zusammen.
- Es gibt nur einen Modus (wie auch bei χ^2 -Verteilung, t-Verteilung sowie F-Verteilung.)
- Die Verteilung nähert sich asymptotisch der x-Achse (geht also von $-\infty$ bis $+\infty$).
- Zwischen den zu den Wendepunkten gehörenden x-Werten befinden sich ungefähr 2/3 der Gesamtfläche.
- Zwischen den Werten $\bar{x} + s$ (= Streuung) und $\bar{x} - s$ liegen 68,26% aller Werte, im Bereich $\bar{x} \pm 2s$ liegen 95,44%.
- Binomialverteilung ist für $n \Rightarrow \infty$ gleich der Normalverteilung (S. 78 Bortz)

Fesche Bilder von Normalverteilungen im Bortz auf S. 74.

Normalverteilungen unterscheiden sich durch

- a.) unterschiedliche Erwartungswerte (μ) [also Mittelwerte] und
- b.) unterschiedliche Streuungen (σ).

Zwei Normalverteilungen mit gleichem σ und μ sind identisch. Die Normalverteilung wird somit durch die beiden Parameter σ und μ festgelegt. Ihre Dichtefunktion lautet:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-(x-\mu)^2 / 2\sigma^2}.$$

Damit klar ist, was diese Funktion da oben soll:

Mithilfe von Funktionen kann man durch Einsetzen des x-Wertes den y-Wert bestimmen und erhält somit eine wunderschöne Kurve, wie z.B. bei der Parabel: $f(x) = x^2$

Diese Funktion da oben gibt dann den y-Wert an, der aber gleichzeitig die Wahrscheinlichkeit des Eintretens dieses (eingesetzten) x-Wertes (in einem Intervall ΔX) angibt. (in der Tabelle als „Ordinate“ bezeichnet.) Über ein Integral dieser Dichtefunktion kann die Fläche berechnet werden (das ist ja der Sinn von Integralen). Alles klar?

Grundprinzip mathematisch ausgedrückt:

Die Wahrscheinlichkeitsfunktion einer stetigen Verteilung wird als Wahrscheinlichkeitsdichte oder **Dichtefunktion** bezeichnet. Erst über Integrale dieser Dichtefunktion können Wertebereichen Wahrscheinlichkeiten zugeordnet werden.

Standardnormalverteilung (STNV):

Die Standardnormalverteilung (STNV) ist eine ganz normale Normalverteilung; wird aber als Grundlage aller Vergleiche zwischen Normalverteilungen genutzt, da sie einen Mittelwert $\mu = 0$ und eine Streuung $\sigma = 1$ [Abk.: $N(0, 1)$] hat.

Durch eine z-Transformation kann jede Normalverteilung in die Standardnormalverteilung überführt werden; somit sind Vergleiche zwischen allen Normalverteilungen möglich!!

Ihre Dichtefunktion lautet:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2}.$$

[Man hat hier x durch z ersetzt, um zum Ausdruck zu bringen, dass diese Funktion nur zur Standardnormalverteilung gehört.]

Diese Dichte (siehe „Ordinate“) und deren Verteilungsfunktion (siehe „Fläche“) ist im Bortz in Tabelle B (S. 768-772) wiedergegeben. [Einen z -Wert in der Tabelle nachzusehen setzte ich als bekannt voraus.]

Das hier ist übrigens die Integralformel, die benötigt wird, um die zu den z -Werten zugehörigen Flächenanteile [von $-\infty$ bis a] bei der Standardnormalverteilung (STNV) auszurechnen:

$$p(z < a) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2} dz$$

Schön, nicht? Die anderen Formeln sind noch komplizierter (das ist aber, ebenso wie diese Formel, nicht relevant für die Prüfung – nur zur Anschauung gedacht :-).

Die **Bedeutsamkeit der Normalverteilung (NV)** lässt sich auf 4 Aspekte zurückführen:

- die NV als empirische Verteilung: eine Reihe von Messungen verteilt sich meistens **annähernd** normal, z.B. Körpergröße, -gewicht, Testleistungen. (Im 19. Jahrhundert entdeckt, heute relativiert.)
- die NV als Verteilungsmodell für statistische Kennwerte: **wenn man genug Stichproben zieht verteilen sich die Mittelwerte der Stichproben normal (Inferenzstatistik!)**.
- die NV als mathematische Basisverteilung: Verknüpfungen u.a. zu χ^2 -, t - sowie F -Verteilung.

- die NV in der statistischen Fehlertheorie: kaum eine Messungen ist absolut fehlerfrei. Misst man andauernd, erhält man irgendwann normalverteilte Werte. Der Mittelwert ist das wahrscheinlichste richtige Ergebnis (Strategie der alten Römer u.a. beim Aquäduktbau.).

Stetige Verteilungen 2: χ^2 -Verteilung, t-Verteilung sowie F-Verteilung:

χ^2 -Verteilung:

Gegeben sei eine normalverteilte Zufallsvariable z mit $\mu = 0$ und $\sigma = 1$, also eine STNV. Das Quadrat dieser Zufallsvariablen bezeichnen wir als χ^2_1 -verteilte Zufallsvariable.

$$\chi^2_1 = z^2$$

(Die untere Zahl am χ^2 kennzeichnet die Anzahl der normalverteilten Zufallsvariablen z .)

Wenn wir jetzt (theoretisch unendlich) viele χ^2_1 -Werte aus zufällig gezogenen z -Werten nach der oben stehenden Formel ermitteln, erhalten wir eine stetige χ^2_1 -Verteilung. (Dichtefunktion steht nicht im Bortz) [Ein Bild verschiedener χ^2 -Verteilungen im Bortz auf S. 80.]

$$\chi^2_n = z_1^2 + z_2^2 + \dots + z_n^2$$

χ^2 -Verteilungen haben:

- in Abhängigkeit von der Anzahl der z^2 - Variablen unterschiedliche Freiheitsgrade bzw. df (degrees of freedom)= n
- eine Variationsbreite von 0 bis $+\infty$
- bei $df = n$ einen Erwartungswert von $m = n$, eine Streuung von $\sqrt{2n}$ und eine Schiefe (ob es linkssteil oder rechtssteil ist.) von $\sqrt{8/n}$.

Mit größer werdendem n nähert sich die χ^2 -Verteilung einer NV mit $m = n$ und $\sigma = \sqrt{2n}$ an.

t-Verteilung:

Aus einer STNV wird ein z Wert und aus einer hiervon unabhängigen χ^2 -Verteilung ein χ^2_n -Wert gezogen. Der folgende Quotient definiert einen t_n -Wert:

$$t_n = \frac{z}{\sqrt{\chi^2_n / n}}$$

Ein Bild einer t-Verteilung im Bortz auf S. 81.

t-Verteilungen:

- unterscheiden sich durch Anzahl der Freiheitsgrade (= Freiheitsgrade der χ^2 -Verteilung; die steckt ja in den Werten drinnen – siehe Formel).
- nähert sich mit wachsenden Freiheitsgraden (n) der Normalverteilung an.

- haben einen Erwartungswert $\mu = 0$
- sind symmetrisch
- gehen bei $df \rightarrow \infty$ in eine NV über (Grenzpunkt Definitionssache)

F-Verteilung:

Gegeben sei eine χ^2 -Verteilung mit $df_1 = n_1$ und eine weitere, unabhängige χ^2 -Verteilung mit $df_2 = n_2$. Der Quotient von 2 zufällig aus diesen beiden Verteilungen entnommenen χ^2 -(Einzel-) Werten, multipliziert mit dem Kehrwert des Quotienten ihrer Freiheitsgrade (d.h. die df sind Zähler/Nenner-mäßig vertauscht), wird als F-Wert bezeichnet:

$$F_{(n_1, n_2)} = \frac{\chi_{n_1}^2}{\chi_{n_2}^2} \cdot \frac{n_2}{n_1}$$

F-Verteilungen:

- sind asymmetrische, stetige Verteilungen mit einer Variationsbreite von 0 bis $+\infty$.
- unterscheiden sich durch Zähler- (n_1) und Nenner- (n_2) Freiheitsgrade

Ein Bild von F-Verteilungen ist im Bortz auf S.82.

Noch ein paar Zusammenhänge:

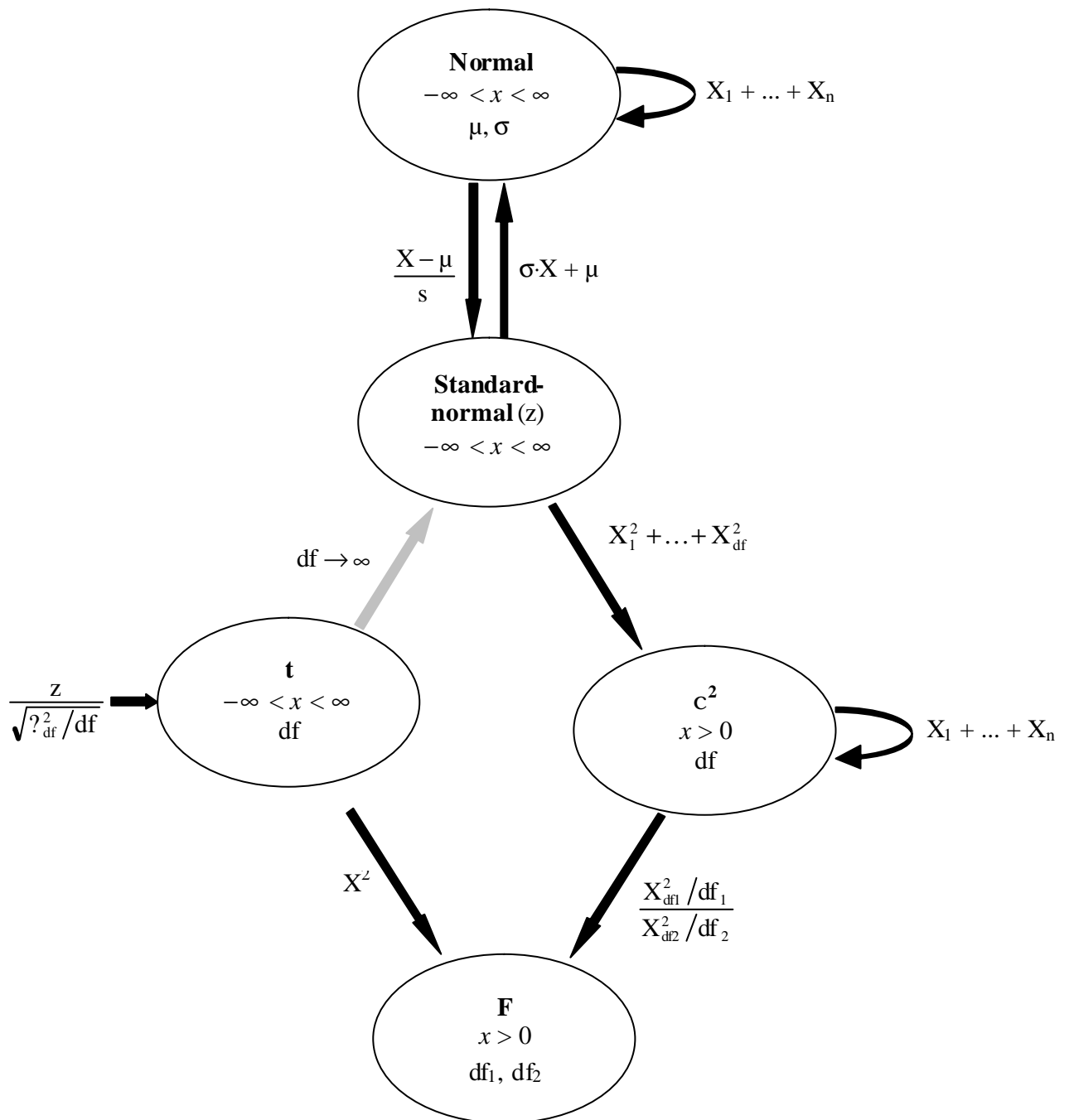
- Ein F-Wert in der Tabelle bei $df_1 = 1$ und $df_2 = \infty$ entspricht einem z-Wert der STNV bei halbiertem Alphaniveau (da die STNV zweiseitig ist muss man beide Grenzwerte
- Eine quadrierte t-Verteilung mit $df = n$ ist mit der F-Verteilung für einen Zähler- und n Nennerfreiheitsgrade identisch. Als Formel:

$$t_n^2 = \frac{\chi_1^2}{\chi_2^2} \cdot \frac{n}{1} = F_{(1, n)}$$

- Zwischen einer χ^2 -Verteilung und einer F-Verteilung besteht der folgende Zusammenhang:

$$F_{(n, \infty)} = \chi_n^2 / n$$

Zusammenhänge von NV, STNV, χ^2 -, t- und F-Verteilung:



Stichprobe und Grundgesamtheit:

Die meisten Untersuchungen in u.a. der Psychologie sind dahin ausgerichtet, aufgrund einer kleinen (in einem Experiment) untersuchten Stichprobe Schlüsse auf die Allgemeinheit zu ziehen. Der sich hiermit befassende Teilbereich der Statistik wird als **Inferenz-** oder **schließende Statistik** bezeichnet.

Inferenzstatistik hat das Ziel, auf Basis von Stichprobenergebnissen Aussagen über die zugrundeliegende Population zu treffen.

Zur Inferenzstatistik zählen

- die Schätzung von Populationsparametern und
- das Testen (Überprüfen) von Hypothesen.

Begriffe der Inferenzstatistik:

- **Grundgesamtheit** oder **Population** bezeichnet alle potenziell untersuchbaren Elemente (Personen, Objekte), die ein gemeinsames Merkmal oder eine gemeinsame Merkmalskombination aufweisen (z.B. Männer, Studenten, Deutsche, VW Golf, Gartenzwerge, etc.).
- **Stichprobe** bezeichnet eine nach einer bestimmten Auswahlmethode gewonnene Teilmenge von Elementen aus der Population.
Inferenzstatistische Verfahren erfordern, dass die Stichprobe mittels einer Zufallsprozedur gewonnen wird.
- **Statistische Kennwerte**, wie die Maße der zentralen Tendenz oder die Maße der Dispersion können für Stichproben wie für Populationen (Grundgesamtheiten) ermittelt werden.
Die Kennwerte einer Population bezeichnen wir als **Parameter**, bzw. **Populationsparameter**.
Ein Stichprobenkennwert wird als **Punktschätzer** bezeichnet.
- **Unbekannte Parameter** (sind in der Regel unbekannt) werden mit den Stichprobenkennwerten geschätzt.
- **Zur Unterscheidung** von Stichprobenkennwerten und Populationsparametern werden Kennwerte mit lateinischen und Parameter mit griechischen Buchstaben notiert. (manchmal werden für Parameter einfach die großen Buchstaben der Kennwerte genutzt; also N statt n . Die Bezeichnung der Kennwerte variiert nicht.)

	Stichprobenkennwert:	Populationsparameter:
Mittelwert:	\bar{x}	μ
Varianz:	s^2	σ^2
Streuung:	s	σ
Korrelation:	r	ρ

Mit endlichen Populationen – z.B. die Psychologie studenten im 3. Semester der Uni Bonn – zu arbeiten ist ausgesprochen schwierig (letztlich ist jede Population endlich, aber ab gewissen Größen – z.B. alle Studenten Deutschlands – gilt sie statistisch als unendlich), dementsprechend versucht man mit unendlichen Populationen zu arbeiten.

- Damit eine Population als unendlich gilt, sollte sie mindestens 100x so groß sein wie die Stichprobe.

Stichprobenarten:

Wie genau kann gewährleistet werden, dass eine Stichprobe eine Grundgesamtheit möglich genau repräsentiert?

Eine Stichprobe kann in bezug auf die Grundgesamtheit entweder in bezug auf

- a.) alle Merkmale (globale Repräsentativität) oder
- b.) bestimmte Merkmale (spezifische Repräsentativität)

repräsentativ sein.

- **Zufallsstichprobe:** Jedes Element der Grundgesamtheit kann mit gleicher Wahrscheinlichkeit ausgewählt werden (Ziehen mit Zurücklegen).
Eine absolut fehlerfreie Zufallsstichprobe ist fast unmöglich; meist aufgrund zu kleiner Stichprobengrößen, mangelnder Repräsentativität und mangelnder Bereitschaft der Menschen, an einem Experiment teilzunehmen.
- **Klumpenstichprobe:** Elemente werden nicht einzeln, sondern in natürlichen Gruppen (Klumpen) ausgewählt [z.B. alle Alkoholiker in Kliniken].
Hierbei ist zu beachten, dass die Klumpen zufällig ausgewählt werden (vgl. mit dem Bsp.: die Auswahl der Kliniken müsste zufällig sein.), ein einzelner Klumpen reicht nicht aus.
- **Geschichtete Stichprobe:** Elemente werden nach relevanten Merkmalen in Schichten (Strata) vorgruppiert. Man überlegt, welche Merkmale relevant für die Untersuchung sind (z.B. Geschlecht, etc.), und wählt innerhalb dieser Merkmale zufällig die Leute aus (mit dem Klumpen-Verfahren geht es auch, ist aber statistisch unsauberer.).
Proportional geschichtete Stichprobe: entsprechend der Merkmalsverteilung in Population zusammengestellt, z.B. 40% Singles, 60% Ehepaare.
Generell führt eine nach relevanten Merkmalen geschichtete Stichprobe zu besseren Schätzwerten der Populationsparameter als eine einfache Zufallsstichprobe.

Verteilung von Stichprobenkennwerten:

Wir berechnen bei einer Stichprobe der Größe n den Mittelwert \bar{x} . Wie genau können wir nun daraus den Mittelwert der Population erkennen?

- bei zwei Stichproben aus einer Population: je weiter die Mittelwerte der Stichprobe voneinander entfernt sind, desto geringer die vermutlich richtige Schätzung des Populationsmittelwertes.
- wenn wir unendlich viele Stichproben haben, dann entsteht eine **Stichprobenkennwerteverteilung**, je geringer die Streuung (= Standardfehler $s_{\bar{x}}$) ist, desto genauer ist die Schätzung.

Die Stichprobenkennwerteverteilung (SKV):

Gedankenexperiment: Eine Population mit μ und σ sei gegeben. Aus dieser Population werden wiederholt (theoretisch unendlich oft) Stichproben derselben Größe n mit Zurücklegen gezogen und innerhalb jeder Stichprobe wird z.B. der Stichprobenkennwert \bar{x} berechnet. Die Verteilung der so gewonnenen Kennwerte ist die Stichproben(kennwerte)verteilung von \bar{x} .

Wenn man genug Stichprobenkennwerte hat, ist deren Mittelwert mit dem der Population identisch.

Der Mittelwert einer Stichprobenverteilung (SKV) eines Kennwertes T heißt Erwartungswert E(T).

Entspricht der Erwartungswert eines Kennwertes dem Wert des zugehörigen Populationsparameters, $E(T) = \theta$, so wird dieser Kennwert als erwartungstreuer Schätzer bezeichnet.

Mittelwert:

Der Mittelwert ist ein erwartungstreuer Schätzer für μ , da gilt: $E(\bar{x}) = \mu$.

Varianz:

Die Stichprobenvarianz s^2 ist kein erwartungstreuer Schätzer: $E(s^2) \neq \sigma^2$.

s^2 unterschätzt σ^2 systematisch um den Faktor: $\frac{n-1}{n}$

Daher muss s^2 wie folgt korrigiert werden (doch, der Bruch ist richtig herum*):

$$\hat{\sigma}^2 = s^2 \cdot \frac{n}{n-1}$$

Das Dach über dem Sigma zeigt an, dass es sich um den Schätzer eines Parameters handelt.

$$\hat{\sigma}^2 \text{ ist erwartungstreu: } E(\hat{\sigma}^2) = \sigma^2.$$

Die Populationsvarianz schätzt man also durch die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert (Quadratsumme) geteilt durch $n-1$ (Freiheitsgrade der Varianz):

$$\hat{s}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Jetzt ist eine Berechnung des geschätzten Standardfehlers möglich (siehe nächste Seite)

*** Einfach:** Weil s^2 unser Sigma systematisch unterschätzt, müssen wir s^2 vergrößern.

Mathematisch korrekt: s^2 unterschätzt systematisch das Sigma:

$$\hat{s}^2 \cdot \frac{n-1}{n} = s^2$$

Und nun stellt die Formel mal um.

Der Standardfehler:

Ist die Streuung einer \bar{x} -Verteilung ziemlich klein, ist die Chance, mithilfe eines Stichprobenkennwertes \bar{x} den Mittelwert der Population μ zu treffen, relativ groß. Ist jedoch die Streuung groß, ist die Chance auf einen Treffer gering. Deswegen benötigt man den Wert der Streuung der Stichprobenkennwertverteilung, und das ist der Standardfehler.

Die Streuung einer Stichprobenkennwertverteilung wird als **Standardfehler des Mittelwertes** (also des Kennwertes) bezeichnet.

Der Standardfehler des Mittelwertes hängt von der Populationsvarianz σ^2 und der Stichprobengröße n ab:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$$

- Standardfehler verändert sich proportional zur Streuung in Population.
- Standardfehler verringert sich mit zunehmendem Stichprobenumfang.

Schätzung des Standardfehlers des Mittelwertes:

In der Regel ist die Varianz einer Population σ^2 nicht bekannt. Daher wird der Standardfehler aus den Stichprobendaten geschätzt: (eine Schätzung über den Durchschnitt der Stichprobenvarianzen wäre ja nicht erwartungstreu.)

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{s^2}{n-1}}$$

Zentrales Grenzwerttheorem:

Der zentrale Grenzwertsatz sagt nun etwas über die Form aus:

Verteilungen von Mittelwerten aus Stichproben des Umfangs n , die sämtlich derselben Grundgesamtheit entnommen werden, **gehen** mit wachsendem Stichprobenumfang **in eine Normalverteilung über**.

Vereinbarungsgemäß hält man ein $n \geq 30$ für hinreichend, um die Stichprobenverteilung des Mittelwertes als Normalverteilung zu behandeln, ungeachtet der Verteilungsform der zugrundeliegenden Population.

Ein Beispiel: man werfe zwei sechsstellige Würfel, und lese deren Ergebnis ab. Die 7 kommt z.B. häufiger vor als die 2, da es bei der 7 mehrere Kombinationsmöglichkeiten gibt (Bortz S. 93). Es ergibt sich eine in Richtung NV bewegende Verteilung:

Ergebnis:	2	3	4	5	6	7	8	9	10	11	12
p(x):	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Entsprechendes gilt für **JEDE** beliebige Verteilungsform (**bei Mittelwerten**).

Stichprobenkennwerte müssen nicht nur Mittelwerte sein, sondern es sind auch Mediane und Varianzen möglich. **Bei Varianzen s^2 ist die Stichprobenverteilung χ^2 -verteilt.**

Mittelwert (\bar{x}) der Stichprobenkennwerteverteilung der Mittelwerte:

- entspricht m der Population bei hinreichend großen Stichprobenumfängen.
- Hat ebenfalls bei $m \pm s_{\bar{x}}$ 68,26% und bei $m \pm 2s_{\bar{x}}$ 95,44% aller Messwerte liegen.

[wobei $s_{\bar{x}}$ = Streuung der Mittelwerte der Stichproben (= Standardfehler)]

Das bedeutet man kann: „Bei $m \pm 2s_{\bar{x}}$ liegen 95,44% aller Messwerte“, auch mathematisch formulieren, und zwar:

$$m - 2 \cdot s_{\bar{x}} \leq \bar{x} \leq m + 2 \cdot s_{\bar{x}}$$

Und schon sind wir bei einem **Wertebereich** (mehr dazu später).

Kriterien der Parameterschätzung: Erwartungstreue, Konsistenz, Effizienz, Suffizienz:

Nach welchen Kriterien können wir entscheiden, ob ein statistischer Kennwert einen brauchbaren Schätzer für einen Parameter darstellt?

Was sind die Eigenschaften, die eine gute Schätzung auszeichnen?

Erwartungstreue:

Ein statistischer Kennwert schätzt einen Populationsparameter erwartungstreu, wenn der Mittelwert der Kennwerteverteilung (**Kennwerte sind nicht nur Mittelwerte!**) bzw. deren Erwartungswert dem Populationsparameter entspricht (siehe auch „Stichprobenkennwerteverteilung“).

Konsistenz:

Von einem konsistenten Schätzwert sprechen wir, wenn sich ein statistischer Kennwert mit wachsendem Stichprobenumfang dem Parameter, den er schätzen soll, nähert.

[Demzufolge ist z.B. die Standardabweichung konsistent, aber nicht erwartungstreu; der Mittelwert hingegen konsistent und erwartungstreu.]

Effizienz:

Je größer die Varianz der SKV, desto geringer ist die Effizienz des entsprechenden Schätzwertes. [konkret: es geht hier also um die **Präzision** einer Schätzung.]

[z.B. schätzt eine Mittelwertsverteilung effizienter m als eine Medianverteilung. (= eine Verteilung aus den Medianen der Stichproben).]

Suffizienz:

Ein Schätzwert ist suffizient oder erschöpfend, wenn er alle in den Daten einer Stichprobe enthaltenen Informationen berücksichtigt, so dass durch Berechnung eines weiteren statistischen Kennwertes keine zusätzlichen Informationen über den zu schätzenden Parameter gewonnen werden kann.

[Da der Median nur ordinale Dateninformationen berücksichtigt, der Mittelwert aber intervallskalierte Dateninformationen berücksichtigt, ist der Mittelwert der erschöpfendere Schätzer.]

Betrachtet man alle vier Kriterien, sieht man, dass \bar{s}^2 und \bar{x} die besten Schätzer für s^2 und m sind. Sie sind erwartungstreu, konsistent und erschöpfend. Bei NVen sind sie auch am Effizientesten.

Methoden der Parameterschätzung: Meth. d. kleinsten Quadrate, Maximum-likelihood-Methode:

Wie finde ich den besten Schätzer für eine Population?

Methode der kleinsten Quadrate:

Die Methode der kleinsten Quadrate wird noch im Punkt „Verfahren zur Überprüfung von Zusammenhangshypothesen“ thematisiert, trotzdem eine kurze Beschreibung:

Nehmen wir an, wir suchen den besten Schätzer für μ . (Wir wissen natürlich, dass das \bar{x} ist.) Den gesuchten Wert nennen wir a .

$$f(a) = \sum_{i=1}^n (x_i - a)^2 = \min.$$

Wir lösen das Ganze auf (Rechnung im Bortz S.97 & 98), und erhalten:

$$a = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

Demnach ist \bar{x} der beste Schätzer für μ .

Maximum-likelihood-Methode:

Mit der Maximum-likelihood-Methode finden wir für die Schätzung unbekannter Parameter Stichprobenkennwerte, die so geartet sind, dass sie die Wahrscheinlichkeit (engl.: likelihood) des Auftretens in der Population maximieren.

Messen wir 5x, und zwar

$x_1 = 11$, $x_2 = 8$, $x_3 = 12$, $x_4 = 9$ und $x_5 = 10$

Wenn wir jetzt eine NV haben, ist es unwahrscheinlich, dass die Population einen Mittelwert von $\mu = 20$ hat.

Nach der Maximum-likelihood-Methode würde sich herausstellen, dass $\bar{x} = 10$ der beste Schätzer für μ ist.

[Eingehendere Beschreibung im Bortz auf den S. 98 & 99]

Intervallschätzung: Wertebereich, Konfidenzintervalle:

Bei einem \bar{x} -Wertebereich versucht man herauszufinden, ob eine Stichprobe (d.h. eigentlich der Mittelwert \bar{x} einer Stichprobe) zu einer Population gehört.

D.h. bei einer NV, also $n \geq 30$, liegen Werte zu bestimmten Bedingungen um den Mittelwert m herum.

Bei $m \pm S_{\bar{x}}$ liegen 68,26% und

bei $m \pm 2S_{\bar{x}}$ 95,44% aller zufällig gezogenen Stichprobenmittelwerte in dem Bereich ± 2 Streuungen um den Mittelwert m .

D.h. bei $m = 100$ und $S_{\bar{x}} = 5$, läge bei ± 2 Streuungen zu 95,44 % die gezogene Stichprobe zwischen 90 und 110. [eigentlich ist das falsch, ein Wert kann nur drin oder draußen liegen, also ist die Wahrscheinlichkeit entweder 1 oder 0 – ist aber nicht so wichtig (= Mathematiker-Pedanterie)].

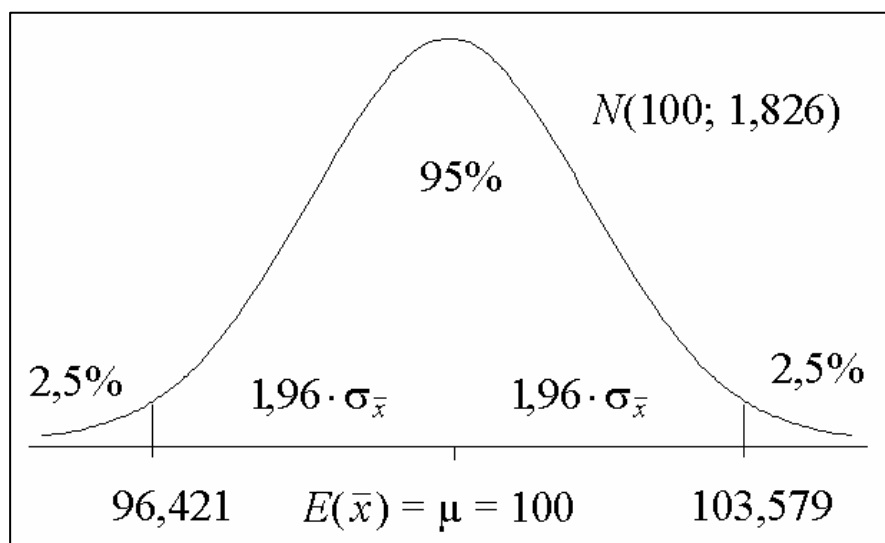
Diesen Bereich – hier 90 bis 110 – nennt man \bar{x} -Wertebereich.

m und $S_{\bar{x}}$ sind hierbei festgelegte Werte. Das Einzige, was variieren kann, ist \bar{x} . (Weswegen man es ja auch \bar{x} -Wertebereich nennt.)

Da man aber lieber mit standardisierten NVen rechnet, arbeitet man meist mit z -Werten. Die Gleichung sieht folgendermaßen aus:

$$m + z_{(\alpha/2)} \cdot S_{\bar{x}} \leq \bar{x} \leq m + z_{(1-\alpha/2)} \cdot S_{\bar{x}}$$

- $(\alpha/2)$ haben wir hier, da ein zweiseitiges Intervall gesucht wird, also z.B. auf jeder Seite 2,5% abgeschnitten wird.



In der Standardnormalverteilung liegen 95% der Fläche zwischen $z = \pm 1,96$.

Konfidenzintervalle:

Konfidenzintervalle verhalten sich genau umgekehrt zum \bar{x} - Wertebereich:

Bei einem Konfidenzintervall versucht man herauszufinden, ob eine Population zu einer Stichprobe gehört (die möglichen m - Werte werden gesucht); und nicht wie beim Wertebereich, ob eine Stichprobe zu einer Population gehört (die möglichen \bar{x} - Werte werden gesucht).

- Konfidenzintervalle sind demnach Bereiche, in denen sich Populationsparameter befinden, die als Erzeuger einer Stichprobe (mit xx% Wahrscheinlichkeit) in Frage kommen.
- Man nimmt hier 95%ige oder 99%ige Wahrscheinlichkeiten (= **Konfidenzkoeffizienten**).

Das Konfidenzintervall kennzeichnet denjenigen Bereich eines Merkmals, in dem sich 95% oder 99% (legt man vorher fest) aller möglichen Populationsparameter befinden, die den empirisch ermittelten Stichprobenkennwert \bar{x} erzeugt haben können.

$$\bar{x} + z_{(a/2)} \cdot s_{\bar{x}} \leq m \leq \bar{x} + z_{(1-a/2)} \cdot s_{\bar{x}}$$

Gesucht sind nun hierbei die Grenzen des Konfidenzintervalls.

- Ein Konfidenzintervall Δ_{crit} wird bestimmt nach $\Delta_{\text{crit}} = m + z_{(a/2)} \cdot s_{\bar{x}}$

Ein Delta steht immer für eine Differenz, d.h. die (Unter-) Grenze des Konfidenzintervalls hat den Abstand Delta entweder zum Mittelwert oder zur anderen (Ober-)Grenze.

- bei kleinen Stichproben (unter 30) sind die „z-Werte“ nicht standardnormalverteilt, sondern t-verteilt (wenn die Population normalverteilt ist), also muss man t-Werte aus der Tabelle einsetzen; Freiheitsgrade = $n - 1$ (siehe „t-Test“)

Bedeutung des Stichprobenumfangs:

Eine (weitere) Bestimmungsgröße für die Breite eines Konfidenzintervalls ist der Stichprobenumfang: **Je größer die untersuchte Stichprobe, desto kleiner das Konfidenzintervall.**

- Die Halbierung eines Konfidenzintervalls macht einen vierfachen Stichprobenumfang erforderlich.
- Die benötigten Stichprobenumfänge können erheblich gesenkt werden, wenn statt einer reinen Zufallsstichprobe eine sinnvoll geschichtete Stichprobe gezogen wird (=> Standardfehler wird kleiner => kleinere Stichprobenumfänge möglich.)

Formulierung und Überprüfung von Hypothesen:

Bei der Formulierung und Überprüfung von Hypothesen entsteht ein zentrales Problem:

Inwieweit können postulierte Eigenschaften (= Hypothesen) der Population (Theorie) durch Stichproben (Empirie) bestätigt werden?

Welche Kriterien gibt es dort? Bis wann ist ein Stichprobenwert (z.B. \bar{x} , s^2 , etc.) „gerade noch mit der Theorie übereinstimmend“?

Um diese u.ä. Fragen geht es in diesem Kapitel.

Alternativhypothesen:

Die Brauchbarkeit von Theorien ist davon abhängig, inwieweit sie sich in Untersuchungen von Teilaussagen, die aus ihr abgeleitet wurden, bewährt.

Ist eine Theorie (noch) nicht empirisch bewiesen worden, stellt man **Hypothesen** auf, um präzisere Vorhersagen machen zu können. Eine Hypothese ist also eine Deduktion der Theorie (Deduktion = Ableitung des Einzelfalls aus dem Allgemeinen). [Deswegen soll man Hypothesen ja auch vorher aufstellen, sonst leitet man nämlich aus dem Einzelfall bzw. den Einzelfällen das Allgemeine ab.]

Hypothesen beinhalten Aussagen, die mit anderen Theorien in Widerspruch stehen können oder den Wissensstand erweitern.

Hypothesen, die in diesem Sinne „innovative“ Aussagen beinhalten, werden als Gegen- oder Alternativhypothesen bezeichnet.

Diese Hypothesen müssen dann (nur noch) überprüft werden.

Varianten von Alternativhypothesen:

Je nach Art der Hypothesenformulierung unterscheidet man zwischen

- Unterschiedshypothesen und
- Zusammenhangshypothesen.

Unterschiedshypothesen prüft man bei *Häufigkeitsvergleichen* und *Mittelwertsvergleichen* (siehe „Verfahren zur Überprüfung von Unterschiedshypothesen“) und

Zusammenhangshypothesen bei *Korrelationsrechnung*

(siehe „Verfahren zur Überprüfung von Zusammenhangshypothesen“)

Gerichtete und ungerichtete Hypothesen:

Gerichtete Hypothesen haben irgendeine Richtungsvorgabe, d.h. es sind Formulierungen z.B. „größer, besser, schneller, schlechter, wandelbarer, etc. als“ vorhanden.

Ungerichtete Hypothesen haben keine Richtungsvorgabe, sie postulieren nur irgend einen Unterschied. Typische Formulierungen wären: „ist ungleich, verschieden, anders, etc. zu“

Spezifische und unspezifische Hypothesen:

Wenn wir eine bestimmte Größe bei einer Hypothese angeben, haben wir eine **spezifische Hypothese**. Z.B. „Es ist (mindestens) um den Betrag x besser als vorher.“ (= Unterschiedshypothese) oder „Der Zusammenhang (=Korrelation) beträgt (mindestens) x .“ (=Zusammenhangshypothese). **Unspezifische Hypothesen** haben demnach keine expliziten Angaben.

- Spezifische Hypothesen kommen in der psychologische Forschung meistens nur in Verbindung mit gerichteten Hypothesen vor.

Die Alternativhypothese sollte – soweit sich das inhaltlich rechtfertigen lässt – so präzise wie möglich formuliert sein. Die wenigsten Vorkenntnisse verlangt eine

unspezifische ungerichtete Hypothese	gefolgt von einer
unspezifischen gerichteten Hypothese	gefolgt von einer
spezifischen gerichteten Hypothese	

Wie das bei einer spezifischen ungerichteten Hypothese (Es unterscheidet sich mindestens um x .) aussieht, war nicht erwähnt (ist wohl auch eher selten.).

Statistische Hypothesen:

- Zur Überprüfung müssen wissenschaftliche Hypothesen **vor** einer Untersuchung aufgestellt und in eine statistische Hypothese überführt werden.
- Hypothesen werden für Populationen, also mit Parametern formuliert.
- Diese beiden Punkte gelten auch für die Nullhypothese.

Alternativhypothesen heißen üblicherweise H_1 .

Bei einer Unterschiedshypothese (hier Mittelwert) heiße sie z.B.

„Psychologiestudenten sind mit diesem neuen Skript (m_1) besser in der Methodenlehreprüfung als vorher (m_0).“ Mathematisch formuliert:

$$H_1 : m_0 < m_1$$

Bei einer Zusammenhangshypothese heiße sie z.B.

„Psychologiestudenten werden eher wahnsinnig (r_1) (korrelieren mehr mit Irrenhausbesuchen) als andere Studenten (r_0).“ Mathematisch formuliert:

$$H_1 : r_0 < r_1$$

Die Nullhypothese:

In Abhängigkeit zur Alternativhypothese wird eine konkurrierende Hypothese gebildet, die sogenannte **Nullhypothese (H_0)**. Die **Nullhypothese** ist eine Negativhypothese, die **behauptet, dass die zur H_1 komplementäre Aussage richtig sei. Die Nullhypothese ist die Basis der klassischen Prüfstatistik, da von dort aus entschieden wird, ob die H_1 stimmt oder nicht.**

(Mehr dazu im Bortz S. 109 & 110)

Fehlerarten bei statistischen Entscheidungen: Alpha- & Beta-Fehler:

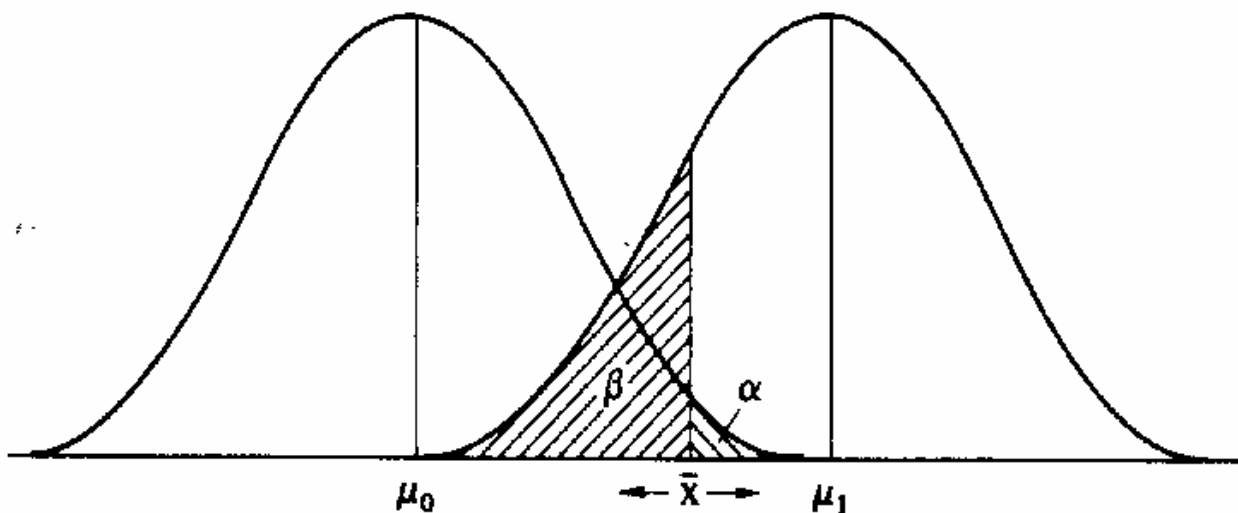
Stehen Null- und Alternativhypothese fest, kann die Untersuchung beginnen. Aber welche der Hypothesen gilt nun? **Die Entscheidung hierüber wird noch erschwert, da sich das Ergebnis der Untersuchung auf eine Stichprobe bezieht, während wir eigentlich Aussagen über eine Population treffen wollen.**

Damit ist nicht auszuschließen, dass wir aufgrund der Stichprobenauswahl Fehler machen: Insgesamt gibt es 4 Möglichkeiten:

		In der Population („in Wahrheit“) gilt ...	
		H_0	H_1
Entscheidung für ...	H_0	Richtige Entscheidung	β -Fehler
	H_1	α -Fehler	Richtige Entscheidung

a -Fehler oder Fehler erster Art: Die H_1 wird angenommen, obwohl eigentlich die H_0 gilt.

b -Fehler oder Fehler zweiter Art: Die H_0 wird beibehalten, obwohl eigentlich die H_1 gilt.



α - und β -Fehlerwahrscheinlichkeit bei einer (rechts-) gerichteten Hypothese.

Signifikanzaussagen:

Signifikanzaussagen sind Aussagen, ob eine Hypothese – meist die H_1 – auf der Basis festgelegter Grenzen (meist bei 5% oder 1% Irrtumswahrscheinlichkeit) gilt oder nicht.

Da Stichprobenergebnisse (zufällig) stark von der Population abweichen können, setzt man Signifikanzen fest, um zu einer bestimmten Wahrscheinlichkeit eine Hypothese (Aussage) beweisen zu können.

Man bezeichnet ein **Signifikanzniveau von $\alpha = 5\%$ als signifikant**, ein

Signifikanzniveau von $\alpha = 1\%$ als sehr signifikant. (Das ist so festgelegt worden.)

Inwieweit man α festlegt, ist eine methodische Sache (Brisanz der Untersuchung => kleines α , junge Wissenschaft mit innovativen Ideen => großes α)

Da die H_0 die Basis der klassischen Prüfstatistik ist und wir die H_1 beweisen wollen, ist für uns (erst einmal) der α -Fehler von Bedeutung. Deswegen wird erst einmal die Wahrscheinlichkeit des α -Fehlers – die sogenannte **Irrtumswahrscheinlichkeit** – bestimmt.

Bestimmung der Irrtumswahrscheinlichkeit:

Irrtumswahrscheinlichkeiten sind bedingte Wahrscheinlichkeiten, d.h. Wahrscheinlichkeiten für das Auftreten eines Ereignisses (hier das Ergebnis unserer Stichprobenuntersuchung) unter der Bedingung, dass die Nullhypothese zutrifft.

Aus der Population ziehen wir (theoretisch) unendlich viele Stichproben – es ergibt sich eine SKV mit dem Mittelwert μ_0 , der dem der Population entspricht.

Bei dieser SKV trennen wir (bei rechtsseitiger H_1 & $\alpha = 0,05$) rechts 5% ab.

Diese Teilfläche stellt den Bereich dar, bei der wir uns zu α % (hier 5%) irrtümlicherweise zugunsten der H_1 entscheiden würden. Die Größe der Irrtumswahrscheinlichkeit wird mit P gekennzeichnet.

Mittels einer z -Transformation lässt sich dann der genaue (Grenz-) Wert errechnen.

Nun ändern wir bei einer Stichprobe die Bedingungen, damit unsere H_1 wirkt (z.B. H_1 : neue Lehrmethode => die wird nun angewandt). Landet der Mittelwert der Stichprobe im Bereich P , nehmen wir die H_1 an.

Das Ganze war die Theorie.

Da aber die Kosten ins Astronomische steigen würden (unendlich viele Stichproben), gehen wir von einer NV der Mittelwerte aus (die SKV soll also eine NV sein).

Dann berechnen wir eine z -Transformation mit dem Wert unserer – unter H_1 -Bedingungen getesteten – Stichprobe. Diesen z -Wert schlagen wir in der Tabelle nach und sehen, wie signifikant er ist. Ist er geringer als unsere Signifikanzniveau α , können wir die H_1 annehmen.

Beispiel: Nach der z -Transformation unser Experimentalergebnisse

$$z = \frac{\bar{x} - m_0}{s_{\bar{x}}} \quad [s_{\bar{x}} \text{ kann auch geschätzt sein.}]$$

erhalten wir den z -Wert 2,29.

Dieser Wert schneidet links von sich 98,90% aller z -Werte ab. D.h. bei unserem rechtsseitigem Test und einem Signifikanzniveau von $\alpha = 5\%$ könnten wir die H_1 annehmen, bei einem Signifikanzniveau von $\alpha = 1\%$ müsste die H_0 beibehalten werden (obwohl sie nur zu 1,1% stimmen würde).

- **Ein nicht-signifikantes Ergebnis heißt nicht, dass die H_0 richtig ist;** z.B. bei einer Verfehlung des 5%-Niveaus um 1% ist die H_0 nur mit einer Wahrscheinlichkeit von 6% richtig !!
Deswegen sagt man auch „Beibehaltung“ statt „Annahme“ der H_0 .

Die Irrtumswahrscheinlichkeit α ist immer vorhanden, sie verringert sich bei

- 1.) größer werdender Diskrepanz von $\bar{x} - m_0$ (je ausgefallener die Stichprobe)
- 2.) Vergrößerung des Stichprobenumfangs: nur wenn n so groß wie die Population ist, ist die Entscheidung für H_1 oder H_0 sicher (aber da alles bekannt ist fällt sie so-undso weg).
- 3.) kleiner werdender Populationsstreuung s bzw. \hat{s} (je weniger sich die Leute unterscheiden.)

Die statistische Hypothesentestung führt zu keinen Wahrheiten, nur zu Wahrscheinlichkeiten.

Unspezifische Nullhypothesen:

Meist haben wir eine spezifische H_0 , also $m_1 = m_0$. Es kann aber auch unspezifische Nullhypothesen geben, also bei einem rechtsseitigem Test $H_0: m_1 > m_0$. Hat man jedoch festgestellt, dass $m_1 = m_0$ verworfen werden kann, so gilt das auch für $m_1 > m_0$.

Es genügt also, wenn eine unspezifische H_1 an einer spezifischen H_0 getestet wird.

Einseitige und zweiseitige Tests:

Zweiseitige Tests sind genauer als Einseitige, da das α -Niveau konstant bei z.B. 1% liegt (es wird also 1% der Verteilung abgeschnitten), und eine zweiseitige Testung durchgeführt wird (es werden also auf beiden Seiten $\alpha/2$ % abgeschnitten).

Zweiseitige Tests haben den Nachteil, dass sie keine Richtungen, sondern nur Ungleichheiten formulieren.

Einseitige Test werden eher signifikant als Zweiseitige – was in der psychologischen Forschung ja auch gerne gesehen wird.

Kann nicht klar entschieden werden, ob ein Sachverhalt besser durch eine gerichtete (einseitig) oder ungerichtete (zweiseitig) Hypothese erfasst wird, muss in jeden Fall zweiseitig getestet werden.

Statistische Signifikanz und praktische Bedeutsamkeit:

Ist n hinreichend groß und also $\hat{\sigma}_{\bar{x}}$ entsprechend klein, wird jeder noch so kleine hypothesenkonforme Unterschied signifikant.

Ein Beispiel: eine neue Lehrmethode führte bei 100 Schülern zu durchschnittlich 42 Punkten im Abschlusstest. Der Mittelwert aller Tests vor der neuen Lehrmethode betrug 40 Punkte, die Streuung 0,8 Punkte. Die H_1 (die neue Lehrmethode ist besser) wäre hierbei auf einem 1%igen Alphaniveau signifikant.

Wenn wir jetzt statt 100 Schülern 10.000 untersuchen würden, wäre selbst ein Unterschied von 0,19 Punkten bei einem 1%igen Alphaniveau signifikant
Aber hat das noch irgendeine Bedeutung?

Umgekehrt gilt auch, dass große (und bedeutsame) Unterschiede bei zu kleinen Stichproben nicht signifikant werden, also „unentdeckt“ bleiben.

Deswegen benötigt man Verfahren, die praktische Bedeutsamkeit von Untersuchungen erkennen zu können. Das geht über die Feststellung der:

- Effektgrößen und
- Teststärke

Die korrekte Anwendung eines Signifikanztests und die Interpretation der Ergebnisse unter dem Blickwinkel der praktischen Bedeutsamkeit sind essentielle und gleichwertige Bestandteile der empirischen Hypothesenprüfung.

Effektgröße:

Der in der H_1 behauptete Unterschied wird als Effekt bezeichnet. Um Effekte verschiedener Untersuchungen vergleichbar zu machen, werden Effekte zu Effektgrößen standardisiert:

$$\varepsilon = \frac{\mu_1 - \mu_0}{\hat{\sigma}}$$

Das Epsilon (ε) steht für Effektgröße.

Die Effektgröße gibt also an, inwieweit – in einem speziellen Test – μ_1 von μ_0 entfernt sein sollte, um von einem praktisch bedeutsamen Effekt sprechen zu können.

Wenn man vor einer Untersuchung eine Effektgröße festlegt, hat dies den Vorteil (neben dem größeren Aufwand), dass der Stichprobenumfang kalkulierbar ist:

$$n = \frac{(z_{(1-\alpha)} - z_{\beta})^2}{\varepsilon^2}.$$

Um eine geeignete Teststärke zu erzielen, sollte vor einer Untersuchung eine Effektgröße festgelegt und n darauf abgestimmt werden.

Der Beta-Fehler:

Die Festlegung des β -Fehlers erfolgt wie die Festlegung des α -Fehlers (zu finden unter dem Punkt „Bestimmung der Irrtumswahrscheinlichkeit“), nur dass das Ganze mit vertauschten Hypothesen stattfindet.

D.h. eigentlich sollte man unendlich viele H_1 -Stichproben (= unter der Bedingung H_1) durchführen, man geht aber von einer NV aus und macht einen z -Test.

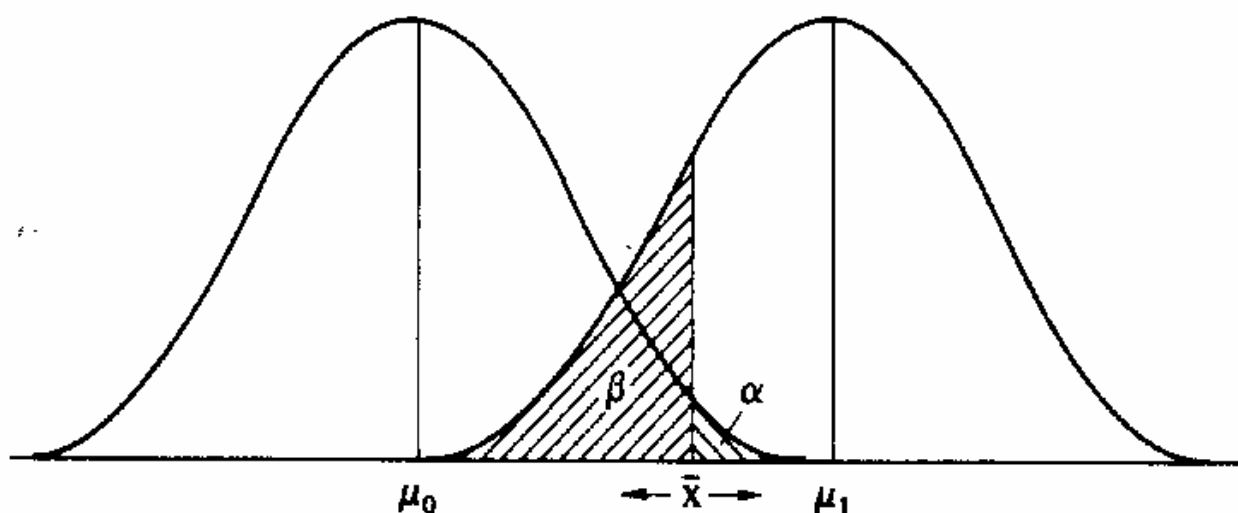
(z.B. bei $H_1: m \geq 110$ wäre $m = 110$)

Der Beta-Fehler stellt den Bereich dar, bei der wir uns zu β % irrtümlicherweise zugunsten der H_0 entscheiden würden.

Unspezifische Alternativhypothesen:

Die Festlegung des β -Fehlers, die mit einer Entscheidung zugunsten der H_0 verbunden ist, kann bei unspezifischen H_1 nicht bestimmt werden.

Hier das Bild einer spezifischen H_1 samt Alpha- & Beta-Fehler-Bereich:



Teils ist der Beta-Fehler durch die Wahl des Alpha-Fehlers festgelegt, d.h. er fängt an der kritischen Grenze (hier links davon) an, die durch die Wahl des Alpha-Niveaus entstanden ist. Der β -Fehler kann aber auch festgelegt sein. So können Indifferenzbereiche entstehen.

Indifferenzbereiche:

Wenn Alpha- und Beta-Niveau festgelegt sind, kann es bei sehr kleinen bzw. sehr großen Effektgrößen (entweder große bzw. kleine Streuung und/oder nahe beieinander bzw. weit entfernt liegenden Mittelwerte von H_1 und H_0) zu Indifferenzbereichen kommen.

Das Stichprobenergebnis befindet sich dann in einem Bereich, für den

- weder die H_0 noch die H_1 abgelehnt werden können (Überschneidung der Kurven) oder
- sowohl die H_0 als auch die H_1 abgelehnt werden müssen (kaum Kurvenüberschneidung, d.h. nur noch die ganz ganz flachen Bereiche der Kurven von H_1 und H_0 überschneiden sich).

Teststärke („power“):

Die Teststärke ist die der korrekten Entscheidung für die H_1 zugeordnete Wahrscheinlichkeit, d.h. **die Teststärke gibt an, mit welcher Wahrscheinlichkeit ein Signifikanztest zugunsten einer spezifischen H_1 entscheidet.**

Die Teststärke hat den Wert $1 - b$

Da sich **a** und **b** gegenseitig beeinflussen, ist die Teststärke für **a** = 0,05 größer als für **a** = 0,01.

Eigenschaften der Teststärke:

Sie vergrößert sich:

- mit wachsender Differenz von m_0 und m_1 (also auch mit wachsender Effektgröße)
- bei kleiner werdender Merkmals (=Populations-) streuung.
- mit größer werdendem **a** (sofern **b** nicht festgelegt ist).
- mit wachsendem Stichprobenumfang (da der Standardfehler kleiner ist) $\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$
- Bei gerichteter Hypothese hat ein einseitiger Test größere Power als zweiseitiger Test.

Bedeutung der Stichprobengröße:

Im Allgemeinen kann man sagen, dass große Stichproben immer günstiger sind als kleine, da sie zu genaueren Ergebnissen führen. Da große Stichproben auch mehr kosten, reicht ein „optimaler“ Stichprobenumfang.

Stichprobenumfänge sind dann optimal, wenn sie bei gegebenem **a, **b** und **e** eine eindeutige Entscheidung über die Gültigkeit von H_0 oder H_1 sicherstellen.**

Für Stichprobenumfänge, die kleiner sind als der „optimale“, existiert ein \bar{x} -Wertebereich, der sowohl mit H_0 als auch mit H_1 vereinbar ist. Für größere Stichproben gibt es Werte, die weder mit H_0 noch mit H_1 vereinbar sind.

Monte-Carlo-Studien, Bootstrap-Technik:

Es gibt statistische Kennwerte, deren mathematischer Aufbau so kompliziert ist, dass es bis heute nicht gelungen ist, deren Standardfehler auf analytischem Wege zu entwickeln.

In diesen Fällen können sogenannte Monte-Carlo-Studien oder die Bootstrap-Technik eingesetzt werden, mit denen die unbekannte H_0 -Verteilung des jeweiligen Kennwertes auf einem PC simuliert wird.

Die Monte-Carlo-Methode hat zwei (für uns) vorrangige Anwendungsvarianten:

- die Erzeugung der H_0 -Verteilung eines statistischen Kennwertes
(man simuliert per PC eine Population, und der PC zieht immer weiter Stichproben, er simuliert eine SKV. Mit dieser künstlichen SKV kann man überprüfen, ob der reale untersuchte Kennwert signifikant ist oder nicht.)
- die Überprüfung der Folgen, die mit der Verletzung von Voraussetzungen eines statistischen Tests verbunden sind.
(da in der empirischen Forschung oftmals Testvoraussetzungen verletzt werden, kann man einen Test auf seine „Robustheit“ testen, d.h. beeinflusst die Verletzung der Testvoraussetzungen das Ergebnis des Tests oder nicht.)

Monte-Carlo-Studien sind für die empirische Forschung äußerst wichtig, da sie –teils– die Entscheidung darüber erleichtern, wann ein bestimmter Test eingesetzt werden kann oder nicht.

Die Bootstrap-Technik:

Die Bootstrap-Technik ist den Monte-Carlo-Studien sehr ähnlich; sie unterscheidet sich von der Monte-Carlo-Studie in einem wesentlichen Punkt:

Die Monte-Carlo-Studien kommen zu generalisierbaren Ergebnissen, die Ergebnisse der Bootstrap-Techniken beziehen sich immer nur auf eine bestimmte Untersuchung.

Die Bootstrap-Technik simuliert keine Population, sie verwendet ausschließlich Informationen einer empirisch untersuchten Stichprobe mit dem Ziel, eine Vorstellung über die Variabilität des zu prüfenden Stichprobenkennwertes zu gewinnen.

Die Bootstrap-Technik zieht aus den einzelnen Werten der empirischen Stichprobe immer neue Stichproben, bildet davon Mittelwert und Varianz und berechnet schließlich ein Konfidenzintervall für diese Stichproben. Daran wird dann die H_0 überprüft.

Verfahren zur Überprüfung von Unterschiedshypothesen:

Bei den Verfahren zur Überprüfung von Unterschiedshypothesen werden normalerweise zwei Stichprobenergebnisse miteinander verglichen; Vergleiche eines Stichprobenmittelwertes mit einem Populationsmittelwert gibt es aber auch (meist ob ein Stichprobenkennwert zu einer Population gehört oder nicht).

Dabei gibt es eine Menge von Testverfahren. Bevor man eines auswählt, muss noch das Skalenniveau festgelegt werden; kann man sich nicht zwischen zweien entscheiden, sollten Tests (also unterschiedliche Tests, die dasselbe messen) für beiden Skalenniveaus durchgeführt werden. Sind die Ergebnisse gleich, kann das höhere Skalenniveau nicht angezweifelt werden. Unterscheiden sich die Ergebnisse, muss das anforderungslosere Skalenniveau gewählt werden.

Verfahren für Intervalldaten:

Die folgenden Verfahren wie z-Test, t-Test, F-Test und χ^2 -Test für eine Varianz, vorgestellt in den Punkten:

- „Vergleich eines Stichprobenmittelwertes mit einem Populationsparameter“
- „Vergleich zweier Stichprobenmittelwerte aus unabhängigen Stichproben“
- „Vergleich zweier Stichprobenmittelwerte aus abhängigen Stichproben“
- „Vergleich einer Stichprobenvarianz mit einer Populationsvarianz“
- „Vergleich zweier Stichprobenvarianzen“

sind Verfahren für Intervalldaten.

Vergleich eines Stichprobenmittelwertes mit einem Populationsparameter (z-Test, t-Test):

Eine Zufallsstichprobe des Umfangs n mit einem Mittelwert \bar{x} wird berechnet.

Nun soll die Hypothese geprüft werden, ob die Stichprobe zu einer Grundgesamtheit mit bekanntem Populationsparameter m_0 gehört.

Aufstellen der Hypothesen:

$$H_0: m_0 = m$$

$$H_1: m_0 \neq m$$

Bei $n \geq 30$ und bekannter Populationsvarianz σ^2 (und somit bekanntem Standardfehler) sowie bei $n \geq 50$ und geschätzter Populationsvarianz $\hat{\sigma}^2$ (und somit geschätztem Standardfehler) kann man einen **z-Test** durchführen:

Also entweder

$$z_{emp} = \frac{\bar{x} - m}{\sigma_{\bar{x}}} \quad \text{oder} \quad z_{emp} = \frac{\bar{x} - m}{\hat{\sigma}_{\bar{x}}}.$$

Man berechnet z_{emp} und z_{krit} . Wenn $z_{\text{emp}} > z_{\text{krit}}$, nehmen wir (bei rechts- und ungerichteter H_1 , also auch hier) die H_1 an.

Zu beachten ist, dass sich das α bei z_{krit} bei zweiseitigen Tests halbiert.

Kleine Stichproben:

Erfüllen unsere Stichprobe und die Population die oben genannten Kriterien für den z-Test nicht, kann man auf einen **t-Test** ausweichen.

Die Hypothesenaufstellung und das Verfahren erfolgen wie beim z-Test, nur halt jetzt mit t-Werten statt z-Werten.

Nicht zu vergessen:

Werden Stichproben des Umfangs n aus einer normalverteilten Population (Grundgesamtheit) gezogen, verteilen sie sich entsprechend einer t-Verteilung mit $n-1$ Freiheitsgraden ($df = \text{degrees of freedom}$).

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}, \text{ mit } df = n-1.$$

Jetzt werden t_{emp} und t_{krit} berechnet, nur abreitet man mit der t-Tabelle statt der z-Tabelle, der Rest läuft wie beim z-Test.

Zur Erinnerung: je nach t-Test (links- / rechtsseitig / ungerichtet) ist:

$$t_{\text{krit}} = t_{(df; \alpha)} \quad \text{oder} \quad t_{\text{krit}} = t_{(df; 1-\alpha)} \quad \text{oder} \quad t_{\text{krit}} = t_{(df; \alpha/2)}$$

Anzahl der Freiheitsgrade:

Freiheitsgrad = Die Anzahl der bei der Berechnung des Kennwertes frei variierbaren Werte.

Die Anzahl der Freiheitsgrade hängt davon ab, wie viel Werte noch frei variieren können. D.h., wenn 4 von 5 Werten bei einer Additionsaufgabe und das Ergebnis feststehen, dann ist der Wert des 5. Wertes festgelegt, ergo unfrei.

Beispiel: die 4 Werte ergeben aufaddiert 9. Das Endergebnis ist 10. Welche Zahl hat wohl der 5. Wert?

Vergleich zweier Stichprobenmittelwerte aus unabhängigen Stichproben (t-Test):

Sind bei Gruppenvergleichen die Daten beider Populationen unbekannt, vergleicht man anhand von Stichproben.

Der t-Test für unabhängige Stichproben prüft zwei voneinander unabhängige Stichproben des Umfangs n_1 und n_2 , die aus zwei verschiedenen Grundgesamtheiten gezogen wurden.

Die Hypothesen lauten:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0 \text{ (ungerichtet)}$$

[Nullhypothese ausformuliert: Wenn sich beide Populationen nicht unterscheiden, müsste bei (unendlich) vielen Stichproben aus beiden Populationen auch kein Unterschied zwischen den Mittelwerten der beiden SKVen bestehen.]

Die eigentliche Formel lautet (ist eine Art z-Transformation für zwei Stichproben):

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)}} \quad df = n_1 + n_2 - 2$$

Denn im Nenner steht erneut nur der Standardfehler, nur diesmal für die Differenz aus beiden Stichproben.

Und dieser Standardfehler berechnet sich aus dieser einfachen formschönen Formel:

$$\begin{aligned} \hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)} &= \sqrt{\frac{(n_1 - 1) \cdot \hat{\sigma}_1^2 + (n_2 - 1) \cdot \hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

(Beide Versionen – also die obere und untere – sind mögliche Abwandlungen. Die obere Version kann man anwenden, wenn beide Populationsvarianzen schon geschätzt wurden.)

Voraussetzungen (für einen t-Test für unabhängige Stichproben):

- Die Verteilungen der Stichprobenmittelwerte sind NV, bei kleineren Stichproben müssen die Populationen (goodness-of-fit-Test) normalverteilt sein.
- homogene Varianzen
- Unabhängigkeit der Stichproben
- Gleiche Stichprobengrößen (im Allgemeinen)

Noch etwas: ab $n_1 + n_2 \geq 50$ kann man auch in der z-Tabelle statt in der t-Tabelle nachsehen.

Vergleich zweier Stichprobenmittelwerte aus abhängigen Stichproben (t-Test):

Abhängig sind Stichproben, deren Elemente paarweise einander zugeordnet sind. Dies geschieht durch Messwiederholung oder Zuordnung (Parallelisierung) nach sinnvollen Kriterien:

Beispiele für „paarweise“ wären:

- Intelligenz wird bei denselben Probanden zweimal erhoben und es soll untersucht werden, ob sich der Intelligenztestwert von Messung zu Messung verändert hat (**Messwiederholung**).
- Intelligenz wird bei Zwillingen gemessen.
- Stichprobengruppen werden **parallelisiert** (matched samples)

Beim t-Test für abhängige Stichproben wird berücksichtigt, dass die Varianz des ersten Messzeitpunktes die des zweiten beeinflusst (oder umgekehrt oder beides).

Dadurch können Mehrfachmessungen (z.B. durch Vorwissen bei einem Wissenstest) vermieden werden.

Voraussetzungen (für einen t-Test für abhängige Stichproben)

- Die Stichprobenverteilung der mittleren Differenzen ist eine NV (uns interessiert also nicht die Verteilung von Mittelwerten im möglichst vielen Stichproben, sondern die Verteilung der Differenzen der Mittelwerte.).
- Bei kleineren Stichproben müssen die Populationen normalverteilt sein.

Die Hypothesen lauten:

ungerichtet	rechtsseitig	linksseitig
$H_0: \mathbf{m}_d = 0$	$H_0: \mathbf{m}_d \leq 0$	$H_0: \mathbf{m}_d \geq 0$
$H_1: \mathbf{m}_d \neq 0$	$H_1: \mathbf{m}_d > 0$	$H_1: \mathbf{m}_d < 0$

Vorgehen:

Zunächst werden die Differenzen zwischen den Messwerten der ersten und der zweiten Erhebung berechnet:

$$d_i = x_{i1} - x_{i2}.$$

Dann wird der Mittelwert der Differenzen berechnet:

$$\bar{x}_d = \frac{\sum_{i=1}^n d_i}{n}, \text{ mit } n: \text{Anzahl der Messwertpaare.}$$

Dann der geschätzte Standardfehler ermittelt:

$$\hat{\sigma}_{\bar{x}_d} = \sqrt{\frac{\hat{\sigma}_d^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{x}_d)^2}{n-1} \cdot \frac{1}{n}} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{x}_d)^2}{n \cdot (n-1)}}$$

Die eigentliche Formel lautet (ist eine Art z-Transformation ohne μ , da $\mathbf{m}_d = 0$):

$$t_{emp} = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}}, \text{ mit } df = n-1 \text{ und dem Standardfehler}$$

Statistische Entscheidung (für alle t-Tests):

	ungerichtet	rechtsseitig	linksseitig
Annahme der H_1 , falls ...	$ t_{emp} > t_{(df; \alpha/2)} $	$t_{emp} > t_{(df; 1-\alpha)}$	$t_{emp} < t_{(df; \alpha)}$

Bei einem gerichteten t-Test für abhängige Stichproben ist darauf zu achten, dass die inhaltliche und die statistische Alternativhypothese dieselbe Richtung behaupten; das hängt davon ab, wie man die Differenzen bildet (der Fehler passiert bei Aufgabenblatt 10 in B2).

Deswegen bei $d_i = x_{i1} - x_{i2}$ immer die H_1 -Werte bei x_i einsetzen (dann behaupten die inhaltliche und die statistische Alternativhypothese immer dieselbe Richtung.).

Aus Monte-Carlo-Studien geht hervor, dass der t-Test für abhängige Stichproben wie auch der t-Test für unabhängige Stichproben äußerst robust auf Verletzungen seiner Voraussetzungen reagiert.

Bei einem t-Test für unabhängige Stichproben geht das bei

- ähnlichen Grundgesamtheiten, möglichst eingipflig und symmetrisch
- großen Stichprobenunterschieden, solange die Varianzen gleich sind.

Bei einem t-Test für abhängige Stichproben sollte

- die Korrelation positiv sein.

Vergleich einer Stichprobenvarianz mit einer Populationsvarianz (χ^2 -Test für eine Varianz):

Manchmal kann die Varianz einer Stichprobe stark von der der Population abweichen, der Mittelwert z.B. aber identisch sein. U.a. aufgrund dessen kann es interessant sein, Varianzvergleiche durchzuführen.

Der folgende Test – ein **χ^2 -Test für eine Varianz** – überprüft, ob eine Stichprobe aus einer Population mit der Varianz $\sigma^2 = a$ stammt.

Voraussetzung:

- Sofern die Messwerte (der Population & der Stichprobe) normalverteilt sind, verteilt sich die Prüfgröße (=die eigentliche Formel) gemäß einer χ^2 -Verteilung mit $n-1$ Freiheitsgraden.

Die Nullhypothese (ist immer so):

$$H_0: \sigma^2 = a$$

Die Alternativhypothesen lauten:

	ungerichtet	rechtsseitig	linksseitig
falls ...	$\chi_{\text{emp}}^2 < \chi_{df, \alpha/2}^2$ oder $\chi_{\text{emp}}^2 > \chi_{df, 1-\alpha/2}^2$	$\chi_{\text{emp}}^2 > \chi_{df, 1-\alpha}^2$	$\chi_{\text{emp}}^2 < \chi_{df, \alpha}^2$

Vorgehen:

Standardfehler ausrechnen, Werte einsetzen nach folgender Formel.

Die eigentliche Formel lautet:

$$\chi_{\text{emp}}^2 = \frac{(n-1) \cdot \hat{\sigma}^2}{a}, \text{ mit } df = n-1.$$

Statistische Entscheidung:

	ungerichtet	rechtsseitig	linksseitig
Annahme der H_1 , falls ...	$H_1: \sigma^2 \neq a$	$H_1: \sigma^2 > a$	$H_1: \sigma^2 < a$

Vergleich zweier Stichprobenvarianzen (F-Test):

Der Vergleich von Stichprobenvarianzen mit der Populationsvarianz kommt in der Praxis nicht so häufig vor. Es werden eher zwei Stichprobenvarianzen miteinander verglichen. Dies geschieht über einen F-Test.

Voraussetzungen

- Normalverteilung der Grundgesamtheiten,
- Unabhängigkeit der Stichproben.

Die Hypothesen lauten:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \quad (\text{ungerichtet})$$

$$H_1 : \sigma_1^2 > \sigma_2^2 \quad (\text{gerichtet})$$

[Die H_0 ausformuliert: es gibt keinen Unterschied in den Populationsvarianzen.]

[Diese Hypothese wird nun anhand von Stichprobenvarianzen überprüft.]

Bei einem einseitigen Test (es kommen fast nur einseitige F-Tests vor) kommt die als größer erwartete Varianz in den Zähler, bei einem zweiseitigen Test (selten) kommt die empirisch größere Varianz in den Zähler.

Die eigentliche Formel lautet:

$$F_{\text{emp}} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \quad \text{mit } df_{\text{Zähler}} = n_1 - 1 \text{ und } df_{\text{Nenner}} = n_2 - 1$$

(Da die Populationsvarianzen nicht bekannt sind, schätzen wir diese.)

Statistische Entscheidung:

	ungerichtet	gerichtet*
falls ...	$F_{\text{emp}} > F_{n_1-1, n_2-1, 1-\alpha/2}$	$F_{\text{emp}} > F_{n_1-1, n_2-1, 1-\alpha}$
Annahme der ...	$H_1: \sigma_1^2 \neq \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$

* es gibt nur rechtsseitige gerichtete Hypothesen.

Verfahren für Ordinaldaten:

Werden in bei den Verfahren für Intervalldaten einige Voraussetzungen nicht erfüllt, oder hat man in der Untersuchung nur Rangreihen erhoben, muss man auf Verfahren für Ordinaldaten zurückgreifen.

Verfahren für Ordinaldaten sind „verteilungsfrei“, sie vergleichen nur Relationen (= Ordinaldaten) = Rangdaten.

Die folgenden Verfahren U-Test und Wilcoxon-Test, vorgestellt in den Punkten:

- „Vergleich von zwei unabhängigen Stichproben hinsichtlich ihrer zentralen Tendenz“ und
- „Vergleich von zwei abhängigen Stichproben hinsichtlich ihrer zentralen Tendenz“

sind Verfahren für Ordinaldaten.

Randomisationstest nach Fisher:

Da die hier vorgestellten Verfahren (U-Test & Wilcoxon-Test) auf der Randomisationsidee von Fisher basieren, wird diese kurz erklärt (am Beispiel unabhängiger Gruppen).

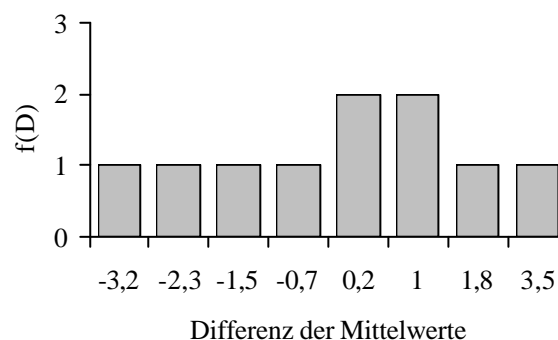
Das ganze Verfahren steht nicht im Bortz (in B3-Folien) und ist auch nur als Einleitung gedacht.

In einem Experiment wurden die Vbn zufällig auf zwei Gruppen (EG und KG) aufgeteilt.

Die H_0 wäre: Die Daten stammen aus derselben Population.

Fishers erstellt eine diskrete Verteilung einer interessierenden Prüfgröße (z.B. $D = \bar{x}_1 - \bar{x}_2$), die auf allen möglichen Aufteilungen aller Messwerte auf die Gruppen basieren. Im Gegensatz zur normalen Stichprobenverteilung, konstruiert Fisher eine Verteilung von Prüfgrößen, die auf Ziehungen **ohne** Zurücklegen basiert.

Konkret geht Fisher alle möglichen Kombinationen durch und bildet pro Gruppe jeweils den Mittelwert. Dann subtrahiert er den einen Mittelwert vom anderen, und erhält eine Verteilung.



Diese Verfahren für Ordinaldaten erstellt also eine Häufigkeitstabelle, an der man mit simpler Wahrscheinlichkeitsrechnung feststellen kann, wie groß die Irrtumswahrscheinlichkeit eines gemessenen Wertes ist.

Vergleich von zwei unabhängigen Stichproben hinsichtlich ihrer zentralen Tendenz (U-Test):

Wenn wir die Beeinträchtigung der Reaktionszeit unter Alkohol mit der Beeinträchtigung der Reaktionszeit unter Alkohol unter Einnahme des Präparates A (soll den negativen Effekt des Alkohols neutralisieren) vergleichen wollen, können wir nicht davon ausgehen, dass die Reaktionszeiten normalverteilt sind.

Dementsprechend können wir nur ein Verfahren für Ordinaldaten anwenden, den sogenannten **U-Test** (der funktioniert auch für ungleich große Stichproben).

Der U-Test ist so etwas ähnliches wie der t-Test für unabhängige Stichproben

Als Beispiel:

Wir messen bei insgesamt 27 Personen die Reaktionszeiten, bei 12 unter Alkohol und bei 15 unter Alkohol unter Einnahme des Präparates A.

Mit Alkohol		Mit Alkohol und Präparat A	
Reaktionszeit (ms)	Rangplatz	Reaktionszeit (ms)	Rangplatz
85	4	96	10
106	17	105	16
118	22	104	15
81	2	108	19
138	27	86	5
90	8	84	3
112	21	99	12
119	23	101	13
107	18	78	1
95	9	124	25
88	7	121	24
103	14	97	11
		129	26
		87	6
		109	20
	$T_1 = 172$		$T_2 = 206$

T_x = Summe der Rangplätze der Bedingung x

\bar{R} = Durchschnitt (Mittelwert) der Rangplätze

$\bar{R}_1 = 14,33$

$\bar{R}_2 = 13,73$

Bestimmung des U-Wertes:

Am Beispiel:

Die erste Person in Gruppe 1 hat den Rangplatz 4. In Gruppe zwei gibt es 13 Personen mit einem höheren Platz. Die zweite Person in Gruppe 1 hat den Rangplatz 17, der von 5 aus Gruppe 2 übertroffen wird.

Die Summe der Rangplatzüberschreitungen ist der U-Wert (also hier: 13+5+..... etc.).

Die Summe der Rangplatzunterschreitungen wird mit U' bezeichnet: $U = n_1 \cdot n_2 - U'$

Da das zählen sehr langwierig ist, geht das auch mit dieser Formel:

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$

Eingesetzt erhalten wir auf das Beispiel bezogen $U = 86$ und $U' = 94$. [U' erhält man, wenn man in der Formel im Zähler n_2 statt n_1 und am Ende T_2 statt T_1 benutzt.] [Zur Kontrolle kann man $U = n_1 \cdot n_2 - U'$ anwenden.]

Wenn n_1 oder n_2 größer als 10 sind, sind die Werte annähernd normal:

Dann erfolgt die Ermittlung des Populationsmittelwertes und des Standardfehlers:

$$m_U = \frac{n_1 \cdot n_2}{2} \quad s_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

Die eigentliche Formel lautet dann:

$$z = \frac{U - m_U}{s_U}$$

Bei unserem Beispiel erhalten wir $z = -0,20$

Dann entnehmen wir der z -Tabelle, dass auf einem Alphaniveau von 5% dieser (negative) Wert kleiner ist als unser empirischer z -Wert \Rightarrow Beibehaltung der H_0 .

Kleinere Stichproben:

Bei kleineren Stichproben bestimmt man nur U und schaut dann in der U -Tabelle nach.

Wichtig:

- Bei verbundenen Rangplätzen (mehrere Personen auf einem Rangplatz) muss die Streuung korrigiert werden (siehe Bortz S. 148)
- **Beim U-Test muss bei ungerichteter Hypothese das Alphaniveau verdoppelt werden.**

Vergleich von zwei abhängigen Stichproben hinsichtl. ihrer zentralen Tendenz (Wilcoxon-Test):

Der Wilcoxon-Test ist so etwas ähnliches wie der t-Test für abhängige Stichproben.

Zuerst wird die Differenz zwischen den zwei Messungen d_i (z.B. „vorher“ & „nachher“) ermittelt. Danach werden die Rangplätze vom Betrag von d_i aufgestellt.

Dann werden die Rangplätze speziell gekennzeichnet, deren Vorzeichen (das Vorzeichen bei d_i) weniger vorkommt.

Dann werden die Summen der speziell gekennzeichneten und der restlichen Rangplätze aufgestellt.

T = Summe der speziell gekennzeichneten Rangplätze

T' = Summe der restlichen Rangplätze

(Paare mit $d_i = 0$ bleiben unberücksichtigt!)

Verbundenen Rängen (z.B. drei Personen auf Platz 2) werden gemittelte Rangplätze vergeben.

Wenn es **keinen** Unterschied gibt, erwarten wir als T-Wert (nicht t-Wert, sondern „T“):

$$m_T = \frac{n \cdot (n + 1)}{4}$$

n = Anzahl der Messwertpaare, wobei $d_i = 0$ unberücksichtigt bleibt.

Je deutlicher der empirische T-Wert (das ist „T“) von m_T abweicht, desto geringer ist die Chance, dass der gefundene Unterschied zufällig zustande kam.

Wenn wir die Summe der speziell gekennzeichneten Rangplätze (T) ermittelt haben, schauen wir in der T-Tabelle (NEIN, NICHT t-Tabelle) beim kritischen Alphaniveau nach. Ist der T-Wert größer, müssen wir die H_0 beibehalten.

Größere Stichproben:

Bei Stichprobengrößen $n > 25$ geht die T-Verteilung (auch jetzt NICHT t-Verteilung) in eine NV über, so dass man mit der z-Tabelle arbeiten kann.

Dazu muss nur noch die Streuung errechnet werden:

$$s_T = \sqrt{\frac{n \cdot (n + 1) \cdot (2 \cdot n + 1) - \sum_{i=1}^k \frac{t_i^3 - t_i}{2}}{24}}$$

k = Anzahl der Rangbindungen und t_i = Länge der Rangbindung i

Und schon kann man die Formel

$$z = \frac{T - m_T}{s_T} \quad \text{angewendet werden.}$$

(Diesen Formelwald muss man aber nicht können.)

Der Wilcoxon-Test setzt im Gegensatz zum U-Test Intervallskalenniveau voraus. Trotzdem ist er ein Verfahren für Ordinaldaten. Wieso?

Der Wilcoxon-Test vergleicht Ränge, er kommt im Prinzip mit Ordinaldaten aus. Das Problem ist nur, dass er Rangplatzdifferenzen bildet, und so Ordinaldatenniveau dafür nicht mehr ausreicht. Trotzdem ist er jedoch ein nichtparametrisches Verfahren und wird somit unter Verfahren für Ordinaldaten gehandelt.

Verfahren für Nominaldaten:

Nominaldaten sind dann nötig, wenn Häufigkeitsunterschiede im Auftreten bestimmter Merkmale bzw. Merkmalskombinationen analysiert werden sollen

Da in fast allen im Bortz in Kapitel 5.3 behandelten Verfahren Prüfstatistiken vermittelt werden, die (annähernd) χ^2 -verteilt sind, werden die Verfahren zur Analyse von Häufigkeiten gelegentlich vereinfachend als χ^2 -Methoden bezeichnet.

 χ^2 -Verfahren dienen also der Analyse von Häufigkeiten.

χ^2 -Verfahren können natürlich auch für ordinal- und intervallskalierte Variablen eingesetzt werden, wenn die Häufigkeit bestimmter Merkmale untersucht werden soll. Die Merkmale werden dann wie nominalskalierte Merkmale behandelt, und somit ist auch das Ordinal- oder Intervallskalenniveau im Eimer.

- Die erwarteten Häufigkeiten repräsentieren immer die jeweilige Nullhypothese.

Übersicht über die χ^2 -Verfahren

	1 Merkmal	2 Merkmale	M Merkmale
2fach gestuft	a) einmalige Untersuchung: eindimensionaler χ^2 -Test b) zweimalige Untersuchung: McNemar χ^2 -Test c) mehrmalige Untersuchung: Cochran-Q-Test	4-Felder χ^2 -Test	Konfigurationsanalyse für alternative Merkmale
mehrfach gestuft	eindimensionaler χ^2 -Test: Vgl. einer empirischen mit einer theoretischen Verteilung	$k \times l$ - χ^2 -Test	Konfigurationsanalyse für mehrfach gestufte Merkmale

Vergleich der Häufigkeit eines zweifach gestuften Merkmals (eindimensionaler χ^2 -Test):**Einmalige Untersuchung:**

An einer Uni sind im Fach Psychologie 869 männliche (Wer's glaubt, wird selig.) und 576 weibliche Studenten immatrikuliert.

Kann man davon ausgehen, dass dieser Unterschied zufällig zustande kam?

Die Antwort auf diese Frage (und somit der ganze Rechenweg) ist davon abhängig, wie die Nullhypothese formuliert ist. Fragen wir danach, ob das Verhältnis männlich/weiblich an der ganzen Uni so ist, oder, ob das Geschlechterverhältnis im allgemeinen im Bereich Psychologie 50 : 50 ist.

Um es (mir) einfacher zu machen, nehme ich die zweite Hypothese (50 : 50) an.

Aufstellen der Hypothesen:

H_0 : Die empirische Verteilung entspricht der erwarteten Verteilung.

H_1 : Die empirische Verteilung weicht von der erwarteten Verteilung ab.

Zeichenerklärung:

f = Häufigkeit

f_{bj} = beobachtete Häufigkeit in Kategorie j und

f_{ej} = erwartete Häufigkeit in Kategorie j ,

k = Anzahl der Kategorien des Merkmals.

p_{ej} = erwartete Wahrscheinlichkeit des Merkmals j .

Dann Abstimmung der erwarteten Häufigkeiten:

$$f_{ej} = p_{ej} \cdot n.$$

D.h. bei unserem Beispiel wäre die Wahrscheinlichkeit für (z.B.) Männer 50%, es gibt insgesamt 1445 Psychologiestudenten, also

$$f_{ej} = 0,5 \cdot 1445 = 722,5$$

Eigentliche Formel:

$$\chi_{\text{emp}}^2 = \sum_{j=1}^k \frac{(f_{bj} - f_{ej})^2}{f_{ej}} \quad \text{mit } df = k - 1$$

An dieser Formel erkennt man die Grundstruktur aller χ^2 -Verfahren: Alle χ^2 -Verfahren laufen auf einen Vergleich von beobachteten und erwarteten Häufigkeiten hinaus, wobei die erwarteten Häufigkeiten jeweils die geprüfte Nullhypothese repräsentieren.

Unsere Werte eingesetzt:

$$\chi^2 = \frac{(869 - 722,5)^2}{722,5} + \frac{(576 - 722,5)^2}{722,5} = 59,41$$

Feststellen der Freiheitsgrade:

$df = k - 1$ es gibt 2 Kategorien (Männer/Frauen), also $df = 1$

Die Freiheitsgrade eines χ^2 -Wertes entsprechen der Anzahl der Summanden(wie viele Werte gibt es für f_b ? Hier zwei, nämlich Männer und Frauen) **abzüglich der Bestimmungsstücke der erwarteten Häufigkeiten, die aus den beobachteten Häufigkeiten abgeleitet wurden** (es gibt hier nur ein gemeinsames Bestimmungsstück: der Stichprobenumfang n : Die Summe der beobachteten und erwarteten Häufigkeiten ergibt jeweils n).

Feststellen von χ^2_{krit} :

$$\chi^2_{krit} = \chi^2_{(df; \alpha)}$$

Bei einer gerichteten Hypothese in einem χ^2 -Verfahren wird das Alphaniveau beim Nachschlagen in der Tabelle verdoppelt; soll heißen: bei $\alpha = 0,05$ schlage ich in der Tabelle bei 10% nach. Das liegt daran, dass alle χ^2 -Verfahren einseitig getestet werden, denn man betrachtet nur die rechte Seite der Verteilungen.

Die Überprüfung einer gerichteten Hypothese ist bei χ^2 -Verfahren nur möglich, wenn der resultierende χ^2 -Wert einen Freiheitsgrad hat.
Deswegen kann man bei einer gerichteten Hypothese auch mit der STNV testen, sofern man aus dem empirischen und kritischen χ^2 -Wert die Wurzel zieht.
Das Ergebnis eines χ^2 -Tests erhalten wir auch, wenn die Häufigkeit der Alternative 1 (z.B. hier weiblich) über die Binomialverteilung geprüft wird.
Ab $n < 10$ ist das nötig (bei zwei Merkmalen).

Entscheidung:

Falls $\chi^2_{emp} < \chi^2_{k-1, 1-\alpha}$, Beibehaltung der H_0 : Es ist davon auszugehen, dass die Verteilung des untersuchten Merkmals nicht von der unter der H_0 erwarteten Verteilung abweicht.

Entscheidung bei unserem Beispiel:

χ^2_{krit} ist bei uns 3,84; χ^2_{emp} ist bei uns 59,41 \Rightarrow Annahme der H_1

Voraussetzungen:

- Jede untersuchte Einheit kann eindeutig einer Kategorie zugeordnet werden.
- Die erwarteten Häufigkeiten sind in jeder Kategorie größer als 5. Im Falle von $k = 2$ sollten in jeder Kategorie 10 Elemente erwartet werden.

Weitere tolle Modifikationen und Verfahren:**Kontinuitätskorrektur:**

Die Kontinuitätskorrektur, auch Yates-Korrektur genannt, schätzt χ^2 besser, da Häufigkeiten diskret, χ^2 -Werte aber stetig verteilt sind. D.h. jedes $|f_b - f_e|$ wird um $-0,5$ vermindert. Die Kontinuitätskorrektur wirkt der Tendenz nach konservativ (schützt H_0).

Der McNemar-Test:

Dieser Test wird bei zweimaliger Untersuchung und Prozentunterschieden benutzt, **er berücksichtigt nur diejenigen Fälle, bei denen eine Veränderung eingetreten ist.**

Der Cochran-Q-Test:

ist eine Erweiterung des McNemar-Tests, gilt auch für mehrmalige Untersuchungen.

Nur anzuwenden bei $n \cdot m > 30$

(Wer sich das alles – Kontinuitätskorrektur, McNemar-Test & Cochran-Q-Test – antun möchte: steht alles im Bortz auf S. 155-158)

Vergleich der Häufigkeit eines k-fach gestuften Merkmals (eindimensionaler χ^2 -Test):

Diese Variation des eindimensionalen χ^2 -Tests ist mit dem eindimensionalen χ^2 -Test für zweifachgestufte Merkmale ziemlich identisch.

Es gibt zwei Möglichkeiten:

- 1.) **Test auf Gleichverteilung (z.B. je 50% Männer/Frauen) und**
- 2.) **Test auf andere Verteilungsformen.**

Für beide Möglichkeiten gilt:

Die Formel ist mit dem des eindimensionalen χ^2 -Test für zweifachgestufte Merkmale identisch, es gibt halt nur mehr Bedingungen j , ergo wird die Berechnung langwieriger.

$$\chi^2_{\text{emp}} = \sum_{j=1}^k \frac{(f_{bj} - f_{ej})^2}{f_{ej}}, \text{ mit } df = k-1$$

Bei der Frage, ob sich Variable 1 von den anderen unterscheidet, wird von den Variablen 2 bis k der Durchschnitt gebildet (fungiert als f_e) & mit Variable 1 (als f_b) über den χ^2 -Test verglichen.

Die Voraussetzungen sind ebenso identisch, nur dass man bei weniger als fünf erwarteten Häufigkeiten pro Kategorie (wie beim eindimensionalen χ^2 -Test für zweifachgestufte Merkmale) anstelle der Binomialverteilung nun eine **Multinomialverteilung** rechnen sollte.

All das oben genannt galt noch für beide Möglichkeiten. Die erste Möglichkeit ist hiermit abgeschlossen.

Das Folgende bezieht sich auf:

Tests auf andere Verteilungsformen.

Sofern die H_0 nicht behauptet, dass eine Gleichverteilung vorherrscht (z.B. je 50% Männer/Frauen), haben wir es mit Test für andere Verteilungsformen zu tun, denn **es ergibt sich bei steigender Variablenanzahl eine theoretische Verteilung, die mit einer empirischen (beobachteten) Verteilung verglichen wird.**

**Der Vergleich einer empirischen mit einer theoretischen Verteilung heißt
Goodness-of-fit-test!**

Beim Goodness-of-fit-Test werden die erwarteten Häufigkeiten aus einer theoretischen Verteilung abgeleitet:

Man kann so z.B. prüfen, ob sich eine empirische Verteilung normalverteilt oder z.B. nach der Poisson-Verteilung verteilt.

- Der Goodness-of-fit-Test dient also z.B. der Überprüfung der Voraussetzung, dass Grundgesamtheit normalverteilt (H_0 : normalverteilt, H_1 : nicht normalverteilt)

Problem: die H_0 als „Wunschhypothese“

- wenn der χ^2 -Wert auf 5%-Niveau signifikant ist, heißt das, dass die Wahrscheinlichkeit für eine Normalverteilung der Werte kleiner als 5% ist, wenn es nicht signifikant ist, bedeutet es lediglich, dass es mit mehr als 5% Wahrscheinlichkeit normalverteilt ist - mehr nicht.
Wir wollen aber die H_0 „beweisen“ (sonst will man normalerweise ja die H_1 „beweisen“), deshalb versucht man den Beta-Fehler möglichst klein zu halten, nicht Alpha (ist wurscht).
Da aber die Hypothese in diesem Fall nicht spezifisch ist, kann der Beta-Fehler nicht bestimmt werden, deswegen wird einfach der Alpha-Fehler enorm vergrößert, z.B. auf 25%
Wenn wir auf 25%tigem Alphaniveau testen und die H_1 dann immer noch nicht signifikant ist, haben wir H_0 gesichert.
- Beachtet: die H_0 als Wunschhypothese beizubehalten, wird mit wachsendem Stichprobenumfang unwahrscheinlicher!

Der Goodness-of-fit-Test:

Zuerst werden die Daten in Kategorien eingeteilt, d.h. wenn wir einen Wert haben, können wir ihn in eine Kategorie packen. Ein gutes Beispiel wären die gemessenen km/h der Wagen, eingeteilt in 10er-Kategorien, beginnend von 60km/h bis (sagen wir mal) 220 km/h.
(Das Ganze läuft unter der H_0 , dass unsere Kategorien normalverteilt sind.)

Also eine Tabelle:

Intervall	Beobachtete Häufigkeiten	Erwartete Häufigkeiten	$\frac{(f_b - f_e)^2}{f_e}$
60,0-69,9 km/h	4	3	0,33
70,0-79,9 km/h	10	6	2,67
etc.	etc.	etc.	etc.
			Die Summe davon ergibt den empirischen χ^2 -Wert

Der kritische χ^2 -Wert lautet ausgeschrieben $\chi^2_{(k-3, 1-\alpha)}$,

somit sind die Freiheitsgrade $df = k - 3$, wobei k die Anzahl der Kategorien angibt.

Poisson-Verteilung:

Bei einer Frage nach einer Poisson-Verteilung sind die Freiheitsgrade $df = k - 2$.

Vom Prinzip läuft es wie oben, nur werden die erwarteten Häufigkeiten nach der Poisson-Verteilung ermittelt (Formel im Bortz S. 71; der Rest S. 161 & 162)

Vergleich der Häufigkeiten von zwei alternativen Merkmalen (4-Felder- χ^2 -Test):

Werden n voneinander unabhängigen Beobachtungen nicht nur einer, sondern zwei Merkmalsalternativen zugeordnet, erhalten wir eine **4-Felder-Kontingenztafel** bzw. eine **bivariate Häufigkeitsverteilung**.

Und schon sind wir bei einem 4-Felder- χ^2 -Test.

(den kann man auch schon $k \times l$ - χ^2 -Test nennen)

Ein Beispiel: 100 Personen sollen auf die Merkmalsalternativen männlich/weiblich und Brillenträger/Nicht-Brillenträger untersucht werden.

	männlich	weiblich	
mit Brille	25	10	35
ohne Brille	25	40	35
	50	50	100

Jetzt gibt es zwei Möglichkeiten:

- 1.) Vorgegebene erwartete Häufigkeiten und
- 2.) Geschätzte erwartete Häufigkeiten

Vorgegebene erwartete Häufigkeiten:

Dann könnte unsere vorgegebene H_0 lauten:

Brillenträger = 30% in der Bevölkerung

Männer/Frauen ja 50% in der Bevölkerung

Dementsprechend wären unsere Wahrscheinlichkeiten

$p(\text{Brille, Mann}) = 0,15$

$p(\text{Brille, Frau}) = 0,15$

$p(\text{ohne Brille, Mann}) = 0,35$

$p(\text{ohne Brille, Frau}) = 0,35$

Dann setzen wir das ein in folgende Formel:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(f_{b(i,j)} - f_{e(i,j)})^2}{f_{e(i,j)}}$$

und bekommen ein $\chi^2 = 11,90$

Die Freiheitsgrade $df = 2 \times 2 - 1 = 3$

Dementsprechend ist unser kritisches $\chi^2 = 11,34 \Rightarrow$ Annahme der H_1 , Frauen tragen weniger Brillen, Männer mehr Brillen als erwartet.

Geschätzte erwartete Häufigkeiten:

Beim 4-Felder- χ^2 -Test entsprechen die erwarteten Zellenhäufigkeiten den bei stochastischer Unabhängigkeit zu erwartenden:

$$f_{eij} = p_i \cdot p_j \cdot n = \frac{(\text{Summe Zeile } i) \cdot (\text{Summe Spalte } j)}{n}.$$

Der Rest läuft wie bei vorgegebenen Nullhypothesen, nur sind die Freiheitsgrade $df = 1$ df .

Dementsprechend ist auch das kritische c^2 anders.

Auch hier wird übrigens die H_1 angenommen (Bortz S. 164)

Kontinuitätskorrektur:

- Auch hier ist eine Kontinuitätskorrektur nach Yates möglich

Prozentunterschiede:

- laufen genauso wie beim eindimensionalen c^2 -Test

Vergleich der Häufigkeiten von zwei mehrfach gestuften Merkmalen ($k \times l$ - χ^2 -Test):

Im vorherigen Punkt waren beide Merkmale jeweils zweifach gestuft. Beim $k \times l$ - χ^2 -Test ist die Anzahl der Stufen unwichtig

Das Merkmal A ist k -fach gestuft und das Merkmal B l -fach.

Der $k \times l$ - χ^2 -Test überprüft die Alternativhypothese:

- H_1 , dass zwischen zwei Merkmalen ein *irgendwie gearteter* Zusammenhang besteht, gegen die Nullhypothese
- H_0 , dass es zwischen den beiden Merkmalen keinen Zusammenhang gibt bzw. die Merkmale stochastisch unabhängig sind.

Das Verfahren läuft wie beim 4-Felder- c^2 -Test:

$$\chi^2_{\text{emp}} = \sum_{i=1}^k \sum_{j=1}^l \frac{(f_{bij} - f_{eij})^2}{f_{eij}}$$

Beim $k \times l$ - χ^2 -Test entsprechen die erwarteten Zellenhäufigkeiten den bei stochastischer Unabhängigkeit zu erwartenden:

$$f_{eij} = p_i \cdot p_j \cdot n = \frac{(\text{Summe Zeile } i) \cdot (\text{Summe Spalte } j)}{n}.$$

Freiheitsgrade:

Bei geschätzten erwarteten Häufigkeiten: $df = (k - 1) \cdot (l - 1)$

Bei vorgegebenen erwarteten Häufigkeiten: $df = (k - 1) \cdot (l - 1) - 1$

Entscheidung

Falls $\chi^2_{\text{emp}} > \chi^2_{(k-1)(l-1), 1-\alpha}$, Entscheidung für die H_1 : Es ist davon auszugehen, dass die Variablen voneinander stochastisch (irgendwie) abhängig sind.

Voraussetzungen

- Jede untersuchte Einheit kann eindeutig den Kategorien (Zellen) zugeordnet werden.
- Die erwarteten Häufigkeiten sind alle größer als 5.

Der $k \cdot l$ - χ^2 -Test und der 4-Felder- χ^2 -Test lassen sich auch mit Prozentwerten durchführen.

Vgl. d. Häufigkeiten von m alternativ oder mehrf. gestuften Merkmalen (Konfigurationsanalyse):

Verallgemeinern wir das 4-Felder χ^2 auf m alternative Merkmale, erhalten wir eine **mehrdimensionale Kontingenztafel**, die mit der **Konfigurationsanalyse** analysiert werden kann.

Ein Beispiel:

Wir überprüfen, ob Frauen, die in der Stadt leben, überzufällig berufstätig sind ($\alpha = 0.01$).

Also haben wir drei alternative Merkmale ($m = 3$):

A = Stadt (+) vs. Land (–)

B = männlich (+) vs. weiblich (–)

C = berufstätig (+) vs. nicht berufstätig (–)

Merkmal			Häufigkeiten		
A	B	C	f_b	f_e	$\frac{(f_b - f_e)^2}{f_e}$
+	+	+	120	86,79	12,71
+	+	–	15	63,33	36,88
+	–	+	70	95,32	6,73
+	–	–	110	69,56	23,51
–	+	+	160	89,54	55,45
–	+	–	10	65,34	46,87
–	–	+	20	98,35	62,42
–	–	–	135	71,77	55,71
			$n_b = 640$	$n_e = 640$	$\chi^2 = 300,28$

Die erwarteten Häufigkeiten werden geschätzt mit:

$$f_{e(i,j,k)} = \frac{(\text{Summe } A_i) \cdot (\text{Summe } B_j) \cdot (\text{Summe } C_k)}{n^2}$$

In unserem Beispiel lauten die Summen

$$\begin{array}{lll} A(+) = 315 & B(+) = 305 & C(+) = 370 \\ A(-) = 325 & B(-) = 335 & C(-) = 270 \end{array}$$

$$(\text{z.B. } f_{e(+++)} = 315 \times 305 \times 370 / 640^2 = 86,79)$$

Freiheitsgrade: $df = 1$

Der kritische Wert (5,43) ist kleiner als der Empirische (6,73) für unsere These, dass in der Stadt mehr Frauen berufstätig sind.

Da das aber unsere H_0 war, müssen wir nun die H_0 annehmen bzw. beibehalten (nicht davon verwirren lassen, ist ein Fehler bei der Hypothesenaufstellung).

Allgemeine Bemerkungen zu den χ^2 -Techniken:

χ^2 -Techniken gehören – von der Durchführung her – zu den einfachsten Verfahren der Elementarstatistik, wenngleich der mathematische Hintergrund dieser Verfahren komplex ist.

Mit Hilfe der χ^2 -Verfahren werden die Wahrscheinlichkeiten multinomialverteilter Ereignisse geschätzt (wobei die Schätzung erst bei $n \rightarrow \infty$ exakt übereinstimmen).

Deshalb sollte beachtet werden:

- Die einzelnen Beobachtungen müssen voneinander unabhängig sein (Ausnahme: McNemar- & Cochran-Test)
- Jede untersuchte Einheit muss eindeutig einer Kategorie zugeordnet werden können.
- Nicht mehr als 20% der erwarteten Häufigkeiten in den Kategorien sollten kleiner als 5 sein.

Verfahren zur Überprüfung von Zusammenhangshypothesen:

Die Verfahren zur Überprüfung von Zusammenhangshypothesen haben der ua. psychologischen Forschung so viel Impulse wie kein anderes statistisches Verfahren verliehen.

Zusammenhänge sind in der Mathematik und in den Naturwissenschaften bekannt, z.B. verändert sich der Umfang eines Kreises proportional zu seinem Umfang.

Solche exakten, funktionalen Zusammenhänge haben wir in der Psychologie nicht. Deswegen sprechen wir von **stochastischen (zufallsabhängigen) Zusammenhängen**, die nach Höhe des Zusammenhangs unterschiedlich präzise Vorhersagen zulassen.

(Bei einer Korrelation von +1 oder -1 geht unser stochastischer Zusammenhang über in einen funktionalen, deterministischen Zusammenhang.)

Im Kapitel „Merkmalszusammenhänge & –vorhersagen“ habe ich einige Punkte aus dem 6. Kapitel im Bortz vorgezogen – also bitte nicht wundern, wenn jetzt was „fehlt“. Die behandelten Punkte waren, bezogen auf die Bortz-Unterkapitel:

- 6.1 Lineare Regression, dabei: Kovarianz und Regression
- 6.2.1 Kovarianz und Korrelation, dabei: Produkt-Moment-Korrelation; Korrelation und Regression; Beobachtete, vorhergesagte und Residualwerte; Determinationskoeffizient
- 6.4 Korrelation und Kausalität

Die Statistische Absicherung von linearen Regressionen:

Regressionsgleichungen werden auf der Grundlage einer repräsentativen Stichprobe bestimmt, um sie auch auf andere Untersuchungseinheiten, die zur selben Population gehören, anwenden zu können.

Das bedeutet aber, dass eine (auf Basis einer Stichprobe gebildete) Regressionsgleichung generalisierbar sein muss. Aus dieser Population können wir eine durch (unendlich) viele Stichproben wieder eine SKV bilden.

Je größer aber die Streuung (der Standardfehler) dieser SKV ist, desto weniger werden die Werte einer einzelnen Stichprobe (hier: einer einzelnen Regressionsgerade) die Regressionsgerade der Population vorhersagen können.

Die nach der Methode der kleinsten Quadrate ermittelte Regressionsgleichung stellt nur eine Schätzung der folgenden, in der Population gültigen, Regressionsgerade dar:

$$\hat{y}_j^* = \mathbf{b}_{yx} \cdot x_j + \mathbf{a}_{yx}$$

\hat{y}_j^* kennzeichnet hierbei einen \hat{y} -Wert, der aufgrund der Populations-Regressionsgleichung vorhergesagt wurde.

- Wichtig: nicht die Punktabstände zur Geraden minimieren, sondern die Abstände in y-Richtung, damit für die Vorhersage x auf y optimiert, deswegen gibt es für eine Regression von y auf x auch andere Koeffizienten a & b.

Voraussetzungen:

- Unsere SKV verteilt sich (gemäß Abb. 6.7 im Bortz S. 183) als bivariate Häufigkeitsverteilung:
 - a.) Die Verteilung der x-Werte muss für sich genommen normal sein.
 - b.) Die Verteilung der y-Werte muss für sich genommen normal sein.
 - c.) Die zu einem x-Wert gehörenden y-Werte (Arrayverteilungen)* müssen normalverteilt sein.
 - d.) Die Mittelwerte der Arrayverteilungen müssen auf einer Geraden liegen.

* Ein Arrayverteilung von y-Werten (zu natürlich jeweils zu einem x-Wert gehören) bedeutet, das bei nicht perfektem Zusammenhang gleiche Untersuchungen eines bestimmten x-Merkmals verschiedene y-Merkmale aufweisen.

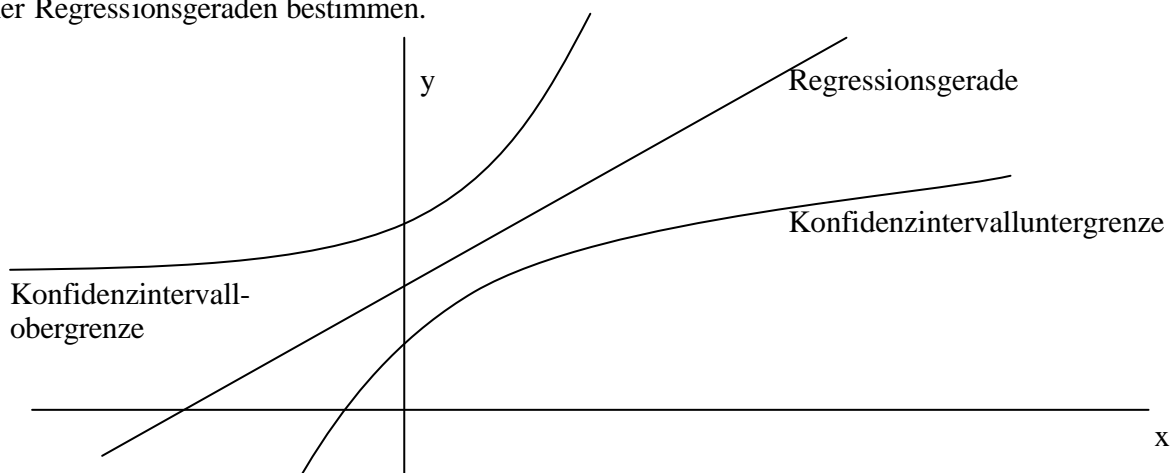
Die Genauigkeit der Regressionsvorhersagen:

Die Genauigkeit der Regressionsvorhersagen wird durch den Standardschätzfehler (STSF) bestimmt. (Formel siehe Bortz S. 185)

Der Standardschätzfehler kennzeichnet die Streuung der y-Werte um die Regressionsgerade und ist damit ein Gütemaßstab für die Genauigkeit der Regressionsvorhersagen. Die Genauigkeit einer Regressionsvorhersage wächst mit kleiner werdendem Standardschätzfehler.

Determinanten der Vorhersagegenauigkeit (durch Konfidenzintervalle):

Durch Konfidenzintervalle kann man die Vorhersagegenauigkeit einzelner Punkt um und auf einer Regressionsgeraden bestimmen.

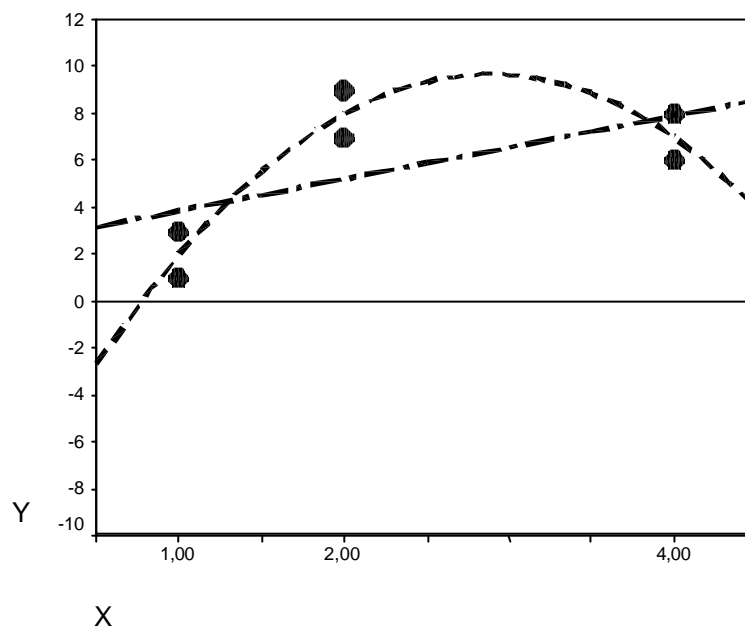
**Bestimmung der Größe des Konfidenzintervalls (KI):**

- Je kleiner der Konfidenzoeffizient ($1 - \alpha$), desto kleiner KI
- Je kleiner der Standardschätzfehler (STSF), desto kleiner KI.
- Je kleiner n , desto größer KI.
- Je kleiner die Varianz x (y ist im STSF), desto kleiner KI.

Nonlineare Regression:

Gelegentlich erwartet man, dass eine nicht-lineare Beziehung eine bessere Vorhersage macht als eine lineare Beziehung. Als Beispiel mit einer nicht- und einer linearen Regression:

X	1	1	2	2	4	4
Y	1	3	7	9	6	8



Hier dürfte die nicht-lineare Beziehung die Punkte besser vorhersagen (Grafik aus den B3-Folien, leicht abgewandelt)

Varianten (einige ausgewählte):**Umgekehrt U-förmige Beziehung** (Bortz S. 189):

Eine umgekehrt U-förmige bzw. parabolische Beziehung wird durch eine quadratische Regressionsgleichung oder ein Polynom 2. Ordnung modelliert.

$$\hat{y} = a + b_1 \cdot x + b_2 \cdot x^2$$

Polynome höherer Ordnung:

Ein (umgekehrt) S-förmiger Zusammenhang lässt sich durch eine kubische Regressionsgleichung bzw. ein Polynom 3. Ordnung anpassen:

$$\hat{y} = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$$

Nonlineare Zusammenhänge, die über ein Polynom 3. Ordnung hinausgehen, können in der u.a. psychologischen Forschung nur sehr selten begründet werden. Die Aufstellung eines solchen Polynoms kann bestenfalls ex post (= ohne theoretische Vorannahmen) sinnvoll sein.

[Die Regressionsgleichung wird einfach um weitere b und x erweitert, wobei der Exponent pro weiterem b um 1 erhöht wird.]

Linearisierung:

Über Logarithmierung kann man die Funktion von Gleichungen ermitteln.

Geht man von einem exponentiellen Zusammenhang von zwei Variablen aus, erhält man die Gleichung:

$$\hat{y} = a \cdot x^b$$

Durch Logarithmierung (siehe Bortz S. 192) gelangt man schlussendlich zu folgender Gleichung:

$$\hat{y}' = a' + b' \cdot x$$

Was verteuft wie eine lineare Regression aussieht.

[Ganz exakt sind die so ermittelten Regressionskoeffizienten leider nicht, sie unterscheiden sich von denen, die wir mit der Methode der kleinsten Quadrate ermitteln würden.]

Interpretationshilfen für r:

Angenommen ein Schulpsychologo ermittelt eine Korrelation von $r = 0,60$ zwischen Abiturschnitt (y) und Intelligenzquotienten (x).

Was bedeutet diese Zahl?

Um das zu veranschaulichen, dichotomisieren wir beide Variablen am Median und erhalten eine 4-Felder-Tafel.

(dichotom = ein Merkmal mit nur zwei Ausprägungen, z.B. Geschlecht. Man kann aber auch intervallskalierte Daten dichotomisieren, d.h. zweiteilen. D.h. hier teilen wir am Median in zwei Hälften.)

Wenn wir dem Schulpsychologen jetzt auftragen, mal zu schätzen, wie sich die Abiturschnitte und IQs bei 200 Schülern verteilen, wenn er am Median teilen soll, so werden bei keinem Zusammenhang ($r = 0$) je 50 Schüler pro Kategorie verteilt sein.

Aber es verteilt sich so bei $r = 0,60$:

		Abiturschnitt (y)		
		< Md _y	> Md _y	
IQs (x)	< Md _x	80	20	100
	> Md _x	20	80	100
		100	100	200

Wieso?

Bei zwei symmetrischen, mediandichotomisierten Merkmalen gibt r an, um wie viel der Fehler minimiert wird.

Im Beispiel betrug der Fehler bei Zufallsverteilung (je 50 n pro Kategorie) 50% Diese 50% reduziert um $r = 0,60$ bzw. 60% von $50 = 30$. Diese 30 werden einer Korrelation von $r = 0,60$ nicht mehr falsch eingeordnet, sondern richtig, d.h.

Die Fehler (50%), die wir bei der Vorhersage von x durch y machen (großes y = großes x), werden bei einer Korrelation von 0,60 um 30 % vermindert.

Mehr Infos Bortz S. 201/202.

k-fach gestufte Merkmale:

Diese Variante ist einfach eine Erweiterung der 4-Felder-Matrix um weitere Kategorien.

Eine Besonderheit dabei ist, dass größere Abweichungen um mehrere Kategorien stärker „bestraft“ werden als andere. Vom Prinzip wie die 4-Felder-Matrix, jedoch aufwendiger zu rechnen. Wer mehr wissen will: Steht alles im Bortz auf S. 203/204.

Korrelation für nichtlineare Zusammenhänge:

Der Zusammenhang zwischen nichtlinearen Variablen lässt sich ermitteln, indem man die Varianz der vorhergesagten Werte durch die Varianz der tatsächlichen Werte teilt:

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Überprüfung von Korrelationshypothese:

Wird aus einer bivariaten, intervallskalierten Grundgesamtheit eine Stichprobe gezogen, kann ein Korrelationskoeffizient r berechnet werden. Er zeigt den linearen Zusammenhang zwischen den zwei Merkmalen an, bzw., quadriert (r^2), den Anteil gemeinsamer Varianz.

Voraussetzungen:

- Grundgesamtheit bivariat normalverteilt
- Kovarianz & Korrelation der Grundgesamtheit bekannt
- Die einzelnen Merkmalsstufenkombinationen müssen voneinander unabhängig sein.

Die ganzen Voraussetzungen sind aber ein bisschen schwierig zu erfüllen (siehe Bortz S.205), dementsprechend beschränkt man sich darauf, dass beide einzelnen Merkmale NV sind, die Arrayverteilungen normalverteilt sind und die Varianzen der Arrayverteilungen homogen ist.

Wenn man diese Voraussetzungen missachtet, muss man mit größerem Alpha- und Beta-Fehler rechnen. Trotzdem ist der Test relativ robust.

Signifikanztest:

Ob eine empirisch ermittelte Korrelation r mit der Hypothese

$$H_0: \rho = 0$$

zu vereinbaren ist, lässt sich mit folgender Prüfgröße testen:

$$t_{emp} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}, \text{ mit } df = n-2.$$

Ab $n > 3$ kann man zeigen, dass der Ausdruck bei $df = n - 2$ t-verteilt ist.

Entscheidung:

	ungerichtet	rechtsseitig	linksseitig
falls ...	$ t_{emp} > t_{(df; \alpha/2)} $	$t_{emp} > t_{(df; 1-\alpha)}$	$t_{emp} < t_{(df; \alpha)}$
Annahme ...	$H_1: \rho \neq 0$	$H_1: \rho > 0$	$H_1: \rho < 0$

Für den zweiseitigen Test einer Korrelation sind in der Tabelle D (Bortz 1999, S. 775) kritische Wert für $\alpha = 0,05$ und $\alpha = 0,01$ angegeben. Diese Werte erhält man, wenn man obige Prüfgröße nach r auflöst:

$$r_{krit} = \frac{t_{krit}}{\sqrt{n-2+t_{krit}^2}}.$$

Fischers Z-Transformation:

Besteht in der Grundgesamtheit ein Zusammenhang $r \neq 0$, erhalten wir für (theoretisch unendlich) viele Stichproben eine rechts- ($r > 0$) oder linkssteile ($r < 0$) SKV der Korrelationen. Eine Nullhypothese mit einem $r \neq 0$ kann somit nichts mehr messen (da asymmetrisch).

Diese rechts- oder linkssteile Verteilung lässt sich durch Fishers Z-Transformation wieder in Normalverteilung transformieren:

$$Z = \frac{1}{2} \cdot \ln\left(\frac{1+r}{1-r}\right) \quad \text{aufgelöst nach } r \quad r = \frac{e^{2Z} - 1}{e^{2Z} + 1}$$

wobei \ln = Logarithmus zur Basis $e (\approx 2,718)$.

Je weiter r von ± 1 entfernt und je größer n ist, desto normalverteilter ist die Kurve nach Fishers Z-Transformation

Weiterhin sind die transformierten Werte Teil einer Verhältnisskala (fester Nullpunkt), man kann nämlich bei einer Korrelation nicht davon ausgehen, dass der Zusammenhang bei $r = 0,4$ halb so groß ist wie bei $r = 0,8$!

Wenn es transformiert wird zeigt sich: die Zuwachsraten im oberen Korrelations-Bereich sind bedeutsamer.

Auch Mittelwerte & Varianzen von einer Korrelation sind nicht interpretierbar, sofern sie nicht transformiert wurden.

**Bei Mittelwertsbildung von (z.B. 3) Korrelationen also erst transformieren, dann Mittel der Fisher-Z- Werte bilden, dann dieses Mittel rücktransformieren:
Dabei werden höhere Korrelationen stärker berücksichtigt als niedrige.**

Es gibt mehrere Variationsmöglichkeiten von Fischers Z-Transformation, sie stehen im Bortz auf den Seiten 210 – 214. Wen es interessiert, der möge es nachschlagen.

Ich glaube aber nicht, dass das für die Prüfung in irgendeiner Form relevant sein könnte.

Spezielle Korrelationstechniken:

Unter dem Punkt „Überprüfung von Korrelationshypothesen“ ging es um die Produkt-Moment-Korrelation zweier intervallskalierter Merkmale.

Das Ganze geht aber natürlich auch mit ordinal- und nominalskalierten sowie dichotomen Merkmalen.

Übersicht der bivariaten Korrelationsarten:

Merkmal y	Merkmal x		
	Intervallskala	dichotomes Merkmal	Ordinalskala
Intervallskala	Produkt-Moment-Korrelation	Punkt-biseriale Korrelation	Rangkorrelation
dichotomes Merkmal	-	Φ - Koeffizient	Biseriale Rangkorrelation
Ordinalskala	-	-	Rangkorrelation

Dazu gibt es noch den Kontingenzkoeffizienten, der den Zusammenhang zweier nominalskalierter Merkmale bestimmt. Da dieser Koeffizient aber kein Korrelationsmaß im engeren Sinn darstellt, ist er nicht in der Tabelle.

I.) Korrelation zweier Intervallskalen:

Berechnung der Produkt-Moment-Korrelation, siehe „Überprüfung von Korrelationshypothesen“

II.) Korrelation einer Intervallskala mit einem dichotomen Merkmal:

A.) Punktbiseriale Korrelation:

Ein Zusammenhang zwischen einem dichotomen Merkmale (z.B. männlich/weiblich) und einem intervallskalierten Merkmal (z.B. Körpergewicht) wird durch eine punktbiseriale Korrelation (auch Produkt-Moment-biseriale Korrelation genannt) erfasst.

Eine punktbiseriale Gleichung erhält man, wenn die Werte 0 und 1 für die beiden dichotomen Variablen in die Produkt-Moment-Korrelations-Gleichung eingesetzt werden. Dadurch vereinfacht sich die Korrelationsformel:

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \cdot \sqrt{\frac{n_0 \cdot n_1}{n^2}}$$

wobei

- n_0, n_1 = Anzahl der Untersuchungseinheiten in den Merkmalskategorien x_0 und x_1 .
 \bar{y}_0, \bar{y}_1 = durchschnittliche Ausprägung des Merkmals y (Körpergewicht) bei den Untersuchungseinheiten x_0 (Männer) und x_1 (Frauen).
 $n = n_0 + n_1$ = Gesamtstichprobe
 s_y = Streuung der kontinuierlichen y -Variablen

Die Signifikanzüberprüfung ($H_0: r = 0$) erfolgt wie bei der Produkt-Moment-Korrelation durch:

$$t = \frac{r_{pb}}{\sqrt{(1 - r_{pb}^2) / (n - 2)}}$$

Der so ermittelte t -Wert ist mit $n - 2$ Freiheitsgraden versehen und in der t -Tabelle mit t_{krit} verglichen.

Die punktbiseriale Korrelation entspricht (als Verfahren zur Überprüfung von Zusammenhangshypothesen) **dem t-Test für unabhängige Stichproben**(als Verfahren zur Überprüfung von Unterschiedshypothesen).

B.) Biseriale Korrelation:

Bei einem künstlich dichotomisiertem Merkmal und einem intervallskalierten Merkmal wird eine biseriale Korrelation (r_{bis}) durchgeführt.

- Biseriale Korrelation (r_{bis}) und Punktbiseriale Korrelation (r_{pb}) hängen statistisch zusammen.
- Es gibt auch noch triseriale und polyseriale Korrelation
- Das steht alles im Bortz auf S. 216-218

III.) Korrelation einer Intervallskala mit einer Ordinalskala:

- Richtige Verfahren hierfür gibt es (noch) nicht. Für die Praxis: Herabstufung der Intervallskala auf Ordinalskalenniveau, Anwendung der Korrelation zweier Ordinalskalen.

IV.) Korrelation für zwei dichotome Variablen:

Wenn die Merkmale x und y jeweils dichotom sind, kann ihr Zusammenhang durch den Phi-Koeffizienten (Φ) ermittelt werden

Da es nur 4 Häufigkeiten gibt...

		Merkmal (y)	
		0	1
Merkmal (x)	0	a	b
	1	c	d

...ist die Berechnung von Φ einfach:

$$\Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a + c) \cdot (b + d) \cdot (a + b) \cdot (c + d)}}$$

Es gibt einen engen Zusammenhang mit dem 4-Felder- χ^2 , deshalb Signifikanzprüfung mittels

$$\chi^2 = n \cdot \Phi^2 \quad \text{mit df} = 1$$

Der Zusammenhang zwischen dem 4-Felder- χ^2 und Phi-Koeffizienten:

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Achtung:

Der Φ -Koeffizient liegt nur im üblichen Wertebereich von -1 bis +1, wenn die Aufteilung der Stichprobe in Alternativen von x der in die Alternativen von y entspricht, also bei identischen Randverteilungen.

Ansonsten gibt es einen geringeren Φ_{maximal} -Wert (ist mit Formel berechenbar)

Bei der Bestimmung von Φ_{max} ist darauf zu achten, dass

- das Vorzeichen von Φ_{max} mit dem Φ_{emp} von übereinstimmt.
- das Vorzeichen von Φ_{max} beliebig ist.

Gelegentlich versuchen Einige die Aufwertung eines empirischen Φ -Wertes am Φ_{max} , um es vergleichbar mit der Produkt-Moment-Korrelation zu machen, das ist aber fragwürdig, da auch die Produkt-Moment-Korrelation nur bei identischen Randverteilungen einen Wertebereich von -1 bis +1 aufweist.

Sind beide Variablen künstliche Dichotomien, arbeitet man mit einer **Tetrachorischen Korrelation** (siehe Bortz S. 220/221)

V.) Korrelation eines dichotomen Merkmals mit einer Ordinalskala:

Die Korrelation eines dichotomen Merkmals mit einer Ordinalskala – auch biseriale Rangkorrelation (r_{bisR}) genannt – wird berechnet, wenn ein Merkmal (x) in künstlicher oder natürlicher Dichotomie vorliegt und das andere Merkmal (y) rangskaliert. (ähnlich dem U-Test)

$$r_{bisR} = \frac{2}{n} \cdot (\bar{y}_1 - \bar{y}_2)$$

wobei

\bar{y}_1 = durchschnittlicher Rangplatz der zu x_1 gehörenden Untersuchungseinheiten

\bar{y}_2 = durchschnittlicher Rangplatz der zu x_2 gehörenden Untersuchungseinheiten

n = Umfang der Stichprobe

Der kritische Wert wird – bei hinreichend großen Stichproben – mit dem U-Test berechnet:

$$z = \frac{U - m_U}{s_U}$$

VI.) Korrelation zweier Ordinalskalen (Spearman's Rho):

Der Zusammenhang zweier ordinalskalierten Merkmale wird durch die **Rangkorrelation** nach Spearman (r_s oder r) erfasst.

Eigentliche Formel:

Errechnung des kritischen Wertes ab $n \geq 30$

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}$$

$$t = \frac{r_s}{\sqrt{(1 - r_s^2) / (n - 2)}} \quad \text{df} = n - 2$$

(Dieselbe Formel wie bei der punktbiserialen Korrelation)

- bei verbundenen Rängen gibt es Modifikationen (Bortz S. 224)

VII.) „Korrelation“ zweier Nominalskalen (Kontingenzkoeffizient):

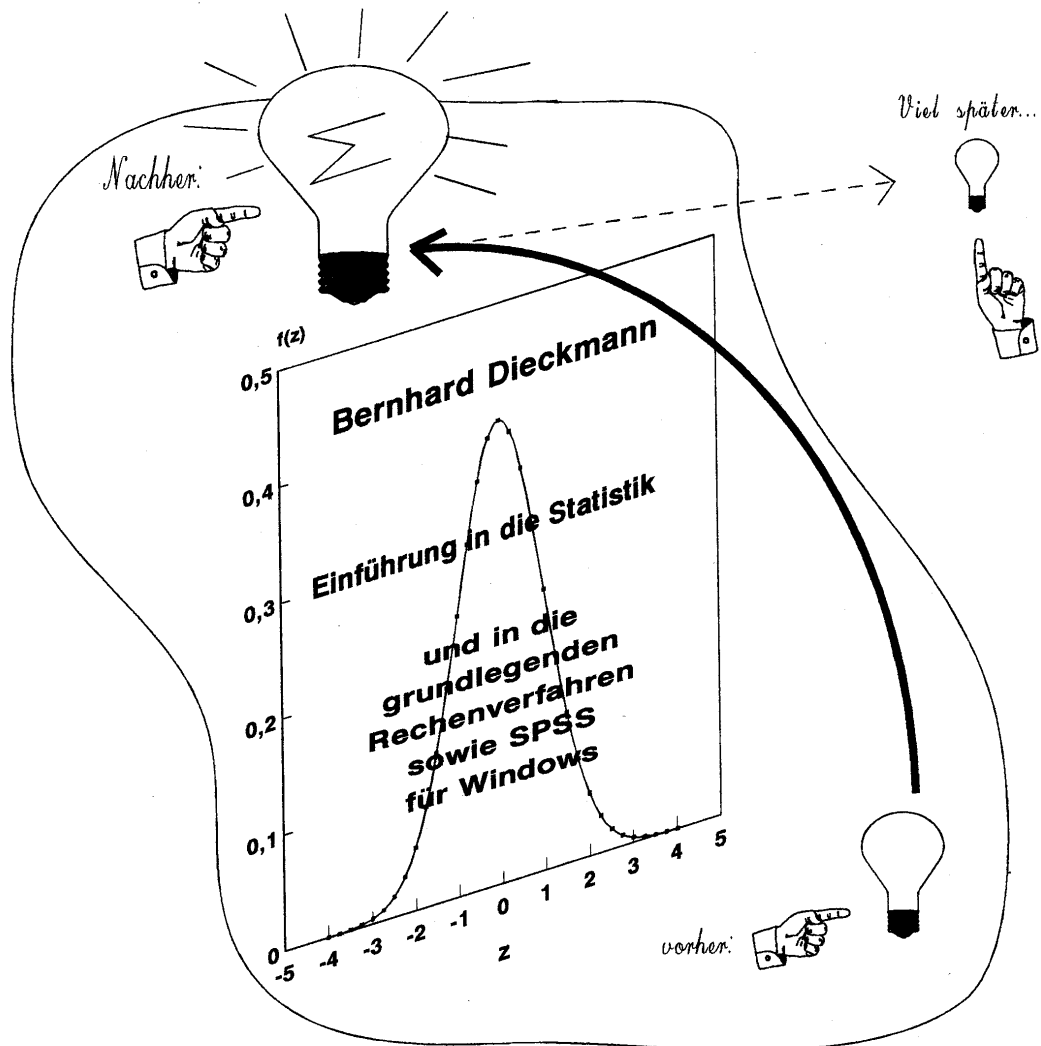
Das bekannteste Maß zur Berechnung des Zusammenhangs zweier Nominalskalen ist der Kontingenzkoeffizient C.

Dieser Test hängt eng mit dem $k \times l$ - χ^2 -Test zusammen, mit dem wir auch die H_0 überprüfen, ob zwei nominalskalierte Merkmale stochastisch unabhängig sind.

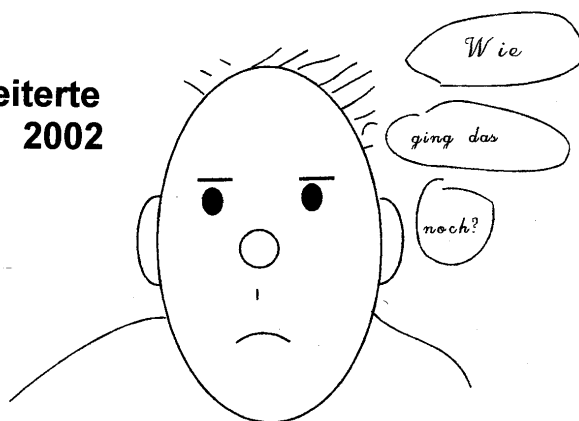
Eigentliche Formel:

$$C = \sqrt{\frac{c^2}{c^2 + n}}$$

wobei $\chi^2 = -\chi^2$ Wert des $k \times l$ - χ^2 -Tests.



6. überarbeitete, erweiterte
Auflage 2002



1. WAS IST EIGENTLICH STATISTIK?	5
1.1. Definitionsversuch für den Begriff Statistik	5
1.2. Das Problem der Verarbeitung von Massenergebnissen durch subjektive Erfahrung	6
1.3. Statistik: Unsicherheit und Ängste	7
1.4. Probleme der Lernmotivation in bezug auf Statistik	8
2. DIE FRAGE DER MEßBARKEIT: MEßNIVEAUS.....	10
2.1. Nominal-Niveau der Messung.....	10
2.1.1. Die Nominalskala.....	10
2.1.2. Die formalen Bedingungen der Klassifikation.....	11
2.2 Ordinal-Niveau der Messung.....	12
2.3 Intervall-Niveau der Messung.....	12
2.4. Ratio- oder Verhältnissniveau der Messung	13
2.4.1. Stetige Merkmale	13
2.4.2. Diskrete Merkmale	14
2.5 Das Problem der Kombination unterschiedlicher Meßniveaus	15
2.6. Qualitative und quantitative Untersuchungen	15
3. EINFACHE BESCHREIBUNGEN VON MASSENEREIGNISSEN - UNIVARIATE VERTEILUNGEN	16
3.1 Der Begriff "Verteilung"	16
3.2 Ordnung und Klassifikation von Daten – graphische Darstellungen.....	20
3.3. Zusammenfassung der Information aus Häufigkeitsverteilungen	24
3.3.1. Daten auf dem Intervall- und Ratio-Niveau der Messung - Lage- und Streuungsmaße	24
3.3.1.1 Maße der zentralen Tendenz und Lagemaße.....	29
3.3.1.2 Arithmetisches , geometrisches Mittel und harmonisches Mittel.....	30
3.3.1.3. Der Median	32
3.3.1.4 Arithmetisches Mittel und Median bei klassifizierten Werten.....	33
3.3.1.5. Weitere Lagemaße: N-tile und Prozentränge	34
3.3.1.6. Modi.....	35
3.3.1.7. Relative, prozentuale und kumulierte Häufigkeiten; Wahrscheinlichkeit.....	36
3.3.1.8. Zur unterschiedlichen Leistung der verschiedenen Mittelwerte.	41
3.3.2. Streuungsmaße: Range, Quartilabstand, durchschnittliche, absolute und quadrierte Abweichungen vom Mittelwert, Varianz und Standardabweichung.....	42
3.3.2.1. Berechnungsweise für die Varianz und die Standardabweichung nach einer vereinfachenden Formel	44
3.3.3. Daten auf dem Ordinal- und Nominal-Niveau. Lage- und Streuungsmaße	47
4. WAHRSCHEINLICHKEIT UND THEORETISCHE VERTEILUNGEN	47
4.1. Begriff Wahrscheinlichkeit	47
4.2 Theoretische Verteilungen	49
4.2.1 Allgemeines	49

4.2.2 Diskrete Ereignisse	49
4.2.3 Stetige Ereignisse.....	49
4.3. Beispiele für theoretische Verteilungen	49
4.3.1 Gleichverteilung - diskret.....	49
4.3.2 Gleichverteilung - stetig.....	50
4.4. Kombinatorik als Voraussetzung zur Herleitung von theoretischen Verteilungen	51
4.4.1. Permutationen	53
4.4.2 Kombinationen.....	54
4.4.2.1. Kombinationen ohne Zurücklegen mit Unterscheidung der Anordnung	54
4.4.2.2. Kombinationen ohne Zurücklegen ohne Unterscheidung der Anordnung.....	54
4.4.2.3 Kombinationen mit Zurücklegen unter Berücksichtigung der Anordnung.....	54
4.4.2.4 Kombinationen mit Zurücklegen ohne Berücksichtigung der Anordnung	55
4.4.2.5. Anwendung dieser Formeln auf ein Beispiel: Urne mit 6 Kugeln	55
EXKURS ZUR BERECHNUNG DER SUMME VON ZAHLENFOLGEN UND DER HÄUFIGKEIT VON WÜRFELSUMMEN; WÜRFELBEISPIELE ZUR KOMBINATORIK	61
4.5. Die Binomialverteilung.....	65
4.6. Die geometrische Verteilung	70
4.7. Die hypergeometrische Verteilung	71
4.8 Die Poissonverteilung.....	73
4.9. Die Normalverteilung und die Prüfung einer Verteilung auf Normalität	73
5. PRÜFVERTEILUNGEN.....	78
5.1. Die Chi-Quadrat-Verteilung	78
5.2 Die t-Verteilung.....	81
5.3 Die F-Verteilung.....	83
6. DER STANDARDFEHLER VON STICHPROBENKENNWERTEN	85
6.1. Allgemeines.....	85
6.2. Der Standardfehler des arithmetischen Mittels	85
6.2.1 Demonstration des Standardfehlers des arithmetischen Mittels.....	85
6.2.2. Zur Bedeutung des Standardfehlers des arithmetischen Mittels	87
6. 3. Der Standardfehler von Prozentsätzen	90
7. STATISTISCHE TESTS	90
7.1. Zur Terminologie.....	90
7.2. Zur Struktur der Statistischen Tests.....	91
7.3. Beispiele	91
7.3.1. Z-Test und t-Test auf Abweichung der Stichprobe von der Grundgesamtheit	91
7.3.1.1. Z-Test.....	91

7.3.1.2. t-Test	93
7.3.2. t-Test für die Differenz zweier Mittelwerte aus unabhängigen Stichproben aus 2 varianzhomogenen Grundgesamtheiten	93
7.3.3. Der F-Test	96
7.3.4. t-Test für die Differenz zweier Mittelwerte aus varianzheterogenen Grundgesamtheiten.....	96
7.3.5. t-Test für Mittelwertsdifferenzen aus <i>abhängigen</i> Stichproben.....	98
7.3.6. Der Binomialtest	100
7.3.7. Chiquadrat-Test.....	103
7.3.7.1. Allgemeines	103
7.3.7.2. Eindimensionales und zweidimensionales Chiquadrat-Design.....	104
7.3.7.3. Standardisierte, korrigierte Chiquadratresiduen.....	107
7.3.8. Mann-Whitney U-Test	109
7.4. Tabellarische Übersicht über kombinatorische Probleme und die Prüfverteilungen	111
8. BIVARIATE VERTEILUNGEN - ZUSAMMENHANGSDARSTELLUNGEN UND -	
MAßE	114
9. KORRELATION UND REGRESSION	114
9.1 Zwei intervallskalierte Merkmale	114
9.1.1. Die Regressionsgeraden und die Regressions-Koeffizienten.....	114
9.1.2. Der Produkt-Moment-Korrelationskoeffizient r	117
9.1.3. Ein Rechenbeispiel.....	118
9.2. Zwei ordinal skalierte Merkmale	119
9.2.1. Kruskals Gamma.....	119
9.2.2. Kendalls τ_b ("tau _b ").	122
9.3 Zwei nominal skalierte Merkmale.....	123
9.3.1 Cramér's V	123
9.3.2 Phi und korrigiertes χ^2	124
10. SIGNIFIKANZTESTS FÜR KORRELATIONS- UND ASSOZIATIONSMAßE.	125
10.1. t-Test für den Produktmomentkorrelationskoeffizienten	125
10.2. Z-Test für Kendalls "tau"	126
10.3. Einschränkungen der Chiquadrat-Tests für Cramers V und Phi.....	126
10.4 Fisher's exakter Test.....	126
11. EINFAKTORIELLE VARIANZANALYSE.....	128
12. EINFÜHRUNG IN STATISTISCHES RECHNEN	133
13. KURZE EINFÜHRUNG IN SPSS 6.0.1 UND IN DIE WICHTIGSTEN ROUTINEN.	
.....	148
13.1. Die „Fenster“ von SPSS.....	148
13.1.1. Das Datenfenster.....	149
13.1.2. Häufigkeitsanalysen als Rechenbeispiel	149
13.1.3. Das Syntaxfenster.	150
13.1.4. Das Ausgabe-Fenster und das Graphik-Karussell.....	151

13.1.4.1. Das Ausgabe Fenster.....	151
13.1.4.2. Das Graphik-Karussell.....	151
13.1.5. Übertragung von Ergebnissen aus SPSS nach WinWord.....	152
13.2. Praktische Fälle.....	153
13.2.1. Untersuchung von Mittelwertsdifferenzen zweier unabhängiger Stichproben.....	153
13.2.2. Untersuchung von Mittelwertsdifferenzen zweier abhängiger Stichproben.....	155
13.2.3. Einfache Varianzanalyse.....	156
13.2.4. Zweidimensionaler CHI-Quadrat -Test mit einer Vierfelder-Tabelle.....	158
13.2.5. Zweidimensionaler CHI-Quadrat -Test mit Kontrollvariable.....	159
13.2.6. Korrelations- und Regressionsanalyse.....	162
13.2.7. Binomialtest.....	164
13.2.8. Kolmogorov-Smirnov-Test auf Verteilungsform bei einer Stichprobe.....	165
4. TABELLEN	166
4.1. Standardnormalverteilung.....	166
4.2. χ^2 -Verteilung	168
4.3. t-Verteilung.....	169
4.4. F-Verteilung für $(1-\alpha)=0,99$	170

1. Was ist eigentlich Statistik?

1.1. Definitionsversuch für den Begriff Statistik

Fragen wir zunächst nach dem alltagssprachlichen Gebrauch:

Wenn wir

- eine **Tabelle** in einem Zeitungsartikel oder einem Buch sehen, in der Zahlen aufgeführt sind, vielleicht sogar Prozentzahlen, und wenn die einzelnen Zahlenangaben dann sogar in eine **graphische Darstellung** übersetzt werden, sprechen wir von einer "**Statistik**".

Neben diesem Sprachgebrauch gibt es aber im Wissenschaftsbereich noch weitere Bedeutungen von Statistik, die im Alltagssprachgebrauch ungewöhnlich sind:

- Statistik wird als die Bezeichnung für eine **Wissenschaft** benutzt. Diese Wissenschaft ist relativ jung. Mitte des 18. Jahrhunderts hielt Gottfried Achenwall in Göttingen Vorlesungen über Staatenkunde, einen Gegenstand, den er auch als Statistik bezeichnete. Inhalt dieser Wissenschaft waren die "wirklichen Merkwürdigkeiten einer bürgerlichen Gesellschaft" (zitiert nach Kern, Horst: Empirische Sozialforschung, Ursprünge, Ansätze, Entwicklungslinien, Beck, München 1982, S. 20). Erst später entwickelt sich die Statistik zu einer selbständigen Disziplin, die, in erster Linie auf wahrscheinlichkeitstheoretischen Grundlagen aufbauend, anstrebt, wissenschaftliche Hypothesen zu überprüfen (schließende oder induktive Statistik). Grundlagen hierfür liefern Beschreibungen von

Realität auf der Grundlage von Messungen und Zählungen (beschreibende oder deskriptive Statistik). Dabei sind die wesentlichen mathematischen Grundlagen schon Ende des 17. Jahrhunderts vorhanden.

- Im englischsprachigen Bereich wird das Wort "statistic" als ein Ausdruck benutzt, der bestimmte **Rechenergebnisse** (z.B. Mittelwerte und Streuungsmaße etc.) bezeichnet.

In diesem Kursus wollen wir unter Statistik die Wissenschaft verstehen, die sich mit der **Organisation und Auswertung von Messungen und Zählungen** beschäftigt, wenn Messungen an **vielen** einzelnen "**Objekten**" vorgenommen werden, die einander **hinreichend ähnlich** sind, und die meistens zufällig aus einer Gesamtheit einander ähnlicher Objekte ausgewählt wurden.

Wenn ich z.B. mit einer Studentin über ihre finanzielle Situation spreche und dabei ihr Einkommen erfahre, dann kann man dieses zwar schon als Meßvorgang interpretieren (es wird eine Zahl, das Einkommen in DM, der Person der Studentin (Erhebungseinheit) zugeordnet; da aber nur eine Einzelmessung vorliegt, existiert noch keine Statistik. Erst wenn ich eine größere Menge von Messungen nebeneinanderstelle, treten Probleme der Statistik auf, d.h. **Statistik hat es immer mit Mengen oder Massen zu tun**. Einzelereignisse sind für sie nicht Gegenstand der Untersuchung oder Betrachtung, sondern nur Massenereignisse.

1.2. Das Problem der Verarbeitung von Massenereignissen durch subjektive Erfahrung

Diese Unterscheidung zwischen dem einzelnen Ereignis und dem "**Massenereignis**" ist für den Alltagsgebrauch ungewöhnlich. Die persönliche Erfahrung ist in der Regel nie mit Massenereignissen konfrontiert, sondern immer nur mit einer Folge von Einzelereignissen, wobei diese Ereignisse nicht einmal ähnlich sein müssen.

Wir fahren morgens in die Universität und benutzen dabei z.B. die U-Bahn und betrachten die Umgebung der Fahrstrecke. In den Institutsräumen der Universität lesen wir Akten und Bücher. In der Mensa begutachten wir das Essen etc.: Ein Einzelereignis reiht sich an das andere. Die Gesamtheit und die Verknüpfung dieser Einzelereignisse führt dann zu so etwas wie subjektiven Erfahrungen, ohne daß hieraus jemals Massenereignisse im Sinne der Statistik werden.

Massenereignisse sind der subjektiven Erfahrung nur sehr selten zugänglich. Ein gut gefülltes Fußballstadion z.B. wäre so ein Massenereignis. Die hier dem Betrachter begegnende Masse ist nicht geordnet, hunderte oder tausende von Fußballfans sitzen oder stehen auf den Bänken in einem Fußballstadion und äußern ihre Begeisterung.

Die Eindrücke, die wir aber von solchen Massenereignissen haben können, sind sehr ungenau. Wir können etwa sagen: "Schätzungsweise waren es (so und so viele)"; "etwa 1/3 hatte Fähnchen oder andere Zeichen der Vereinszugehörigkeit dabei"; "jeder Zehnte trug ein Transparent"; "die Mehrheit war besonders bunt angezogen"; "etwa jeder fünfte Fan war stark alkoholisiert".

Alle diese Angaben beruhen auf groben Schätzungen; eine genaue Analyse des Massenereignisses "gefülltes Fußballstadion" ist aus der Perspektive des einzelnen Zuschauers in der Regel nicht möglich. Dennoch bilden sich aus der subjektiven Einschätzung individueller Erlebnisse und der ungenauen Speicherung von Eindrücken solcher Massenereignisse Urteile über das Leben, die Gesellschaft, die Menschen, über soziale Verhältnisse, soziale Normen etc., und diese subjektiven Eindrücke haben im Alltagsleben eine starke Wirkung auf Urteile. Besonders wenn wir als Mitglied einer bestimmten Fan-Gemeinde Beobachtungen anstellen, ist die Gefahr sehr groß, daß wir auf bestimmte Merkmale oder bestimmte Gruppen beispielsweise von Personen besonders achten und die dabei gewonnenen Eindrücke ohne wirkliche Kenntnis der Gesamtheit unzulässig verallgemeinern. Hier stoßen wir schon auf wichtige Unterschiede zwischen Alltagserfahrung und wissenschaftlichem Urteil.

Will man so entstandenen (Vor)urteilen argumentativ begegnen, dann ist es wichtig, solche Massenereignisse genau zu beschreiben. Hierzu bedarf es allerdings eines methodisch geregelten, nachvollziehbaren Vorgehens, das in seinen einzelnen Schritten begründbar ist. Dieses Vorgehen macht „Massen-Ereignisse“ für unsere Vorstellungskraft verfügbar, die sich ihr sonst entziehen.

Wenn wir uns zum Beispiel vorstellen, eine Nation habe eine Gesamtbevölkerung von 70 Millionen Menschen, dann sehen wir nur noch eine unübersehbare Menschenmenge vor uns, und die individuelle Fassungskraft wird von solchen Angaben in jedem Falle überfordert. Dennoch ist die arbeitsmarktpolitische Tatsache, daß Ende 1997 mehr als 4 Millionen Arbeitslose von der Bundesanstalt für Arbeit registriert worden sind, von großer Bedeutung. Politische Bedeutsamkeit ist nicht an unser Vorstellungsvermögen gebunden.

Sehr viele Verwaltungs- und politische Vorgänge verursachen solche politisch bedeutsamen Tatsachen, werden durch sie bedingt, und man kann sagen, daß bestimmte Zahlenwerke der amtlichen Statistik, insbesondere aber die Statistiken unserer Ministerien, Ergebnisse und zugleich Steuerinstrumente des politischen Lebens sind.

1.3. Statistik: Unsicherheit und Ängste

Man kann nun feststellen, daß Zahlen, **die sich der subjektiven Vorstellungskraft entziehen und gleichzeitig Steuerungsinstrument unseres politischen Lebens sind**, im öffentlichen Bewußtsein große Unsicherheit und Ängste hervorrufen.

Dies könnte u.a. folgende Ursachen haben:

Zahlen wecken zwar den **Anschein der exakten Messung**, weil wir mit ihnen in erster Linie im Rahmen des Mathematik- und Physikunterrichts vertraut gemacht worden sind. Wir sind gewöhnt, **auf Zahlen zu vertrauen**. Es macht uns unsicher und erregt Angst, wenn wir Zahlen nicht vertrauen können. Was aber heißt das, **daß wir einer Zahl vertrauen** können ?

Ein paar Beispiele hierfür:

Ich werde gefragt, wie groß ich bin. Ich stelle zwei Messungen an, eine morgens, eine abends. Morgens messe ich 1,78 m, abends 1,76 m. Ich gebe vorsichtshalber meine Größe mit 1,77 m an...

Ich frage im Supermarkt, was das Brot kostet. Ich bekomme die Auskunft, daß hier 20 Brotsorten vorrätig sind, und daß der Preis zwischen 3,50 DM und 4,50 DM pro Kilo schwankt. Bei der Konkurrenz stelle ich die gleiche Frage. Dort sagt man mir, der Preis schwanke - je nach Sorte - zwischen 3,30 DM und 4,60 DM. Was kostet das Brot nun ?

Ich will in Inflationszeiten einen Brief frankieren. Der Postbeamte sagt mir, eine Briefbeförderung koste 1000 Reichsmark. Am nächsten Tag komme ich wieder. Da kostet eine Briefbeförderung bereits 20000 Reichsmark.

Ich werde im Auto bei einer Geschwindigkeitsüberschreitung geblitzt. Mein Tacho zeigt 45 km/h. 6 Wochen später bekomme ich ein nettes Foto von mir zugeschickt, zugleich mit dem Hinweis, ich sei mit 40 km/h geblitzt worden, und davon würden noch 4 km/h (zum Ausgleich von Meßfehlern) abgezogen. Man werfe mir daher die Überschreitung der Höchstgeschwindigkeit von 30 km/h um 6 km/h vor.

Alle diese Beispiele zeigen Situationen, in denen auf eine Zahl aus verschiedenen Gründen nicht ganz vertraut werden kann, sondern nur ein bißchen. Die Gründe können sehr verschieden sein. Es lohnt sich, sie zu analysieren: Hier waren die Gründe rhythmisches Schwanken, Vielfalt verschiedener Ereignisse in zwei Gruppen, schnelle Änderung eines Zustandes, Verwendung unterschiedlicher Meßinstrumente und Definitionen.

Wir merken uns daher, daß wir auf Zahlen nur vertrauen dürfen, wenn wir sicher sein können, die gleichen **Definitionen** für die bezeichneten Tatbestände (Erhebungseinheit, Dimension) zu verwenden wie unsere Kommunikationspartner und auch mit den gleichen Meßinstrumenten arbeiten. Zudem müssen die Zustände, die wir mit Zahl beschreiben, zeitlich stabil sein, wenn wir künftig die gleichen Zahlen verwenden wollen, oder die Prozesse müssen wiederholbar sein, deren Ablauf wir beschrieben haben.

Unter den oben genannten Bedingungen ist die Bedingung der gleichen Definition die wichtigste und in der sozialwissenschaftlichen Praxis zugleich die schwierigste. Wie jeder Messung liegt jeder empirisch bedeutsamen Zahl eine Reihe von Definitionen zugrunde, durch die sie zustandekommt. Bei sozialen Tatsachen sind diese **Definitionen in der Regel kontrovers**, da unterschiedliche Interessen am Ergebnis der Definition bestehen. Wird z. B. im Fernsehen um Zahlen gestritten, handelt es sich oft um einen Streit über die Wahl der politisch jeweils "nützlichen" Definition. Von solchen Feinheiten versteht natürlich der Fernsehzuschauer nichts, und er denkt, man stritte über Tatsachen, nicht aber über Begriffe. In seinem Gedächtnis bleibt der Streit über die Zahl erhalten, nicht aber der Definitionsstreit.

Gewöhnlich liegt jeder größeren statistischen Erhebung und natürlich auch jeder Stichprobe eine Reihe verschiedenartigster **Fehler** (z.B. Meß- und Übertragungsfehler, Fälschungen) zugrunde, deren Ausmaß nur selten genau bestimmt werden kann. Von daher sind die Verwaltungen und Regierungen nicht selten in der dummen Situation, sich auf Zahlenmaterial stützen zu müssen, das sehr wenig zuverlässig ist. Politische Fehlentscheidungen sind aus diesem Grunde gelegentlich statistisch bedingt, und man gibt leicht der Statistik die Schuld für politisches Versagen. Dennoch tun die Verwaltungen in der Regel so, als hätten sie verlässliche Daten. Das gehört zu ihrem Prestige.

Bestimmte statistische Angaben bekommen gelegentlich hohen Symbolwert: Wenn z. B. die Zahl der Arbeitslosen über 4 Millionen ansteigt, der Benzinpreis höher als 1,00 EURO pro Liter liegt oder sogenannte Maastricht-Kriterien über drei Prozent wachsen. Dies wissen die Parteien natürlich, und sie tun alles, um die eigenen politischen Ergebnisse nach Möglichkeit zu beschönigen oder zu verstecken. Die Kunst der Fälschung statistischer Angaben durch ihre Frisierung vor der Veröffentlichung ist hoch entwickelt, mindestens genauso hoch wie die Abneigung des normalen Zeitungslesers gegenüber einer tabellarischen Darstellung technischer oder sozialer Sachverhalte. Das schöne Bild einer Grafik, das **vermeintlich schnellen Überblick** verschafft, wird der zahlenmäßigen Einzelangabe, die dann auch noch Absolut- und Prozentzahlen nebeneinanderstellt, in der Regel vorgezogen. Der schnelle Leser fragt nicht nach Definitionen, Bezugsgrößen für Prozentzahlen. Er will nicht mühsam rechnen, vergleichen, analysieren, sondern er will unterhalten sein und hält das auch noch für Information. Vor allem will er keine Angst haben müssen. Er will durch Zahlen "**beruhigt**" werden. Statt Information oder Entertainment nennt man diese beruhigende Unterhaltung neuerdings "**Infotainment**".

Wenn sich angesichts dieser Probleme und Ängste die Statistik auf komplizierte Definitionen, wissenschaftliche und politische Erörterungen der Erhebungskonzepte und noch kompliziertere Rechnungen stützt, dann bekommt sie - vor allem auch an den Universitäten - in der Regel den Beigeschmack einer **Geheimwissenschaft**, die nichts mit dem Leben zu tun hat.

Das Verdikt der Alltagssoziologie, Statistik sei falsch, mechanistisch, borniert, patriarchalisch, Produkt eines falschen Bewußtseins und was der Plattitüden mehr sind, ist mit rationalen Argumenten in der Regel nicht zu entkräften, weil – wie Fachleute gelegentlich erfahren müssen - die Kritiker oft gar nicht wissen, was sie selber mit der Kritik meinen und von Statistik zu wenig verstehen, um ihre eigene Meinung zu belegen. Die meisten Kritiker werden sich erst mit der Statistik anfreunden, wenn sie ihrer selbst bedürfen.

1.4. Probleme der Lernmotivation in bezug auf Statistik

Die schnelle Ablehnung der Statistik und des Statistiklernens ist also verständlich wegen der **zahlreichen Negativerlebnisse**, die das Alltagsleben dem Benutzer und dem der Statistik Unterworfenen anbietet.

Werde ich in der Volkszählung zahlenmäßig erfaßt, dann muß ich damit rechnen, daß die Informationen über mich nicht unbedingt in dem Sinne benutzt werden, wie ich mir das vorstelle.

Der Verwendungszusammenhang meiner Daten ist - von bestimmten Verwendungsverboten abgesehen - unklar; wer welche Daten wann verwendet, ist unbekannt, und wehren kann ich mich gegen einen abstrakten Gegner kaum noch.

Andererseits möchte ich als politisch aktiver Bürger gern zuverlässige Informationen über Sozialstrukturdaten haben, mit denen ich meine oder meiner Gruppe Rechte öffentlich einfordern kann. Ich möchte beispielsweise nachweisen können, in welchem Maße Frauen in bestimmten Ämtern und Berufsbereichen überrepräsentiert oder unterrepräsentiert sind, ich möchte die Menge der Sozialhilfeempfänger und das Ansteigen dieser Menge wissen, um gegen unsoziale politische Praktiken argumentieren zu können.

Weiterhin möchte ich wissen, in welchem Maße sich der Lebenshaltungsindex, die Luftbelastung, der Ölpreis etc., verändern.

Wir können also sehen, daß die Nachfrage nach Informationen nicht nur auf Seiten des Staates, sondern auch auf Seiten des informationsbedürftigen Staatsbürgers vorhanden ist, so daß eigentlich die Bereitschaft bestehen müßte, statistische Auskünfte zu geben, weil man sie ja auch selbst haben will.

Andererseits steht im Zeitalter leistungsfähiger Datenverarbeitungsanlagen die Ohnmacht des einzelnen der Statistik unterworfenen Menschen als immer gewichtigeres Argument im Raum, wenn über Statistik, das Lernen von Statistik und die Benutzung von Statistiken nachgedacht wird. Nur wird hierbei nicht berücksichtigt, daß es für relativ kleine Organisationen und Vereine immer leichter wird, einen eigenen statistischen Apparat aufzubauen, mit dessen Hilfe auch Argumente zur Kritik offizieller Statistiken produziert werden können, indem eigene Stichprobenuntersuchungen zur Kontrolle politisch gefärbter Aussagen angestellt werden.

Mit allen diesen eher soziologischen und politologischen Fragen der Benutzung von Information in Wirtschaft, Bürokratie und Politik **hat sich die theoretische Statistik** nur am Rande **beschäftigt**. Mit dieser Enthaltensamkeit steht sie nicht allein, sondern sie teilt diese Enthaltensamkeit generell mit allen Naturwissenschaften und technischen Wissenschaften, deren Interesse in erster Linie die unmittelbare Sache ist, um deretwillen sie forschen.

Die Folgen der Anwendung naturwissenschaftlicher und technischer Forschungsergebnisse sind in der Regel nicht Gegenstand der gleichen Naturwissenschaften, und sie entziehen sich auch in der Regel ihrem methodischem Instrumentarium.

Wir können diese Tatsache zwar beklagen, aber nicht kurzfristig ändern. Anfänger der Statistik, besonders in pädagogischen und psychologischen Forschungseinrichtungen, erliegen oft der Versuchung, die einschlägige mißbräuchliche Benutzung der Statistik, für die der Nachweis aus so ziemlich jeder Tageszeitung zu erbringen ist, als Argument gegen das Erlernen von Statistik ins Feld zu führen.

Es entstehen in Anfängerseminaren dann stundenlange und meist fruchtlose Diskussionen darüber, daß das eigentliche Lernen von Statistik darin bestehen müßte, zu wissen, wie der Mißbrauch funktioniert und wie man ihn verhindern könnte. Die Statistiker selbst geben hierauf nur teilweise befriedigende Antworten, im übrigen müßten sie darauf verweisen, daß diese Themen Gegenstand anderer Wissenschaftsdisziplinen sind, insbesondere der Juristerei (siehe Wirtschaftskriminalität), der Politologie, der Verwaltungswissenschaft, der Soziologie, der Zeitungswissenschaften etc.

Alles dieses müßte man im Grunde mitstudieren, wenn man einen hinreichend vollständigen Eindruck von der Bedeutung konkreter statistischer Angaben haben wollte und vom konkreten Umgang mit ihnen.

Da das Erlernen von Statistik leider all diese Aspekte traditionell nicht enthält und auch wegen der geringen Zeit, die meistens für diesen Stoff zur Verfügung steht, nicht enthalten kann, soll in diesem folgenden Kursus nur der Versuch gemacht werden, die offensichtlich naheliegenden Mißbrauchsmöglichkeiten zu benennen und einige Grundsätze ständig mitzubedenken, die einen seriösen Gebrauch von Statistik möglich machen.

Der **Zweck** des Statistiklernens kann darin bestehen, daß wir in der Lage sind, aufgrund entsprechender Lernprozesse einerseits **statistische Veröffentlichungen sehr kritisch zu lesen** und sehr schnell jene Momente politisch beabsichtigter Ungenauigkeit identifizieren zu können, die uns dann gestatten, kritische Argumente zielbewußt aufzugreifen.

Ein weiterer Zweck des Statistiklernens kann darin bestehen, die Studierenden auf **eigene wissenschaftliche Arbeiten vorzubereiten, in denen ein wissenschaftlich kontrollierbarer Umgang mit Statistik gefordert wird**.

Statistische Methoden dringen in den Sozialwissenschaften parallel zu qualitativen Forschungsmethoden immer weiter vor. Ein Gegensatz zwischen qualitativen und quantitativen Forschungsmethoden wird in jüngerer Zeit immer mehr verneint. Die Forderung der Methodiker richtet sich zunehmend auf eine Vermittlung und Verbindung zwischen statistischen und qualitativen Ansätzen der Forschung.

Einzelfallanalyse und die Analyse von Massenereignissen müssen immer theoretisch aufeinander bezogen werden. Das eine kann das andere nicht ersetzen, und der eine Ergebnistyp liefert nur unvollständige Aspekte der Wirklichkeit ohne Hinzuziehung des anderen.

Kritisches Lesen und selbständiges handwerkliches Beherrschen statistischer Methoden zum Zwecke eigener wissenschaftlicher Arbeit sind also die wesentlichen Zielsetzungen des Statistiklernens. Diesen Zielsetzungen muß im Bewußtsein von dem gesellschaftlichen Mißbrauch und von den gesellschaftlichen Gefahren der Statistik nachgegangen werden.

Wir können nicht auf das Statistiklernen deshalb verzichten, weil es schwierig ist, und weil Statistik eine Wissenschaft ist, deren Ergebnisse im Zentrum politischer Kontroversen stehen. Im Gegenteil: Wenn wir an der politischen Gestaltung unserer Öffentlichkeit beteiligt sein wollen, ist es geradezu zwingend, daß wir in diesen Bereich systematisch eindringen.

2. Die Frage der Meßbarkeit: Meßniveaus

In Alltagsgesprächen hört man öfters das Argument: „Das ist nicht vergleichbar“, das ist nicht „meßbar“. Sind solche Argumente triftig? Wann - und wann nicht? Grundsätzlich lassen sich nämlich alle Gegenstände unseres Bewußtseins miteinander vergleichen, es kommt hierbei nur darauf an, ob sie **in bezug auf eine Theorie interpretiert werden können, für die sie relevant sind**. So kann man in bezug auf eine Theorie der Beutemachens durchaus Hunde und Katzen vergleichen, in bezug auf eine Theorie des Fliegens durchaus eine Hummel, einen Seeadler und eine Boeing 707. Die subjektive Feststellung: „Das ist nicht vergleichbar“ könnte daher meinen: „Mir fällt keine Theorie ein, die für beide Gegenstände relevant ist, die ich vergleichen möchte. Daraus ist aber nicht zu schließen, daß es keine solche Theorie geben könnte.“

Wie ist das nun mit dem Messen? In der Alltagssprache spreche ich dann von Meßvorgängen, wenn ich aus der Physik bekannte Maßstäbe einsetze, wie Meter (Distanz), Liter (Volumen), Grad (Temperatur), km/h (Geschwindigkeit). Man sagt: „Das Bild ist 30 cm breit und 45 cm lang“, wenn man das Format ausmißt. Sage ich aber: „Das Bild ist sehr schön!“, oder: „Das ist ja ein eigenartiges Grün auf diesem Bild!“, dann würde man - alltagssprachlich - kaum noch von Meßvorgängen sprechen. Dennoch haben die beiden letztgenannten Feststellungen etwas in sich, das in den Sozialwissenschaften mit Messung bezeichnet wird. Bei dem „Grün“ wird dem Bild ein Element einer „Farbskala“ zugeordnet, bei dem „sehr schön“ ein Element aus einer Wertskala. Messung bedeutet in den Sozialwissenschaften, daß ich Zahlen oder Symbole, die ein Element einer Skala sind, nach Maßgabe einer Meßvorschrift, einem Objekt zuordne. Würde ich beispielsweise innerhalb eines Forschungsprojekts eine Farbtabelle in die Hand bekommen, und müßte ich entscheiden, welches Grün der Farbtabelle auf dem Bild ist, oder müßte ich mehrere Bilder, die mir im Rahmen eines Forschungsprojekts vorgelegt werden, nach meinem Schönheitsempfinden in eine Rangordnung bringen, dann kann es sein, daß aus Urteilen, die ich im Alltagsleben fälle, „Messungen“ werden, weil diese Urteile in einen sozialwissenschaftlichen Zusammenhang hineingenommen werden. Hierbei beziehen sich solche Urteile immer auf Positionen einer Skala. Wichtig ist für die sozialwissenschaftliche Messung nun, welche mathematischen Eigenschaften solche Skalen haben („Meßniveau“). Mit welchen mathematischen Methoden man Daten weiterverarbeiten darf, die auf Skalen definiert sind, hängt nämlich vom „Meßniveau“ der entsprechenden Skala ab. Man unterscheidet hier „Nominalniveau“, „Ordinalniveau“, „Intervallniveau“ und „Ratio-Niveau“.

2.1. Nominal-Niveau der Messung

2.1.1. Die Nominalskala

Wie wir schon aus der Alltagserfahrung wissen, können Unterschiede zwischen verschiedenen Menschen, Menschengruppen, Institutionen, sozialen Prozessen - was auch immer Gegenstand der sozialen Forschung sein mag - verschieden genau bestimmt werden. Die Genauigkeit der Bestimmung von Unterschieden zwischen Gegenständen der empirischen Sozialforschung wird mit Hilfe des Begriffs "Meßniveau" charakterisiert. Das Nominal-Niveau der Messung liegt dann vor, wenn nur gesagt werden kann, daß zwei Elemente einer Menge einander gleich oder ungleich sind. Die mathematischen Symbole hier für sind = (gleich) bzw. \neq (ungleich).


Frauen sind in bezug auf ihr Geschlecht Frauen gleich. Sie sind den Männern in bezug auf ihr Geschlecht ungleich. Deutsche sind Deutschen in bezug auf ihre Nationalität gleich. Sie sind Engländern in bezug auf ihre Nationalität ungleich.


Es ist auf der Nominalskala **nicht möglich**, Aussagen zu formulieren wie: Frauen **haben mehr Geschlecht** als Männer, sondern nur: Sie haben ein anderes Geschlecht. Es ist auch sinnlos zu sagen: Engländer haben **doppelt soviel Nation** wie Amerikaner.


Man darf auf Nominalskalen nur sagen: Die X sind in bezug auf Z (Skalename) den Y gleich oder ungleich.


2.1.2. Die formalen Bedingungen der Klassifikation

Allen Klassifikationen können (neben anderen Skalen) Nominalskalen zugrundeliegen. Unabhängig von Meßniveau gelten bei allen Klassifikationen die „formalen Bedingungen“ der Klassifikation „Eindeutigkeit, Vollständigkeit“ und „Echtheit“. Die folgenden Ausführungen sollen diese Begriffe erläutern. Nehmen wir an, wir wollten die folgenden sechs schraffierten Flächen klassifizieren (einer Skala zuordnen),


 (Schraffur: von links rechts fallend)

 (Schraffur: von links nach rechts steigend)

 (Schraffur senkrecht)

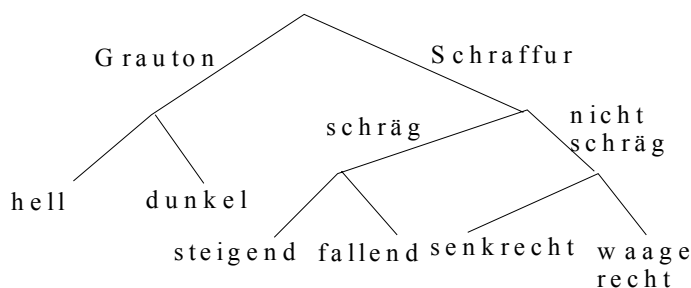
 (Schraffur waagerecht)

 (hellgrau)

 (dunkelgrau)

dann müßte unsere Skala **alle** diese Eigenschaften vorsehen („Prinzip der Vollständigkeit“), sie müßte es gestatten, diese Eigenschaften **eindeutig** zuzuordnen („Prinzip der Eindeutigkeit“), und sie müßte bei der Bildung von Unterklassen möglichst so vorgehen, daß mindestens zwei der gebildeten Unterklassen auch besetzt sind (Prinzip der Echtheit). Eine Klassifikation, die diese Eigenschaften erfüllen würde, wäre die folgende:

Eindeutige und vollständige Klassifikation der oben aufgeführten Flächen:



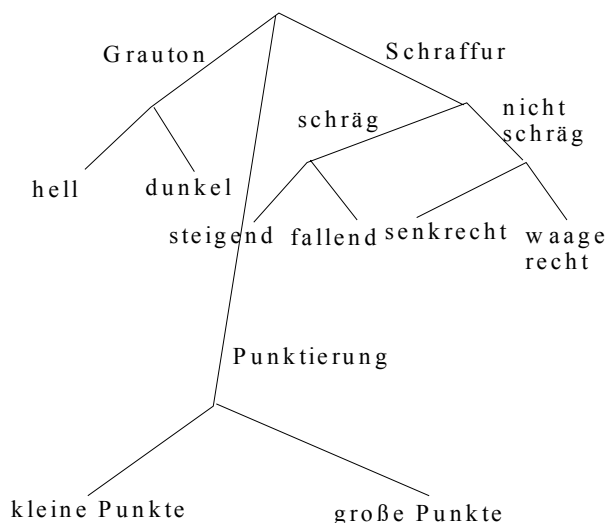
Hätten wir in der Klassifikation nur Schraffuren vorgesehen und keine Grautöne, könnten wir diese nicht zuordnen (die Klassifikation wäre **unvollständig**). Würden wir bei den schräg schraffierten Flächen nicht sagen, daß steigend und fallend **immer von links nach rechts** „steigend“ oder „fallend“ bedeutet, wäre die Klassifikation **uneindeutig**.

Hätten wir noch **eine** weitere **gepunktete Fläche** (die wir oben in der Graphik nicht haben),



dann wäre die Klassifikation auf der 1. Ebene (Grauton, Schraffur, **Punktierung**) entsprechend zu ergänzen. Gäbe es dort keine Differenzierung zwischen verschiedenen Punktierungen, dann wäre die Klassifikation dann

unecht, wenn unter Punktierung z. B. noch die Klassen „große Punkte“ und „kleine Punkte“ gebildet worden wären. Mindestens eine dieser Unter-Klassen wäre dann leer.



In der Praxis der Statistik sollten die Kriterien der Eindeutigkeit und der Vollständigkeit erfüllt sein, das Kriterium der Echtheit ist nur selten erfüllt, da in statistischen Tabellen oft leere Klassen vorkommen. Aber auch das Kriterium der Vollständigkeit wird oft verletzt, wenn bei Umfragen neben anderen Kategorien einer Frage noch eine Kategorie „sonstige“ vorgesehen ist. Dann werden die hierin auftauchenden Antworten nicht in den anderen, zuvor gebildeten Kategorien vollständig erfasst.

2.2 Ordinal-Niveau der Messung

Auf dem Ordinal-Niveau der Messung kommt zu der Gleichheits- oder Ungleichheitsrelation noch die Größer-Kleiner-Relation hinzu. Die mathematischen Symbole hierfür sind: $a > b$ (a ist größer als b), bzw. $a < b$ (a ist kleiner als b). Relation wird hier als aus einem Vergleich gewonnene Beziehung zwischen zwei oder mehreren Elementen einer Menge definiert, die durch ein Zeichen, wie z.B. das Gleichheitszeichen (=), oder das "Größer-als"-Zeichen (>) ausgedrückt werden kann. Vor allem im Bereich von nicht sehr genau angebbaren "Einschätzungen" taucht in der Sozialforschung das Ordinal-Niveau der Messung auf. "Ich ziehe das Musikhören dem Statistiklernen vor", wäre eine solche Relation. Um wieviel lieber mir das Musikhören als das Statistiklernen ist, werde ich kaum angeben können. Wenn ich im Sommer noch lieber Schwimmen gehe als Musik höre, und lieber Musik höre als Statistik lerne, dann hätte ich schon drei Elemente, die ich in eine Rangordnung bringen kann, ohne daß ich angeben kann, um wieviel mir das eine sympathischer ist als das andere. Zwischen den ranggleichen Elementen gilt die Relation gleich (=). Zwischen den rangverschiedenen Elementen gilt neben der Relation $<$ oder $>$ immer gleichzeitig die Relation ungleich (\neq). Insofern kommen auf der Ordinalskala zu den Relationen = und \neq noch die Relationen $<$ und $>$ hinzu, die auf der Nominalskala nicht definiert sind.

2.3 Intervall-Niveau der Messung

In dem Moment, in dem ich Messungen anstelle, die es mir erlauben, **Abstände** (Intervalle) zwischen den verschiedenen Meßpunkten zu benennen, bewege ich mich auf dem Intervall-Niveau der Messung. Sehr viele Tätigkeiten unseres Alltagslebens enthalten Merkmale, die auf diesem Meßniveau definiert sind: Die Geburtstage, die in jedem Ausweis eingetragen sind, kann ich nebeneinander stellen und ihren Abstand in Tagen, Monaten und Jahren berechnen.

Das Wetter, das ich am Morgen beim Weg zur Arbeit erlebe, wird durch Temperatur und Luftfeuchtigkeit (neben anderen Daten) beschrieben. Diese Angaben sind auf dem Intervall-Niveau definiert, d.h. daß Abstände zwischen den verschiedenen Meßpunkten angegeben werden. Unsere Vorstellungskraft ist auch nicht überfordert, wenn wir die Information verarbeiten sollen, daß es morgen um 5 Grad Celsius kälter sein wird als heute. Haben wir aber heute bereits eine Temperatur von - sagen wir - minus 10 Grad Celsius, dann wäre die Angabe, daß es morgen halb so warm sein wird wie heute, allerdings für uns kaum noch interpretierbar, denn die Hälfte von minus 10 Grad wären minus 5 Grad, und das wäre nicht halb so warm, sondern wärmer.

Auf der Temperaturskala, die eine Intervallskala ist, sind eben nur Abstände, aber keine Quotienten berechenbar. Habe ich Abstände, dann kann ich Additionen (+) und Subtraktionen (-) ausführen, das heißt, auf dem Zahlenstrahl, der von links nach rechts steigend definiert ist, um einen bestimmten Betrag nach rechts gehen (Addition) oder nach links gehen (Subtraktion).

Quotienten, Ergebnisse von Divisionen, können nur dann berechnet werden, wenn ein absoluter Nullpunkt der Skala existiert, auf den man sich bei konkreten Rechnungen bezieht.

Auf der Intervallskala kommen zu den Relationen $=$, \neq , $>$, $<$ jetzt die Möglichkeiten der Addition und der Subtraktion hinzu. Divisionen und Multiplikationen sind nicht sinnvoll ausführbar.

2.4. Ratio- oder Verhältnissniveau der Messung

Angaben, mit denen physikalische Größen gemessen werden, sind in der Regel auf dem Ratio- oder Verhältnissniveau definiert. Für einen Physiker ist es nämlich durchaus möglich, wenn er bei der Temperatur vom absoluten Nullpunkt ausgeht (auf der Kelvin-Skala wird der Wert $-273,16^\circ$ Celsius als 0 definiert), solche Aussagen wie doppelt so warm, halb so warm etc. sinnvoll zu interpretieren.

Das Ratio- oder Verhältnissniveau der Messung liegt also bei allen Längenmaßen und Gewichten vor, für die eine feststehende Vergleichsnorm definiert ist. In Statistik-Programmen, in denen man das Meßniveau einer Variable wählen kann, wird das Ratio-Niveau in der Regel als „**metrisch**“ bezeichnet.

Zu den Relationen und Operationen, die auf dem Nominal-, dem Ordinal- und dem Intervallniveau der Messung definiert sind ($=$, \neq , $>$, $<$, $+$, $-$) kommen jetzt noch die Operationen der Multiplikation und der Division hinzu (mathematische Zeichen: $*$ und \div). In der Sprache mathematischer Formeln werden diese Symbole allerdings oft durch andere ersetzt. Statt „ $a * b$ “ wird „ ab “ geschrieben, statt „ $1 \div 2$ “ wird auch „ $1:2$ “ oder „ $\frac{1}{2}$ “ geschrieben, d.h. die Nebeneinanderstellung oder der Bruchstrich ersetzen die Symbole $*$ und \div .

Nominal-Niveau, Ordinal-Niveau, Intervall-Niveau und Ratio-Niveau der Messung stehen untereinander nach dem oben Gesagten in einer Beziehung **aufsteigender Informationsgehalte**; das Nominal-Niveau hat den geringsten Informationsgehalt, das Ratio- oder Verhältnissniveau den höchsten Informationsgehalt. Für das jeweils höhere Meßniveau gelten alle Eigenschaften des geringeren Meßniveaus ebenfalls.

Die Bedeutung der Unterscheidung dieser Meßniveaus für die Statistik besteht darin, daß wir bestimmte Berechnungen z.B. in der Statistik nur anstellen dürfen, wenn das für solche Berechnungen geeignete Meßniveau gegeben ist. So sind die Ausdrücke: „2. Sinfonie von Beethoven und „1. Sinfonie von Schubert“ auf dem Nominalniveau definiert. Wir würden daher nicht auf die Idee kommen, sie zu addieren und aus beiden vielleicht die 3. Sinfonie von Bruckner zu machen.

2.4.1. Stetige Merkmale

Auf einem Intervall- und dem Ratio-Niveau der Messung werden stetige und diskrete Merkmale unterschieden. Stetige Merkmale sind solche, für die im Prinzip jeder beliebige Meßwert im gesamten Meßbereich auftreten kann. So können wir auf dem Thermometer, wenn es nur hinreichend genau geeicht ist, Werte wie 37,5; 38,9 ablesen. Moderne Stoppuhren erlauben die Messung der Zeit bis hin zur Tausendstel-Sekunde. Mit weiterer Steigerung der Genauigkeit werden sämtliche möglichen Zwischenwerte im Prinzip ablesbar, so daß mathematische Funktionen, die Beziehungen zwischen zwei solchen Messungen darstellen, als nicht unterbrochene, gerade oder kurvenförmige Linie dargestellt werden können, deren Wertepaare vollständig

definiert sind. Das hat zur Folge, daß man die Steigung solcher Kurven in jedem Punkt feststellen kann, das heißt, man kann sie „differenzieren“ (sie sind für die Differentialrechnung geeignet).

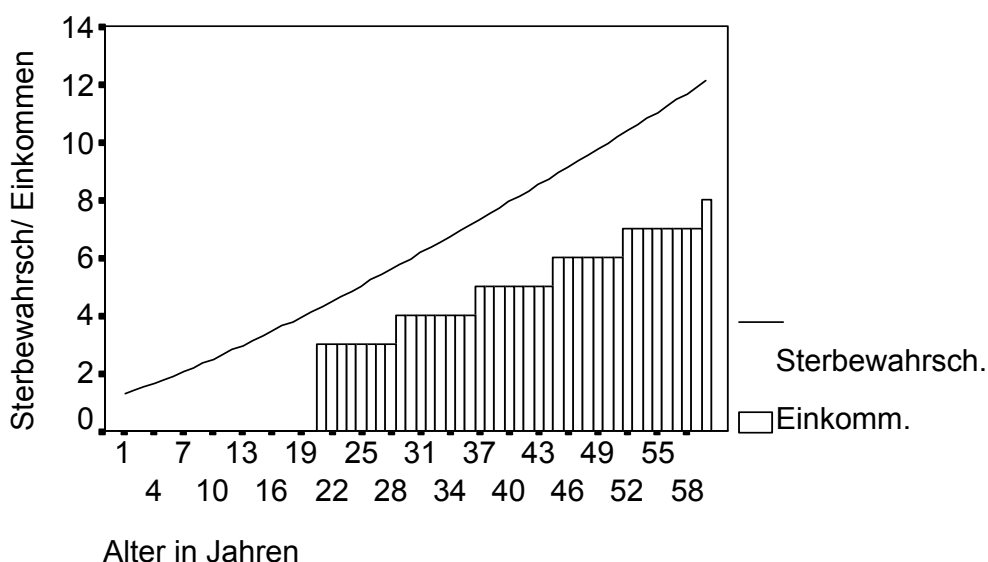
2.4.2. Diskrete Merkmale

Haben wir es bei Temperaturmessungen nicht mit einem Fieberthermometer oder einem sehr genauen Fensterthermometer zu tun, sondern blicken wir auf ein normales Zimmerthermometer, dann sind in der Regel nur Angaben lesbar, die die vollen Grade in Celsius umfassen. Ähnlich ist es mit Zeitangaben: eine Bahnhofsuhr, die keinen Sekundenzeiger hat, springt von Minute zu Minute weiter, so daß nur volle Minutenangaben möglich sind, Zwischenwerte aber ausgeschlossen bleiben. Angaben im Ausweis über das Lebensalter beziehen sich auf den Tag der Geburt, lassen aber genauere Angaben fort. Fragen wir im Fragebogen nach dem Geburtsdatum, dann müssen wir uns in der Regel mit Angaben über das Geburtsjahr bescheiden. Alle Prozesse des Abzählens ganzer Einheiten können nur ganzzahlige Ergebnisse haben, gebrochene Zahlen kommen nicht vor.

Solche Merkmale, die zwar auf dem Intervallniveau definiert sind, jedoch nur ganz bestimmte Werte annehmen können, heißen diskret, was nichts mit "Diskretion" zu tun hat, sondern als statistischer Fachterminus die Tatsache bezeichnet, daß innerhalb einer Meßstrecke nur ganz bestimmte Zwischenwerte (meistens ganze Zahlen) auftreten können. Die graphische Darstellung darf daher im Prinzip bei diskreten Merkmalen keine durchgezogenen Linien oder Kurven verwenden, sondern müßte sich mit der Angabe von Meßpunkten begnügen. Streng genommen lassen sich für diese Meßpunkte auch keine Steigungen angeben, d.h. die Differentialrechnung dürfte nicht angewendet werden.

Die folgende Graphik zeigt zwei Variablen in Abhängigkeit vom Alter: Eine stetige Variable, die Sterbewahrscheinlichkeit, ausgedrückt in Promille/Jahr und eine diskrete Variable, das Einkommen, ausgedrückt in Tausend DM im Monat. Es handelt sich hierbei um ausgedachte Daten. Die Sterbewahrscheinlichkeit wachse hier stetig, das Einkommen wachse von Gehaltserhöhung zu Gehaltserhöhung in einer „Sprungfunktion“, die nicht alle möglichen Meßwerte annimmt:

Alter, Sterbewahrscheinlichkeit und Einkommen in der Kleinstadt Durchschnittshausen.



2.5 Das Problem der Kombination unterschiedlicher Meßniveaus

In der Praxis treten nun leider die verschiedenen Meßniveaus kombiniert auf. Zum Beispiel die Aussage: „Frauen haben im Durchschnitt ein geringeres Einkommen als Männer“ enthält zwei Meßniveaus: das Nominal-Niveau (bei dem Begriff Frauen) und das Ratio-Niveau (bei dem Begriff „Einkommen“). Die Aussage: "Wirtschaftsmanager denken pragmatischer als Universitätsprofessoren" enthält das Nominal-Niveau bei der Unterscheidung zwischen Managern und Professoren und das Ordinal-Niveau bei dem Ausdruck "pragmatischer". Die Berücksichtigung bei der jeweiligen Kombination von Meßniveaus in der Statistik führt zu einer Fülle verschiedener Berechnungsmethoden von "statistischen Zusammenhängen", die sich alle nur deshalb unterscheiden, weil die vorausgesetzte Kombination von Meßniveaus unterschiedlich ist. Die in der untenstehenden Tabelle aufgeführten Zusammenhangsmaße sind eine Auswahl der wichtigsten Zusammenhangs-Berechnungsmethoden, von denen die meisten in dieser Schrift dargestellt sind:

Meßniveau/Rechenverfahren für die Darstellung statistischer Zusammenhänge	Nominal	Ordinal	Intervall/Ratio
Nominal	Chiquadrat-Test, Cramér's V, Phi, Fischers exakter Test	Kruskal-Wallis-Test	t-Test Varianzanalyse
Ordinal		Mann-Whitney –U-test, Kendalls tau, Kruskals gamma	
Intervall/Ratio			Varianzanalyse, Regressions- und Korrelationskoeffizienten

Eine ausführlichere Auflistung solcher Techniken findet man in Sidney Siegel: Nonparametric Statistics, McGraw-Hill, New York 1956.

2.6. Qualitative und quantitative Untersuchungen

Im Alltagssprachgebrauch wird nicht zwischen verschiedenen Meßniveaus unterschieden, sondern das Begriffspaar qualitativ/quantitativ steht in der Regel an der Stelle einer solchen Unterscheidung. Die Begriffe Qualität und Quantität enthalten aber noch mehr an Bedeutung als die rein informationstheoretische Unterscheidung zwischen Meßniveaus. Wir wissen alle, daß mit "Qualität" auch ein Werturteil ausgesprochen wird: Im Alltagssprachgebrauch wird - besonders in einer der „Psychoscene“ nahestehenden Sprache - zwischen einer mehr "kopfbetonten" Analyse und einer mehr "bauchbetonten" unterschieden. Hiermit wird in der Regel nicht nur wertfrei unterschieden, ob mehr der Verstand oder das Gefühl bei bestimmten Urteilen maßgebend war, sondern die mehr "bauchbetonten" Urteile werden in bestimmten Kreisen auch gegenüber der rein durch Argument und Erfahrung angeleiteten Erkenntnis als "höherwertig" angesehen.

Dieser Wertcharakter des Wortes "qualitativ" spielt auch im Sprachgebrauch von Marx eine Rolle, wenn dort vom "Umschlag der Quantität in die Qualität" gesprochen wird. **Diese sehr bildliche („metaphorische“) Sprechweise darf auf keinen Fall verwechselt werden mit einer Sprechweise, die wir in den empirischen Sozialwissenschaften benutzen.** Dies bedeutet nicht, daß wir uns nicht mit der Bedeutung von Metaphern beschäftigen und sie nicht zum Gegenstand qualitativer Analyse machen dürfen, um sie besser verstehen zu können. Es ist aber ein großer Unterschied zwischen der wissenschaftlichen Benutzung und der Analyse von Metaphern. In der Sprache der Statistiker wäre ein Ausdruck wie "Umschlag der Quantität in Qualität" sinnlos. Sinnvoll hingegen wäre z.B. die Beschreibung eines sozialen Prozesses, in dem Frauen sich zum gewaltlosen Widerstand gegen militärische Hochrüstung entschließen, nachdem die Rüstungsausgaben in einem Staat über eine bestimmte Marke geklettert sind. In diesem Fall würde ein Sozialwissenschaftler oder Statistiker zwei zeitlich verbundene Prozesse feststellen: Eine Veränderung der Rüstungsausgaben in der Zeit (Ratioskala) und eine Veränderung der Einstellungen gegenüber der Militärpolitik des Staates (zuerst: Nicht-Wahrnehmung; dann

kämpferisches Engagement) (Nominal-Skala). Wir hätten hier also zwei Bewegungen nebeneinander: Die eine auf der Intervall-Skala, die andere auf der "Nominal-Skala". Keine "Quantität" schlägt hier in eine "Qualität" um, sondern der eine soziale Prozeß bewirkt vielleicht den anderen oder löst ihn aus.

Wir können also festhalten, daß bei der Verwendung von Begriffen aus der Alltagssprache im Bereich statistischer Argumentation größte Vorsicht geboten ist. Die alltagssprachliche - manchmal metaphorische - Begrifflichkeit hat den Nachteil, daß es schwierig ist, sich darauf zu einigen, was genau gemeint ist. Viele Auseinandersetzungen in der Studentenbewegung hatten eine begriffliche Unschärfe aus einem marxistischen Sprachgebrauch als Ursache, nicht immer waren es sachliche Gegensätze. Oft ist es in politischem Sprachgebrauch auch durchaus beabsichtigt, sich nicht auf Definitionen festzulegen, weil so durch Manipulation des Sprachgebrauchs leichter erreicht werden kann, was sonst nur über offene Konflikte durchgesetzt werden könnte. Die Novelle „Animal Farm“ von George Orwell bietet hier die besten Beispiele.

Qualitative Analyse meint - im Gegensatz zu quantitativer Analyse - die komplexe Beschreibung von Einzelfällen, bei der weniger gezählt und gemessen wird, sondern mehr die Struktur einzelner Fälle sprachlich und vielleicht auch graphisch dargestellt wird. (Siehe hierzu u.a.: Flick, Uwe/Kardorff, Ernst von/Keupp, Heiner/Rosenstiel, Lutz von/Wolff, Stephan (Hrsg.): Handbuch Qualitative Sozialforschung. Grundlagen, Konzepte, Methoden und Anwendung. München (Psychologie Verlags Union) ²1995.

Qualitative Analyse liegt auch immer dann vor, wenn es um die Herstellung neuer Begriffe und Kategorien geht, während die quantitative Analyse in der Regel die vorherige Existenz und Operationalisierung von Begriffen voraussetzt (mit Operationalisierung ist die Angabe von Verfahren gemeint, mit deren Hilfe Messungen durchgeführt werden sollen).

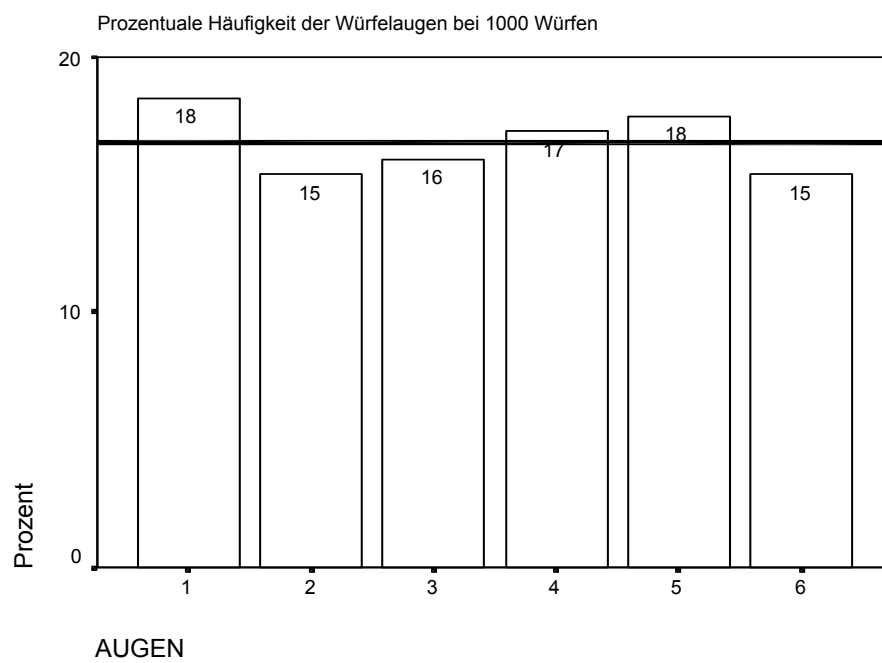
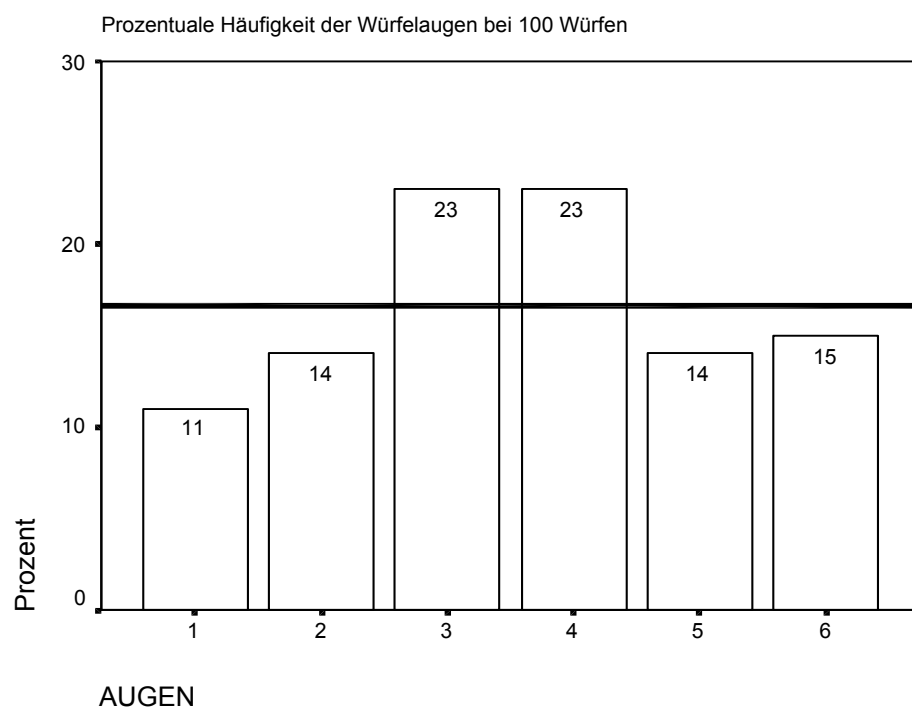
3. Einfache Beschreibungen von Massenereignissen - univariate Verteilungen

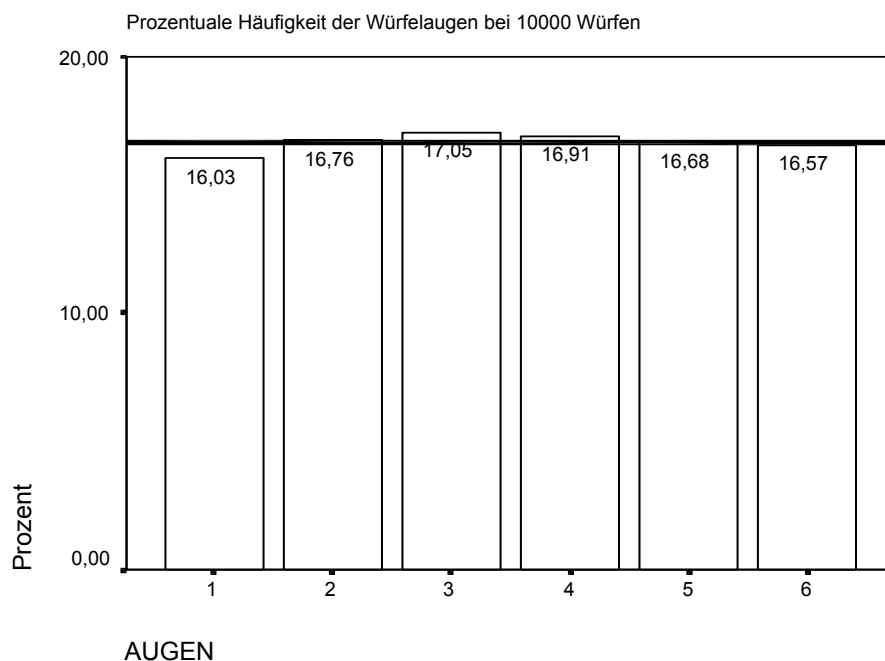
3.1 Der Begriff "Verteilung"

Massenereignisse setzen sich immer aus vielen Einzelereignissen zusammen. In unserem Empiriekurs sind beispielsweise 4 Männer und 16 Frauen, also 20 Personen („Ereignisse“, „Untersuchungsobjekte“). Wollen wir das Merkmal oder die Variable „Geschlecht“ an dieser „Grundgesamtheit“ untersuchen, dann können wir aufgrund der äußeren Erscheinung oder sogar vom Vornamen ausgehend viermal feststellen, daß die „Merkmalsausprägung“ „männlich“ für unsere Untersuchungsobjekte zutrifft, sechzehnmal die Merkmalsausprägung „weiblich“.

Solche Gegenüberstellungen von „Häufigkeiten“ nennen wir **Verteilungen**. Wenn nur eine Variable (Merkmal) betrachtet wird, nennt man die Verteilung „univariat“. Von **empirischen Verteilungen** sprechen wir, wenn wir sie aus wissenschaftlichen Messungen und Zählungen gewinnen. Empirische Verteilungen haben konkrete Merkmale bzw. Variablen zur Grundlage. **Theoretische Verteilungen** sind demgegenüber abstrakte mathematische Strukturen, deren Anwendbarkeit auf konkrete Fälle aufgrund bestimmter mathematischer Voraussetzungen experimenteller Anordnungen (wie z. B. beim Würfel-Experiment), erwartet wird. Kennen wir nämlich die mathematischen Voraussetzungen experimenteller Anordnungen, dann können wir ziemlich genau vorhersagen, wie sich die Ergebnisse des Experiments verteilen. So wissen wir zum Beispiel, daß beim Würfeln die gleichen Eintrittswahrscheinlichkeiten von jeweils $p = 1/6$ auf die Zahlen 1 bis 6 entfallen. Dies bedeutet nicht, daß wir einzelne Würfe vorhersagen könnten, aber die ungefähre empirische Verteilung nach einem Experiment: Wir erwarten aber eine Verteilung, die mit hoher Wahrscheinlichkeit von einer bestimmten Struktur umso weniger abweicht, je größer die Zahl der Versuche (hier: Würfe mit dem Würfel) ist:

Was hiermit gemeint ist, zeigen die folgenden Graphiken. Sie demonstrieren, wie sich die prozentualen Häufigkeiten einzelner Augenzahlen bei immer längeren Würfel-Serien immer mehr der Marke 16,67 % annähern. Wir wissen, daß die Wahrscheinlichkeit beim Würfeln, daß eine der sechs Würfelflächen oben liegt, gleich $1/6$ ist. Rechnet man $1/6$ in Prozente um, erhält man 16,666 ... %. Dieser Wert wird – **nach dem Gesetz der großen Zahl** - bei immer längeren Würfel-Versuchs-Serien immer besser angenähert:

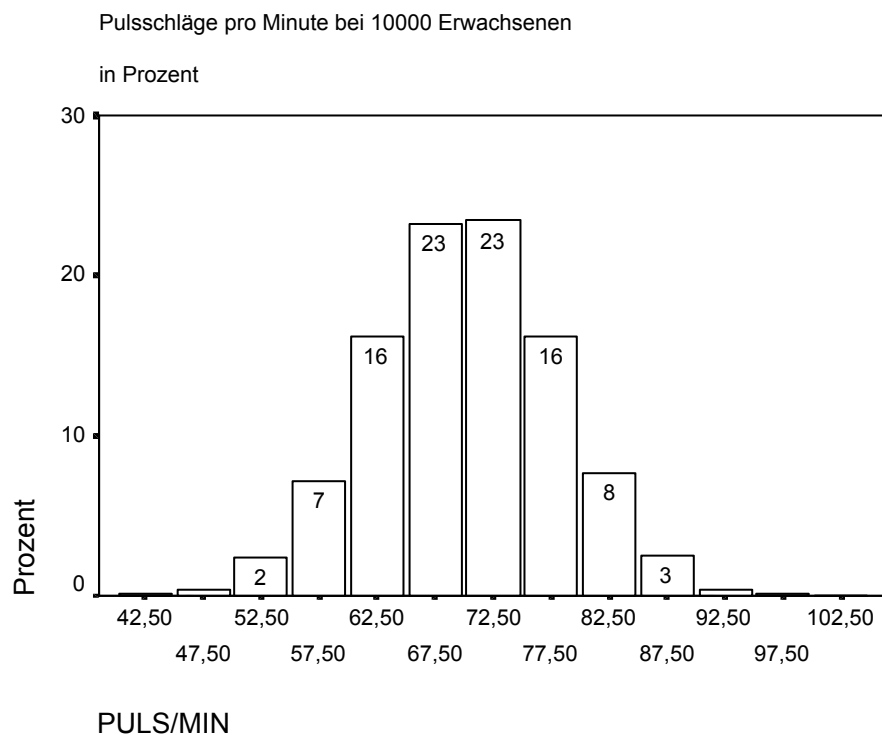




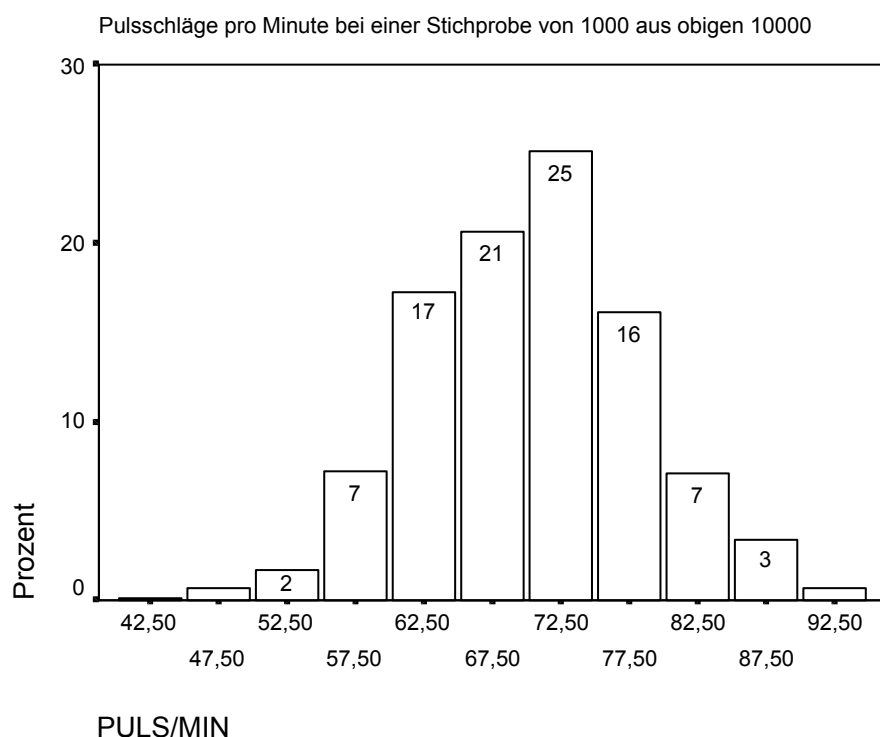
Die Beispielgrafiken – hier für die sogenannte Gleichverteilung - zeigen bei wachsender Zahl der Versuche eine immer bessere Annäherung der prozentualen Häufigkeiten, mit denen die Würfelaugen auftreten, an die hier erwartete prozentuale Häufigkeit von 16,67%.

Auch für ungleichverteilte Häufigkeiten läßt sich dieses Gesetz leicht demonstrieren.

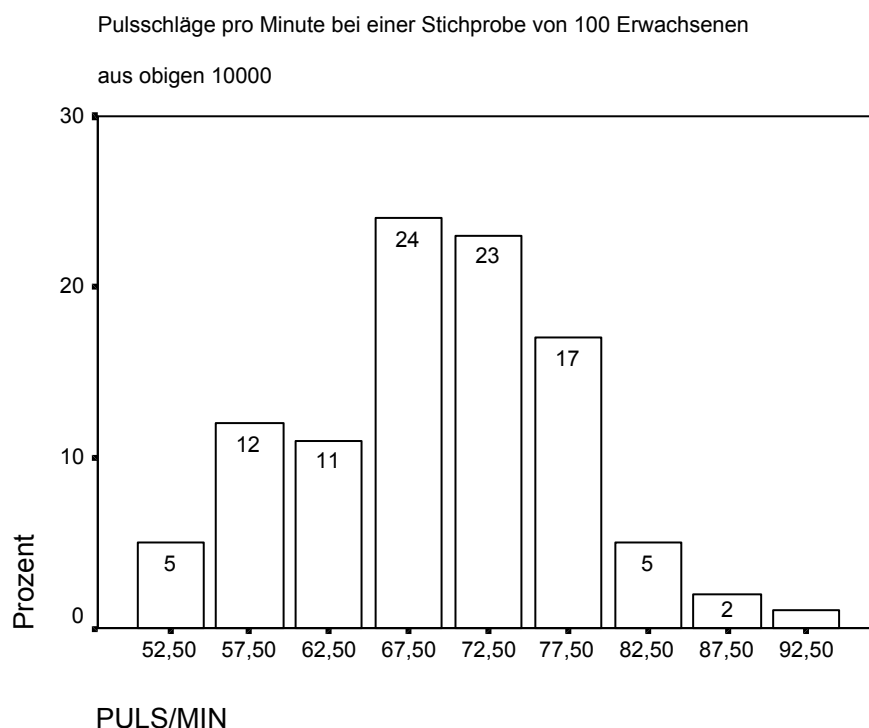
Gegeben sei eine Darstellung der Häufigkeiten des Pulsschlages pro Minute aus einer Kleinstadt mit 10000 Einwohnern:



Ziehen wir aus dieser „Grundgesamtheit“ nun zwei Stichproben, eine im Umfang von 1000 Einwohnern, eine im Umfang von 100 Einwohnern und beobachten wir dann, wie sich die prozentualen Häufigkeiten verändern, die den verschiedenen „Klassen“ von Pulsschlaghäufigkeiten in der Grundgesamtheit zukamen:



Hier haben sich die prozentualen Häufigkeiten der mittleren Klassen deutlich verändert, die der anderen Klassen weniger. Ziehen wir nun noch eine kleinere Stichprobe, die nur noch den Umfang 100 Erwachsene hat, dann verändern sich die prozentualen Häufigkeiten der angegebenen Klassen noch stärker. Die „normalverteilte Grundstruktur“ kann jedoch noch erahnt werden. Wir werden später sehen, daß der Bereich, in dem der „Mittelwert“ mit einer bestimmten Wahrscheinlichkeit auftauchen kann, sich in Abhängigkeit vom Stichprobenumfang gut bestimmen läßt.



Mit Hilfe der Kenntnis der zu einer empirischen Verteilung gehörenden theoretischen Verteilungsfunktionen kann relativ genau vorhergesagt werden, wieviele Einzelereignisse sich in bestimmten Intervallen empirischer Verteilungen befinden werden.

3.2 Ordnung und Klassifikation von Daten – graphische Darstellungen

Im folgenden sollen Daten analysiert werden, die die Altersangaben von Passanten an der Eingangstür unseres Universitätsgebäudes am Tag der offenen Tür (1.6. 1985) wiedergeben. 60 Passanten wurden hier nach ihrem Alter befragt. Es antworteten 60 wie folgt:

26;14;3;8;21;39;4;15;1;24;5;0;33;30;12;40;25;39;12;26;21;25;4;25;
22;22;15;24;32;28;31;31;25;24;34;15;27;13;13;26;38;19;11;8;23;30; 15;19;3;26;17;32;22;20;8;8;13;23;13;12;

Eine solche Datensammlung, die noch ungeordnet ist, nennt man auch *Urliste*.

Wenn wir uns diese Daten flüchtig betrachten, dann sind sie eher verwirrend - etwa wie der Blick aufs Fußballstadion. **Ordnen** wir aber diese Daten der Größe – also dem Alter – nach, dann wird die Sache schon übersichtlicher: In der folgenden Tabelle sind alle Altersangaben der Größe nach aufgeführt.

**Personen am Tag der offenen Tür
in der TUB am 1.6. 1985, nach Alter**

	Anzahl	%
,00	1	1,7%
1,00	1	1,7%
3,00	2	3,3%
4,00	2	3,3%
5,00	1	1,7%
8,00	4	6,7%
11,00	1	1,7%
12,00	3	5,0%
13,00	4	6,7%
14,00	1	1,7%
15,00	4	6,7%
17,00	1	1,7%
19,00	2	3,3%
20,00	1	1,7%
21,00	2	3,3%
22,00	3	5,0%
23,00	2	3,3%
24,00	3	5,0%
25,00	4	6,7%
26,00	4	6,7%
27,00	1	1,7%
28,00	1	1,7%
30,00	2	3,3%
31,00	2	3,3%
32,00	2	3,3%
33,00	1	1,7%
34,00	1	1,7%
38,00	1	1,7%
39,00	2	3,3%
40,00	1	1,7%

Die Altersangaben der Personen sind jetzt der Größe nach aufsteigend geordnet. Jede Alters-Angabe kommt nur einmal vor. Unter den Personen ist ein Baby unter einem Jahr, das mit der Altersangabe 0 geführt wird. 0 bedeutet hier nicht 0, sondern "noch nicht 1 Jahr alt". Weiterhin sehen wir, daß **identische Altersangaben mehrfach, mit einer bestimmten Anzahl oder "Häufigkeit" auftauchen**. Solche Häufigkeiten nennen wir „absolute Häufigkeiten“. Den einzelnen Häufigkeiten ist eine **Prozentangabe** zugeordnet.

Fassen wir die Daten noch weiter zusammen, um sie übersichtlicher darstellen zu können, dann liegt es nahe, die Daten zu diesem Zweck **in Klassen einzuteilen**.

Klassen sind hierbei

- Meßwertgruppen innerhalb jeweils eines Intervalls –
- meistens in einer Menge benachbarter Intervalle (Abstände) auf einer Meßskala, die in der Regel gleich groß sind.

Die Klassenbildung dient der übersichtlicheren Darstellung und der Verdeutlichung von Eigenschaften einer Verteilung.

Wollen wir selber Klassen definieren, schauen wir zunächst nach, wie groß der kleinste Meßwert (Minimum) und der größte (Maximum) sind. Dies sind hier die Altersangaben 0 und 40. Den Abstand zwischen Minimum und Maximum (hier 40) nennt man Variationsbreite oder (aus dem englischen Sprachraum): **RANGE**.

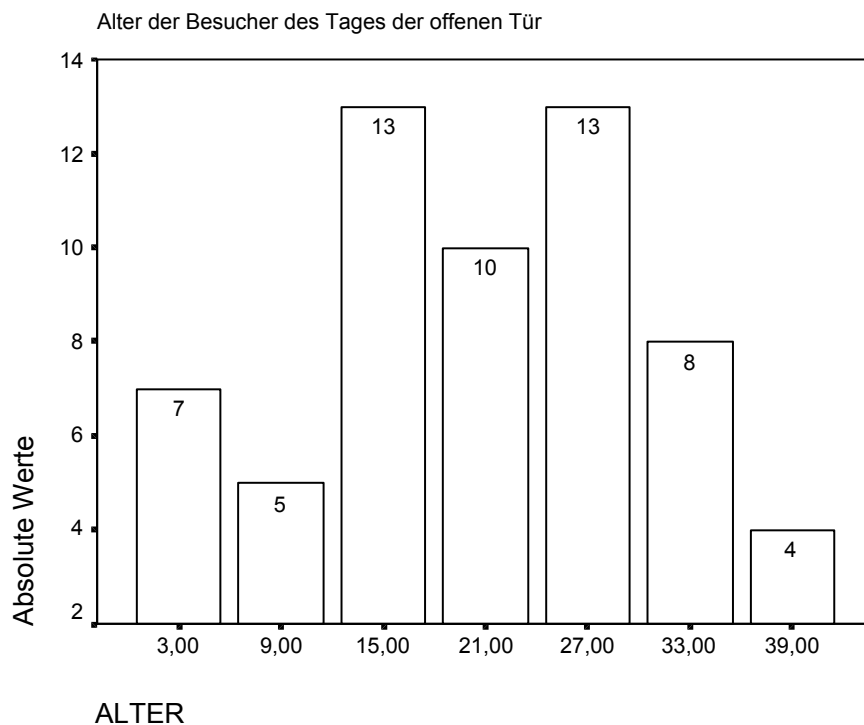
Alle Klassen müssen diesen Raum erfassen, so daß weder der kleinste noch der größte Meßwert unterschlagen werden. Wenn wir 7 Klassen bilden wollen, teilen wir den Range (40) durch 7 und erhalten 5,7. Wir können daher abschätzen, daß wir mit einer **Klassenbreite** von 6 (von 5,7 aufgerundet) arbeiten können, wenn wir ganzzahlige Klassenbreiten haben wollen, und daß wir mit 7 Klassen auskommen werden.

Wo soll aber die **Untergrenze der kleinsten Klasse** liegen? Wenn der kleinste Meßwert exakt 0 wäre, müßte die Untergrenze so gewählt werden, daß der Meßwert 0 noch von der kleinsten Klasse erfaßt wird. Bedeutet aber 0 nicht exakt „Null“, sondern signalisiert 0 nur, daß ein Meßwert zwischen 0 und 1 vorliegt (was hier wohl der Fall sein dürfte), dann können wir die Untergrenze der kleinsten Klasse auch so legen, daß sie zwar oberhalb von 0, aber ganz dicht bei diesem Wert 0 beginnt.

Wenn wir die Klassenbreite 6 einsetzen, erhalten wir folgende Klassen:

Klassengrenzen	Klassen mitten	Häufigkeiten
über 0 bis einschließlich 6	3	$f_1 = 7$
über 6 bis einschließlich 12	9	$f_2 = 5$
über 12 bis einschließlich 18	15	$f_3 = 13$
über 18 bis einschließlich 24	21	$f_4 = 10$
über 24 bis einschließlich 30	27	$f_5 = 13$
über 30 bis einschließlich 36	33	$f_6 = 8$
über 36 bis einschließlich 42	39	$f_7 = 4$

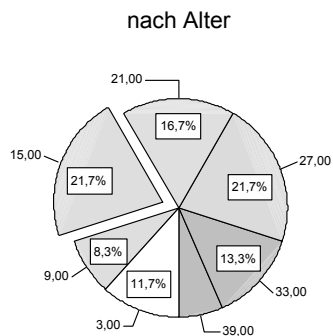
Diese klassifizierten Werte lassen sich nun in verschiedensten Formen auch graphisch darstellen. Die folgende Darstellung nennt man „Balkendiagramm“. In ihm ist jeder **Klassenmitte** ein Balken oder Streifen zugeordnet, **dessen Länge die Häufigkeit symbolisiert**, die der Klasse zukommt, auf deren „Mitte“ er errichtet ist. Die Breite des Balkens ist bedeutungslos; zwischen den Balken kann ein Zwischenraum existieren:



Das Balkendiagramm hat den Vorteil, daß es nicht nur auf Klassen mit definierter Breite angewendet werden kann, sondern auch auf Nominal- und Ordinaldaten. Dort wird der Balken den nominalen „Kategorien“ oder den Rängen zugeordnet.

Für das folgende „Kreisdiagramm“ haben wir die Klasseneinteilung des zuvor gezeigten Stabdiagramms übernommen. Der Winkel einzelner Sektoren richtet sich nach dem Anteil der in ihm enthaltenen Häufigkeit. Ein Sektor ist „ausgerückt“ um ihn hervorzuheben.

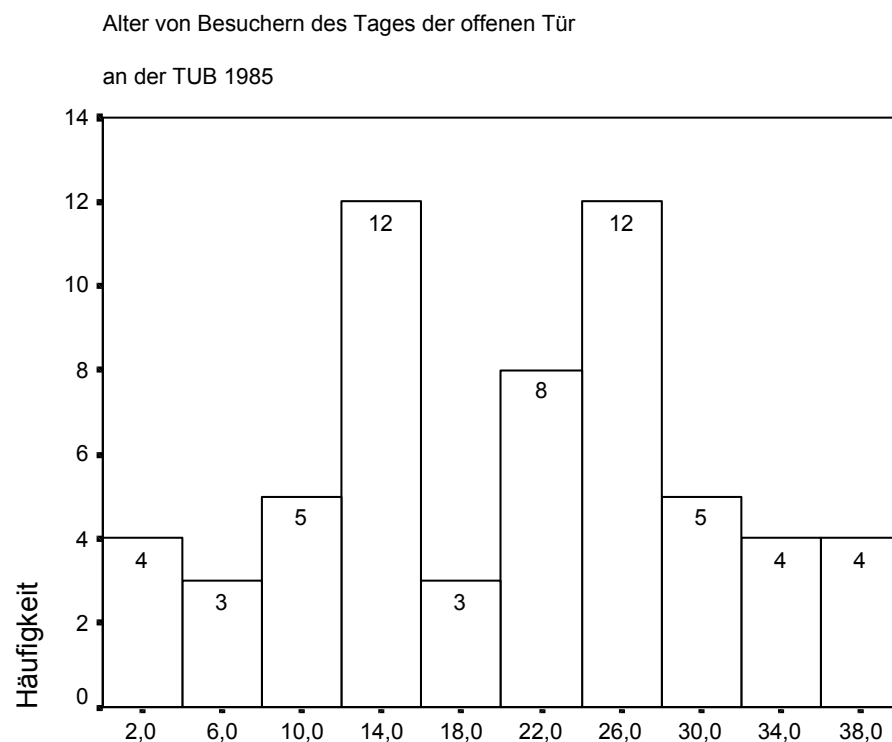
Besucher des Tages der offenen Tür in der TUB



Je nach Klassenbreite erhalten wir unterschiedliche Formen der Verteilung, da sich bei veränderter Klassenbreite auch die Häufigkeiten pro Klasse ändern, wie weiter unten (siehe Ausführungen zum "Modus") gezeigt wird. Hierin liegt insbesondere für die politische Praxis im Umgang mit Statistiken ein wichtiges Gestaltungsmoment: Unerfreuliches wird graphisch unterdrückt, Erfreuliches graphisch aufgebläht.

Das folgende Diagramm ist ein sogenanntes **Histogramm**, in dem einem ganzen Intervall (also nicht nur einem Meßpunkt) ein dieses Intervall voll ausfüllender Streifen zugeordnet wird, dessen Fläche der Häufigkeit entspricht, mit der Fälle im jeweiligen Intervall auftreten. Das unten dargestellte Histogramm hat die **Klassenbreite 4** (und nicht 6, wie das obige Balkendiagramm). Daher sind auch andere Klassengrenzen und andere Häufigkeiten zustande gekommen, die man aus den unten angegebenen Klassenmitten rekonstruieren kann:

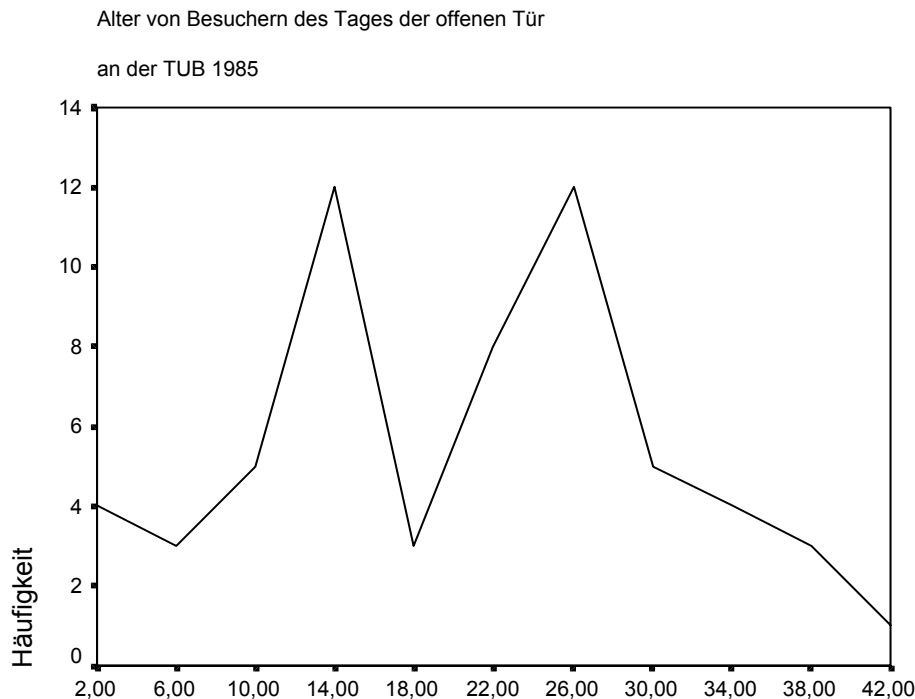
Histogramm



Klassenuntergrenze	Klassenobergrenze	Klassenmitte	Absolute Häufigkeit	Prozentuale Häufigkeit
über 0	bis einschließlich 4	2	4	6,7
über 4	bis einschließlich 8	6	3	5
über 8	bis einschließlich 12	10	5	8,3
über 12	bis einschließlich 16	14	12	20
über 16	bis einschließlich 20	18	3	5
über 20	bis einschließlich 24	22	8	13,3
über 24	bis einschließlich 28	26	12	20
über 28	bis einschließlich 32	30	5	8,3
über 32	bis einschließlich 36	34	4	6,7
über 36	bis einschließlich 40	38	4	6,7
		Summe	60	100

Schließlich kann man die gleichen Daten noch mittels Häufigkeitspolygon darstellen:

Häufigkeitspolygon



Hier werden Punkte über den Klassenmitten in die Graphik im senkrechten Abstand der jeweiligen Häufigkeit eingetragen und mittels Linienzug verbunden.

Zu beachten ist bei dem obigen Häufigkeitspolygon, daß eine andere Klassengrenzenwahl zugrundeliegt als beim Histogramm, nämlich von 0 bis unter 4, von 4 bis unter 8 etc., so daß der Wert 40 nicht mehr in die vorletzte Klasse (bis unter 40) fällt, sondern in die letzte Klasse (40 bis unter 44).

3.3. Zusammenfassung der Information aus Häufigkeitsverteilungen

3.3.1. Daten auf dem Intervall- und Ratio-Niveau der Messung - Lage- und Streuungsmaße

Neben einer graphischen Darstellung von Häufigkeiten gibt es nun Möglichkeiten einer **vergleichsbezogenen** mathematischen Beschreibung von empirischen Verteilungen: Stichproben-Verteilungen werden hierbei durch Kennwerte beschrieben, die die **Form** und die **Lage** der Verteilung charakterisieren.

Bei der **Form** der Verteilung kommt es oft darauf an,

- ob die Verteilung einer sogenannten **Normalverteilung** ähnlich ist oder nicht. Die Normalverteilung wird weiter unten besprochen. Maße, die die Ähnlichkeit mit einer Normalverteilung messen, sind die „Schiefe“ und der „Exzess“.

- Weiterhin kommt es bei der Form darauf an, ob die Verteilung **breit** oder **schmal** ist. Maße, die die Breite betreffen, nennen wir „**Streuungsmaße**“. Die Streuung einer Verteilung wird u.a. auch durch die Differenz von kleinstem und größtem Meßwert charakterisiert.
- Maße, die die **Symmetrie** oder **Asymmetrie** betreffen, lassen sich aus dem Vergleich verschiedener Mittelwerte entwickeln. Asymmetrie liegt dann vor, wenn wir es mit einer Häufung der Meßwerte an der **rechten** oder **linken Seite der Verteilung** zu tun haben.
- Weiterhin kann noch die **Menge der Gipfelpunkte** einer Verteilung von Bedeutung sein.

Bei der **Lage** der Verteilung kommt es darauf an,

- **wo Meßwerte liegen, die die ganze Verteilung gut repräsentieren.** Solche Werte, die die gesamte Verteilung gut repräsentieren, sind in der Regel „Mittelwerte“ oder auch „Maße der zentralen Tendenz“ (s.u.), die dann brauchbar sind, wenn die Form der Verteilung nicht allzu asymmetrisch ist.

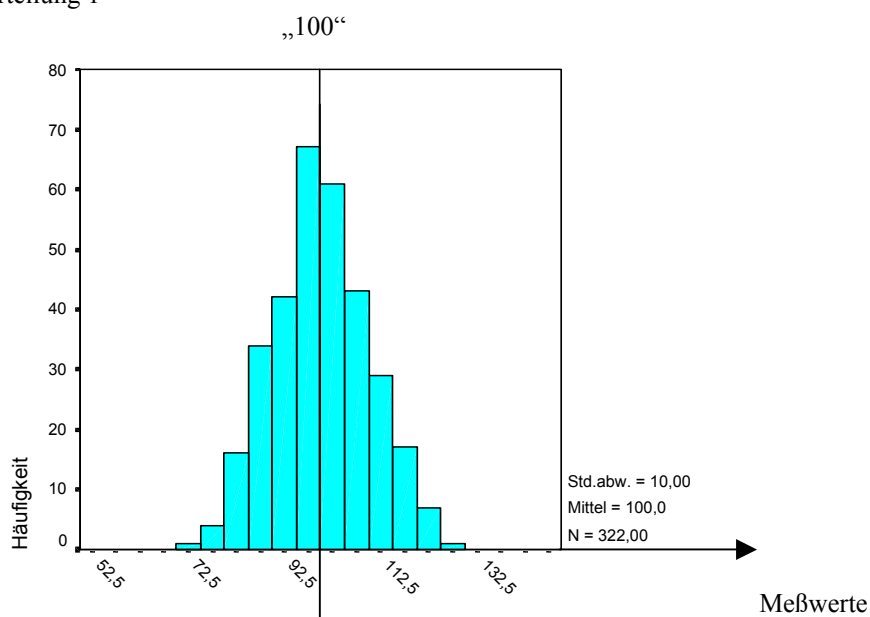
Um zu zeigen, worum es hier geht, werden unten zunächst drei verschiedene Verteilungen **mit dem gleichen Maßstab** der Meßwertachse als **Histogramme** dargestellt, zusammen mit ausgewählten statistischen Maßen, die sie charakterisieren:

Die drei Verteilungen unterscheiden sich. Während „Verteilung 1“ offensichtlich in ihrer Mitte **über dem Meßwert 100** liegt, liegt die Mitte von „Verteilung 2“ **über dem Meßwert 120** – d.h. im Durchschnitt sind die Meßwerte der „Verteilung 2“ **um ca. 20 Meßwertpunkte größer** als die von „Verteilung 1“. Diese „Mitten“ sind in den unten aufgeführten Graphiken durch einen senkrechten Strich oberhalb der Graphik markiert.

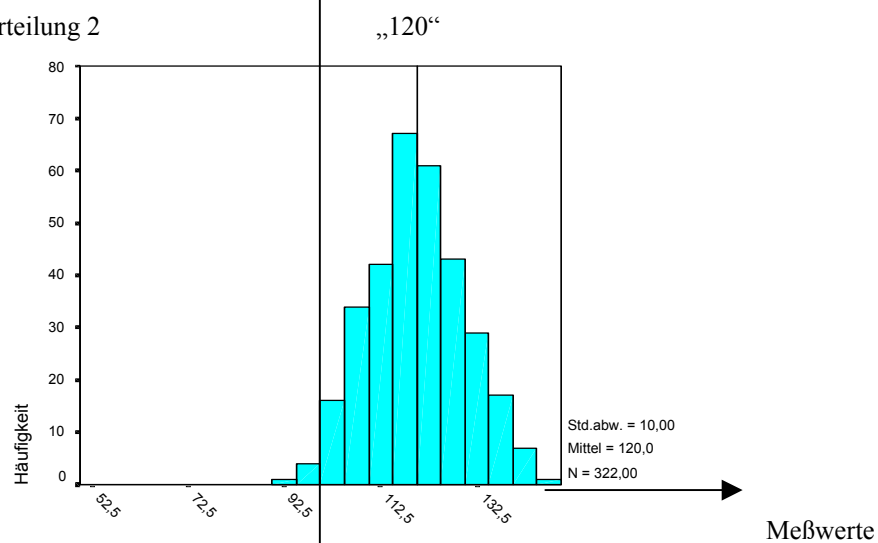
Bei „Verteilung 3“ fällt auf, daß sie **breiter ist** (weiter streut) als die Verteilungen 1 und 2. Außerdem hat sie – im Gegensatz zu den anderen Verteilungen – zwei Gipfelpunkte – also zwei Häufigkeiten, deren Nachbarhäufigkeiten rechts und links niedriger sind als die Häufigkeit am „Gipfelpunkt“.

Drei ähnliche – und doch deutlich verschiedene – annähernd „normale“ Verteilungen:

Verteilung 1



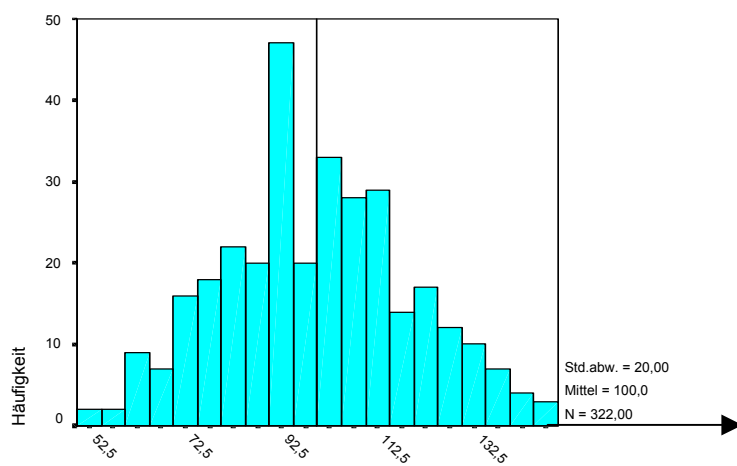
Verteilung 2



Verteilung 3

„100“

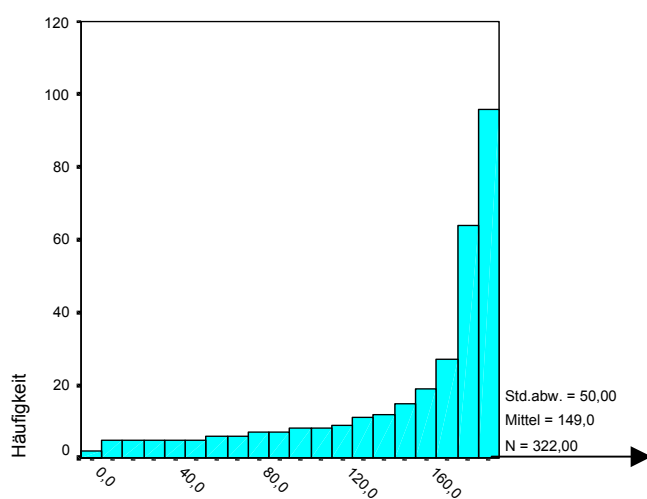




Meßwerte

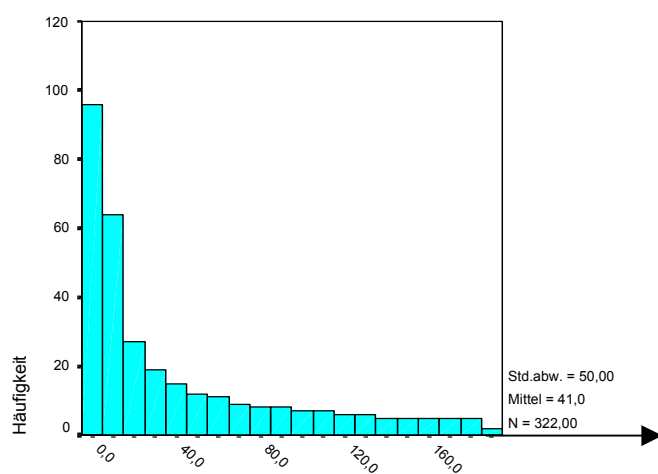
Während die vorherigen Verteilungen noch einigermaßen „symmetrisch“ aussahen, folgen jetzt zwei stark asymmetrische Verteilungen – wiederum Histogramme:

Verteilung 4



Meßwerte

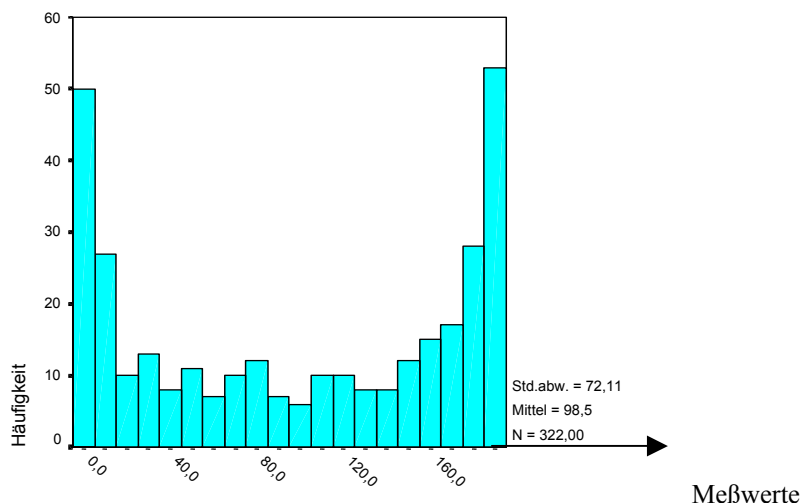
Verteilung 5



Meßwerte

Die folgende Verteilung ist „u-förmig“, aber dennoch ziemlich symmetrisch:

Verteilung 6:



In der „beschreibenden Statistik“ (auch „deskriptive“ Statistik genannt) wird nun der Versuch gemacht, über die graphischen Mittel des Vergleichs von Verteilungen hinaus noch mathematische Hilfsmittel zu konstruieren, die die Lage, die Streuung und die Form von Verteilungen auszudrücken gestatten. Dies hat Vorteile, wenn es darum geht, den Vergleich von Häufigkeitsverteilungen nicht nur durch „Augenschein“ vorzunehmen, sondern auf eine mathematische, besonders aber wahrscheinlichkeitstheoretische Grundlage zu stellen.

Die Problematik, die hierbei auftritt, besteht darin, optische Vergleichsvorgänge auf mathematische Vergleichsverfahren zu reduzieren oder sie durch mathematische Verfahren zu präzisieren. Wenn z.B. die beiden folgenden Flächen bezüglich ihrer Größe verglichen werden, dann kann durch den Augenschein kaum entschieden werden, welche größer ist. Messe ich aber die Kanten und multipliziere sie miteinander, dann kann die Frage genau entschieden werden.



Ähnlich ist es mit den Maßen der zentralen Tendenz:

Nehmen wir an, die „Mitte“ einer Meßwerte-Schar sei ein adäquater Ausdruck ihrer zentralen Tendenz. Was ist aber die „Mitte“? Bei zwei Meßwerten ist das klar: Die Mitte wird durch die Hälfte der Strecke zwischen den Meßwerten bestimmt. Bei drei Meßwerten ist das schon nicht mehr klar: Ist der „mittlere“ Meßwert von dreien immer Repräsentant der Mitte aller drei Meßwerte, unabhängig davon, wie weit er von den beiden äußeren Meßwerten entfernt ist? Und wie ist das, wenn ich mehr als drei Meßwerte habe? Wie ist das bei asymmetrischen Verteilungen, bei denen die größten Häufigkeiten nicht in einem mittleren Intervall liegen, sondern am Rande?

3.3.1.1 Maße der zentralen Tendenz und Lagemaße

Jedes Maß der zentralen Tendenz kann nur alle Werte einer Verteilung durch einen einzigen Wert „ersetzen“. Diehl (Diehl, Joerg M.; Kohr, Heinz U.: Deskriptive Statistik, 12. Auflage 1999, S. 76 f) führt hierzu aus, daß dieser Ersatz immer ein unvollkommener Ersatz ist, und er systematisiert die Fehler, die man bei den verschiedenen Maßen der zentralen Tendenz macht, wenn man sie auf empirische Verteilungen anwendet. Dieses Verfahren der Berechnung und Systematisierung der Fehler macht die Maße der zentralen Tendenz in gewisser Weise in ihrer Brauchbarkeit und Unbrauchbarkeit einschätzbar.

Exkurs zum Summenzeichen

Bevor wir auf die formelmäßige Darstellung der Maße der zentralen Tendenz und der Streuung eingehen, müssen wir ein mathematisches Symbol einführen, das Summenzeichen Σ (großer griechischer Buchstabe „Sigma“). Es steht für die Vorschrift, eine Folge von Meßwerten aufzuaddieren, wobei aus dieser Folge eine Reihe wird (vgl. unten, S. 102).

Eine Folge ist eine Menge von Zahlen, die in einer bestimmten Reihenfolge stehen. Eine arithmetische Folge bzw. eine geometrische Folge sind besondere Folgen, die aus besonderen Regeln hervorgehen (s.u.). Beim Σ wird - wenn dies sich nicht aus dem Zusammenhang ergibt - angegeben, von welchem bis zu welchem Glied der Folge, auf die sich das Summierungszeichen bezieht, die Summierung durchgeführt werden soll. Die einzelnen Glieder einer Folge werden - ihrer Stellung in der Folge nach - durchnummeriert und diese Nummer wird auch als Index bezeichnet, der - in allgemeiner Form - oft durch die Buchstaben i oder j oder k ersetzt wird, wobei diese Buchstaben tiefgestellt gedruckt werden. i, j oder k sind das sogenannte „Laufindizes“.

Beispiel:

Die Folge von Zahlen 3 5 8 2 2 1 0 8 9

kann auch bezeichnet werden mit

$a_1 \ a_2 \ a_3 \ a_4 \ a_5 \ a_6 \ a_7 \ a_8 \ a_9$

Bildet man die Summe der vier letzten Zahlen in dieser Folge, so ließe sie sich mathematisch wie folgt ausdrücken:

$$\sum_{i=6}^9 a_i = a_6 + a_7 + a_8 + a_9 = 1 + 0 + 8 + 9 = 18$$

Wie man sieht, steht unter dem Σ ein Laufindex i, dessen erster Wert für den Summierungsvorgang unter dem Σ und dessen letzter Wert über dem Σ angegeben sind. Weitere Informationen und Übungsbeispiele finden sich zu dieser Symbolik im Anhang.

Das Summenzeichen ist ein in der Statistik sehr häufig benutztes Symbol. Es ist daher sehr wichtig, mit ihm sachgerecht umgehen zu können. Mit ihm werden in der Regel drei verschiedene - ebenfalls in der Statistik verwendete Symbole - verknüpft:

x_i als Symbol für Meßwerte

y_j ebenfalls als Symbol für Meßwerte - aber Meßwerte einer anderen „Variablen“ „ y “.

f_k als Symbol für Häufigkeiten in bestimmten Klassen von Meßwerten.

3.3.1.2 Arithmetisches , geometrisches Mittel und harmonisches Mittel

Das arithmetische Mittel ergibt sich als Quotient aus der **Summe der Meßwerte** ($\sum x_i$), die man durch die **Menge der Meßwerte** (N), [oder bei Stichproben (n)], [oder bei klassifizierten Daten $\sum f_k$] teilt. Das heißt, wir können das arithmetische Mittel „ \bar{x} “ (sprich x-quer) in verschiedenen Schreibweisen ausdrücken:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

bzw.

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i$$

oder
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{\sum_{j=1}^k f_k}$$

Hierbei bedeuten: x_i Meßwerte i , $i = 1, 2, 3, \dots, n$
 n Menge der Meßwerte
 f_k Häufigkeit in der Klasse k .

Das arithmetische Mittel ist im übrigen ein physikalisches Konzept, das auf die Hebelgesetze zurückgeführt werden kann. Es entspricht dem Konzept des Schwerpunktes eines Körpers. Man nennt das arithmetische Mittel deshalb "arithmetisch", weil es aus einer arithmetischen Folge immer genau die Mitte herausucht!

Arithmetische Folgen sind solche, die durch fortlaufende Addition einer Konstante zu einem Betrag entstehen.

Die Folge

7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27.

wäre eine arithmetische Folge, die ab der Startzahl 7 durch fortlaufende Addition der Konstante 2 zum Vorgängerwert entsteht. Ihr mittlerer Wert, der 6. von 11 geordneten Zahlen, ist 17. Er ist identisch mit dem arithmetischen Mittel.

Das **geometrische Mittel** ist in Analogie zum arithmetischen Mittel konstruiert. Es sucht aus einer geometrischen Folge den mittleren Wert heraus.

Hierbei gilt als geometrische Folge eine Folge, die durch fortlaufende Multiplikation einer Startzahl mit dem gleichen Faktor erreicht wird, wie

3,5 17,5 87,5 437,5 2187,5

Diese Folge entsteht aus

$$3,5 * 5^0; \quad 3,5 * 5^1; \quad 3,5 * 5^2; \quad 3,5 * 5^3 \quad 3,5 * 5^4 \dots$$

Das geometrische Mittel sucht nun aus dieser Folge das mittlere Element heraus, also den Wert $3,5 * 5^2 = 87,5$.

Die Formeln für das geometrische Mittel lauten:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i} = \left[\prod_{i=1}^n x_i \right]^{\frac{1}{n}} = 10^{\left(\frac{1}{n} \sum_{i=1}^n \log x_i \right)}$$

Hinter dem letzten Gleichheitszeichen steht die logarithmische Schreibweise des geometrischen Mittels.

Π ist (in Analogie zu Σ) die Vorschrift zu fortlaufender Multiplikation der Meßwerte ($x_1 * x_2 * x_3 * \dots * x_n$). (Π ist hier also nicht die Kreiskonstante 3,14... !) Die logarithmische Schreibweise (s.o.) empfiehlt sich, wenn die Produkte zu groß werden, um noch leicht mit ihnen rechnen zu können. Das ist immer der Fall, wenn n groß ist.

Führt man für das o.a. Beispiel diese Rechnung aus, ergibt sich

$$\begin{aligned} \bar{x}_g &= \left[(3,5 * 5^0 * 3,5 * 5^1 * 3,5 * 5^2 * 3,5 * 5^3 * 3,5 * 5^4) \right]^{\frac{1}{5}} \\ &= (3,5^5 * 5^{10})^{\frac{1}{5}} = 3,5^1 * 5^2 = 3,5 * 25 = 87,5 \end{aligned}$$

Analog zur arithmetischen Folge ergibt sich auf diese Weise bei echten geometrischen Folgen eine Identität von Median und geometrischem Mittel.

Zu beachten ist, daß beim geometrischen Mittel keine Meßwerte verarbeitet werden können, die gleich oder kleiner als Null sind. Tauchen solche Werte auf, dann sind alle Meßwerte vor Berechnung des geometrischen Mittels durch Addition einer Konstante so zu transformieren, daß sie größer als Null sind. Vom geometrischen Mittel dieser transformierten Werte wäre dann die Konstante wieder abzuziehen.

Neben dem arithmetischen und geometrischen Mittel existieren noch weitere Mittelwerte, die auf einer harmonischen Folge aufbauen.

Eine harmonische Folge liegt in folgender Zahlenreihe vor:

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \dots$$

Allgemein ausgedrückt:

$$\frac{1}{x_i} \quad \text{für } x_i = 1 \text{ bis } n.$$

Das harmonische Mittel hat die Formel:

$$HM = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Führt man mit dieser Formel die Berechnung durch, dann ergibt sich das Resultat von 0,25, also $\frac{1}{4}$. $\frac{1}{4}$ ist der 4. Wert der obigen Reihe aus 7 Elementen, d.h. $\frac{1}{4}$ ist der mittlere Wert.

Praktisch wird das harmonische Mittel zum Beispiel auf Meßwerte angewendet, die sich auf Geschwindigkeiten beziehen. Wenn ein Fahrradfahrer die ersten 10 km eine Geschwindigkeit von 15 km/h fährt, die zweiten 10 km eine Geschwindigkeit von 18 km/h und die dritten 10 km eine Geschwindigkeit von 21 km/h, dann käme aus der Rechnung als Durchschnittsgeschwindigkeit heraus: 17,66 km/h.

Diehl und Kohr (Deskriptive Statistik, a.a.O., S. 86) erwähnen noch das von Senders 1958 vorgestellte „Kontraharmonische Mittel“:

$$KHM = \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i}$$

Es wird verwendet, wenn x_i ein Wert ist, der seinerseits eine Häufigkeit bezeichnet wie die Anzahl der Kinder pro Familie.

Beispiel:

Kinder pro Familie x_i	Anzahl der Familien f_i	$f_i x_i$	$f_i x_i^2$
2	20	40	80
3	15	45	135
5	11	55	275

$$KHM = \frac{490}{140} = 3,5$$

Das durchschnittliche Kind kommt aus einer Familie mit 3,5, also 3 bis 4 Kindern (KHM). Die durchschnittliche Familie hat aber nur 3,04 Kinder, wie man mit dem arithmetischen Mittel zeigen kann.

3.3.1.3. Der Median

Der **Median** (M_d) bezeichnet in einer nach Größe geordneten, **aber sonst beliebigen Meßwertfolge** den mittleren Wert der **geordneten** Folge, unabhängig vom Charakter der Folge:

Für die Folge

10,30,50,51,52,53,57 (7 Meßwerte, der Größe nach geordnet, **ungleiche** Abstände)

gilt: $\bar{x}=43,29$; $M_d = 51$

Für obige Folge mit 7 Meßwerten ist der Median der 4. Meßwert. Er teilt die Meßwertfolge in zwei gleich große Mengen.

In dieser Folge wäre $\bar{x} = 43,29$, läge also zwischen dem 2. und 3. Meßwert. Die Ursache für die Abweichung von Median und \bar{x} ist der für diese Folge extrem kleine Wert 10.

Für arithmetische Folgen sind Median und arithmetisches Mittel identisch. Das gleiche gilt für alle symmetrischen Verteilungen. (Hieraus ist jedoch nicht im Umkehrschluß abzuleiten, daß alle Verteilungen symmetrisch sind, für die der Median und das arithmetische Mittel identisch sind. Wie man durch Ausprobieren feststellen kann, gibt es auch leicht asymmetrische Verteilungen mit Identität von arithmetischem Mittel und Median).

Betrachten wir noch einmal unsere Besucher des Tages der offenen Tür. 60 Besucher kamen, der Median mußte also zwischen dem 30. und 31. Meßwert in der geordneten Reihe liegen. Dies sind die Meßwerte 21 und 22. Da in Folgen, die eine geradzahlige Menge von Elementen haben, keine einzelne „mittlere“ Zahl vorkommt, sondern immer zwei mittlere Zahlen, legt man in diesen Fällen den Median als das arithmetische Mittel der beiden mittleren Zahlen fest. Daher ist der Median = 21,5. Das arithmetische Mittel ist für die Rohwerte $\bar{x} = 19,87$. Median und arithmetisches Mittel sind hier also nicht identisch, da die Meßwerte keine arithmetische Folge darstellen.

Wie wirken sich nun Klassifikationsvorgänge auf arithmetisches Mittel und Median aus?

3.3.1.4 Arithmetisches Mittel und Median bei klassifizierten Werten

Klassifizieren wir unsere Werte vom Tag der offenen Tür in Klassen mit der Breite 7, dann zeigt sich folgendes Bild, in dem auch die Summe der Meßwerte pro Klasse berechnet wurde:

Klassengrenzen	Klassenmitten	Häufigkeiten f_k	Summe der Meßwerte pro Klasse
0 bis unter 7	x_1 3,5	f_1 7	24,5
7 " " 14	x_2 10,5	f_2 12	126,0
14 " " 21	x_3 17,5	f_3 9	157,5
21 " " 28	x_4 24,5	f_4 19	465,5
28 " " 35	x_5 31,5	f_5 9	283,5
35 " " 40	x_6 38,5	f_6 4	154,0
		$\Sigma f_k = 60$	$\Sigma f_k x_k = 1211$

Die 1. Klasse von 0 bis unter 7 hat hier die Klassengrenzen 0 und 6,999...also die **Klassenmitte** 3,5. Da die Statistiker hier die Konvention haben, wenn andere Informationen fehlen, anzunehmen, **daß die Werte einer Klasse in ihrer Mitte konzentriert** sind, haben wir bei klassifiziert vorliegenden Werten in diesem Fall davon auszugehen, daß 7 mal der Meßwert 3,5 vorliegt.

Das arithmetische Mittel \bar{x} errechnet sich demgemäß aus

$$\bar{x} = \frac{f_1(x_1) + f_2(x_2) + \dots + f_6(x_6)}{\Sigma f_k}$$

$$= \frac{7 * (3,5) + 12 * (10,5) + 9 * (17,5) + 19 * (24,5) + 9 * (31,5) + 4 * (38,5)}{60}$$

$$= 1211/60 = 20,18 \text{ Jahre}$$

Dieses Rechenergebnis für \bar{x} ist mit dem für die Rohwerte berechneten ($\bar{x} = 19,87$) nicht identisch!

Die Nichtübereinstimmung resultiert aus der Ungenauigkeit der Klassenmitten als Repräsentanten aller Werte pro Klasse.

Die **maximale Abweichung** des arithmetischen Mittels bei klassifizierten Werten vom "wahren Wert" beträgt **eine halbe Klassenbreite**; sie träte ein, wenn alle Rohwerte auf der Klassenuntergrenze oder der Obergrenze konzentriert wären, was praktisch aber nie der Fall ist.

Wie berechnen wir nun den Median bei klassifizierten Werten? Hierzu stellen wir fest, in welche der Klassen er fällt. Er liegt im obigen Beispiel zwischen dem 30. und 31. Wert, da dort $N = 60$ ist.

Diese beiden Werte liegen in der 4. Klasse. Also wäre der Median = 24,5, d.h. identisch mit der Mitte der Klasse, in die er fällt, wenn wir bei der Annahme bleiben, daß sich die Werte auf die Klassenmitte konzentrieren.

3.3.1.5. Weitere Lagemaße: N-til und Prozentränge

Während der Median genauso viele Meßwerte unter wie über sich hat, haben **N-til** die Eigenschaft, die geordneten Meßwerte auch in einem anderen Verhältnis zu teilen, z.B. im Verhältnis 25:75; 10:90; 75:25 und so weiter. Bei 4 N-tilen spricht man von „**Quartilen**“, bei 10 N-tilen spricht man von „**Dezilen**“, man benutzt also die lateinischen Zahlen zur Bildung dieser Begriffe. Wie schon bei der Medianberechnung kommt es auch hier wieder darauf an,

- ob ein konkreter Meßwert das N-til repräsentiert,
- ein Wert zwischen zwei benachbarten Meßwerten,
- oder eine Klassenmitte.

Für unser Beispiel von den Besuchern des Tages der offenen Tür 1985 können wir folgendes feststellen:

Ordnet man den Altersstufen nicht nur die absolute Häufigkeit zu, sondern auch die prozentuale Häufigkeit ($100 \cdot f_k/n$), und zählt man jede weitere prozentuale Häufigkeit zu der Summe der Vorgänger hinzu, dann erhält man die in der letzten Spalte der unten stehenden Tabelle aufgeführten kumulierten prozentualen Häufigkeiten. Sie sind identisch mit den **Prozenträngen**.

Von diesen unterscheiden sich die **N-til** dadurch, daß sie durch die jeweiligen Klassengrenzen beeinflußt sind.

Für die Demonstration des Unterschieds nehmen wir wieder unsere Daten vom Tag der offenen Tür als Beispiel, jetzt mit einer Klassifikation, bei der die Klassenbreiten gleich 6 sind und die Klassen von Null bis unter 6, von 6 bis unter 12 etc. definiert sind:

Altersangaben zum Tag der offenen Tür in der TU-Berlin

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	3,00	7	11,7	11,7	11,7
	9,00	5	8,3	8,3	20,0
	15,00	13	21,7	21,7	41,7
	21,00	10	16,7	16,7	58,3
	27,00	13	21,7	21,7	80,0
	33,00	8	13,3	13,3	93,3
	39,00	4	6,7	6,7	100,0
	Gesamt	60	100,0	100,0	

Für diese Daten sind die rechts in der Tabelle eingetragenen Werte (die kumulierten Prozente), wie schon gesagt, gleich den **Prozenträngen**.

Fragt man nach den N-tilen, dann gibt es für diese Frage 2 Antworten: Entweder, man tut so, als würden die Daten nur aus Angaben über die Klassenmitten bestehen. Aus dieser Annahme resultiert dann folgende Lösung für 6 N-tile (in der nachfolgenden Tabelle mit Percentile bezeichnet):

OFFTU1		
N	Gültig	60
	Fehlend	0
Mittelwert		20,7000
Median		21,0000
Perzentile	16,66666667	9,0000
	33,33333333	15,0000
	50	21,0000
	66,66666667	27,0000
	83,33333333	33,0000

Die 5 in der Tabelle aufgeführten N-tile von insgesamt 6 N-tilen (oben als „Percentile“ bezeichnet), entsprechen den Klassenmitten der Klassen, in die jeweils die N-tile (bzw. Percentile) fallen.

Nimmt man an, daß die Daten sich gleichmäßig jeweils über die gesamte Klassenbreite verteilen, dann resultiert folgende Lösung des gleichen Problems aus obigen Daten:

OFFTU1		
N	Gültig	60
	Fehlend	0
Mittelwert		20,7000
Median		21,0000 ^a
Perzentile	16,66666667	9,3333 ^b
	33,33333333	15,7826
	50	21,0000
	66,66666667	26,2174
	83,33333333	31,8571
a. Aus gruppierten Daten berechnet		
b. Perzentile werden aus gruppierten Daten berechnet.		

Hier sind die N-tile (wiederum in der Tabelle als „Percentile“ bezeichnet) nicht mehr identisch mit den Klassenmitten, sondern nehmen Werte an, die innerhalb der Intervalle der jeweiligen Klassen liegen. Formeln für die Berechnung entsprechender N-tile finden sich bei Benninghaus (Hans Benninghaus, Einführung in die sozialwissenschaftliche Datenanalyse, München 1990, S. 130ff) und Diehl/Kohr (Diehl, Joerg M und Kohr, Heinz U.: Deskriptive Statistik, Verlag Dietmar Klotz, Frankfurt am Main, 12. Auflage, S. 57).

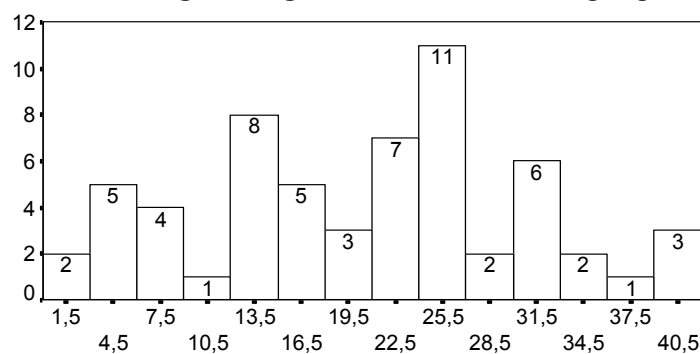
Diese Formeln, deren Handhabung ziemlich umständlich ist, werden hier nicht referiert. Da wir in der Praxis sowieso meistens mit den Rohdaten und nicht mit den klassifizierten Daten rechnen, kommen diese Formeln selten zur Anwendung. Sie sind aber in SPSS.10 verfügbar, und zwar über die Option „Daten sind aus gruppierten Daten berechnet“ (s.o.).

3.3.1.6. Modi

Der **Modus** ist neben den Mittelwerten, N-tilen und Prozenträngen ein weiteres Lagemaß, mit dem wir uns beschäftigen. Er ist die Bezeichnung für "Gipfelpunkte" (Modi) einer Verteilung. **Gipfel sind solche Meßwerte oder Klassenmitten, deren Häufigkeit größer ist als die der jeweils rechts und links benachbarten**

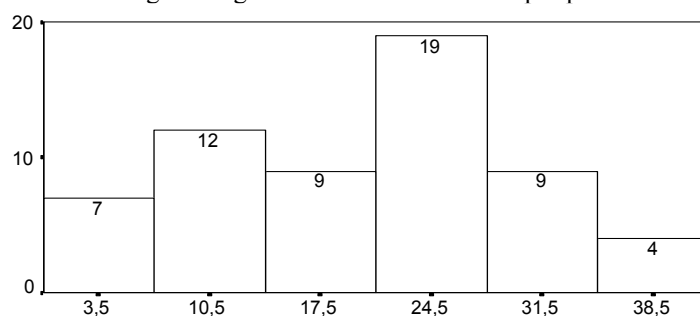
Meßwerte oder Klassenmitten. Menge und Lage der Gipfelpunkte hängen stark von der Definition der Klassengrenzen ab. So erzeugt eine Klassifikation unserer Daten vom Tag der offenen Tür 1985 bei einer Klassenbreite von 3 und 7 unterschiedlich viele Modi, wie die folgenden Graphiken zeigen:

Altersverteilung am Tag der offenen Tür : 5 Gipfelpunkte



Klassenmitten bei Klassenbreite 3

Altersverteilung am Tag der offenen Tür: Zwei Gipfelpunkte



Klassenmitten bei Klassenbreite 7

Während bei der Klassenbreite 3 noch 5 Gipfel vorhanden sind, bleiben nur noch 2 Gipfel bei Klassenbreite 7 übrig. Wie sich hier zeigt, ist der Modus extrem anfällig für Klassifikationsunterschiede. Er scheidet daher für einen Vergleich dann aus, wenn verschiedene Klassenbreiten vorliegen. Weiterhin ist zu betonen, daß der Modus vor allem dann nützlich für einen Vergleich ist, wenn er nur einmal auftritt. Treten mehrere Modi auf, wählt beispielsweise SPSS.10 den kleinsten Meßwert aus, dem ein Modus entspricht. Das hieße für obige Graphik mit den zwei Modi, daß der Wert 10,5 als Modus angegeben werden würde, nicht aber der Wert 24,5, dem der Modus mit der absolut größten Häufigkeit entspricht.

3.3.1.7. Relative, prozentuale und kumulierte Häufigkeiten; Wahrscheinlichkeit

Der Vergleich zwischen Verteilungen aufgrund "absoluter" Häufigkeiten (f_k) ist schwierig, wenn man verschieden große Stichproben einander gegenüberstellt. Um Vergleiche auch zwischen Stichproben und Grundgesamtheiten verschiedenen Umfangs sinnvoll durchführen zu können, ist es nützlich, statt der absoluten Häufigkeiten "relative" Häufigkeiten anzugeben. Hierbei teilt man die absoluten Häufigkeiten einfach durch n (die Menge der Fälle) und schon ist ein Vergleich möglich, allerdings auch wiederum nur auf der Grundlage identischer Klassengrenzen. Die relativen Häufigkeiten sind definiert als

$$\text{Relative Häufigkeit} = \frac{f_k}{n} = \frac{1}{n} * f_k$$

Hierbei gelten die Beziehungen:

$$\sum_{k=1}^l f_k = n ; \quad \frac{\sum_{k=1}^l f_k}{n} = 1$$

Das heißt, die Summe aller relativen Häufigkeiten ist immer gleich 1. k ist hier der Laufindex für die Klassen, der zum Beispiel bei 6 Klassen die Werte 1 bis 6 durchläuft.

Multiplizieren wir die relativen Häufigkeiten mit 100, dann machen wir aus ihnen "prozentuale" Häufigkeiten, also die "Prozente", die oben in der Tabelle schon berechnet wurden.

$$\text{Prozentuale Häufigkeit} = \frac{100 * f_k}{n}$$

Relative Häufigkeiten kennen wir in der Alltagssprache meistens nur als "Wahrscheinlichkeiten": "Ich **wette 10:1**, daß Du nicht gern Statistik lernst, weil nur 1/10 (oder nur 10 %) der Studenten erfahrungsgemäß Spaß an diesem Fach haben".

Die Wahrscheinlichkeit, einen Studenten zu treffen, der das Fach gern hat, müßte bei dieser Wette 1/10, also 0,1 sein oder 10 % sein. Wir sehen, daß die alltagssprachliche Definition der Wahrscheinlichkeit identisch ist mit der Definition der relativen Häufigkeit. Bei der Wahrscheinlichkeit werden die "zutreffenden" Fälle (also im Beispiel die Studenten, die gern Statistik lernen) durch die Menge aller Studenten geteilt, die das Fach lernen; die Formel lautet:

$$p(E) = \frac{\text{Anzahl der für E zutreffenden Fälle}}{\text{Anzahl der insgesamt möglichen Fälle}}$$

Die Formel bedeutet: Die Wahrscheinlichkeit p für das Eintreten des Ereignisses E ist gleich dem Quotienten aus den Fällen, bei denen E eintritt, und allen möglichen Fällen des Experiments.

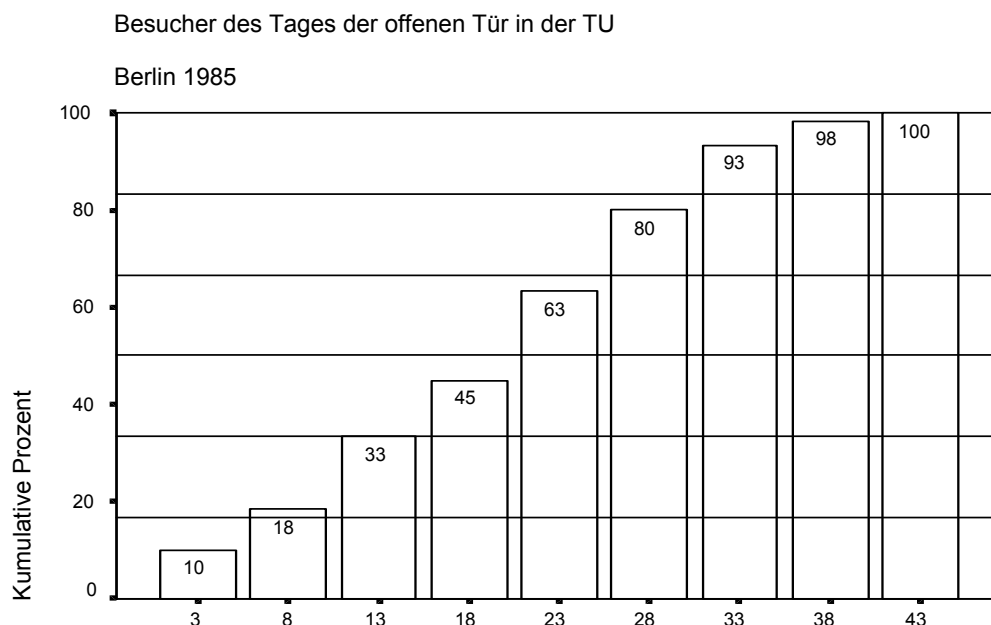
Für die prozentualen Häufigkeiten gilt, daß ihre Summe gleich 100 ist:

$$\left[\frac{\sum_{k=1}^l f_k}{n} \right] * 100 = \frac{\sum_{k=1}^l f_k}{\sum_{k=1}^l f_k} * 100 = 100$$

Weil die Summe der relativen Häufigkeiten gleich 1 ist, leuchtet es ohne weiteres ein, daß die Summe der prozentualen Häufigkeiten immer gleich 100 ist. Allerdings erhalten wir wegen der Rundungsfehler „nach dem Komma“ nur selten den exakten Wert von 100 %, wenn wir Prozente - wie bei den oben gezeigten kumulierten Prozents - addieren.

Die folgende Grafik zeigt die kumulierten Prozente der Altersangaben vom Tag der offenen Tür 1985. Eingezeichnet sind waagerecht die Teilungslinien, die wir bei Sextilen erhalten würden; man kann in der Graphik sehr deutlich den Unterschied zwischen Prozenträngen und N-tilen sehen: Die Prozentränge entsprechen den kumulierten Prozents an der Obergrenze der jeweiligen Klasse, die N-tilen – je nach

Konvention – der Klassenmitte jener Klasse, in die der N-til-Wert fällt, oder einem bestimmten Wert in dieser Klasse:



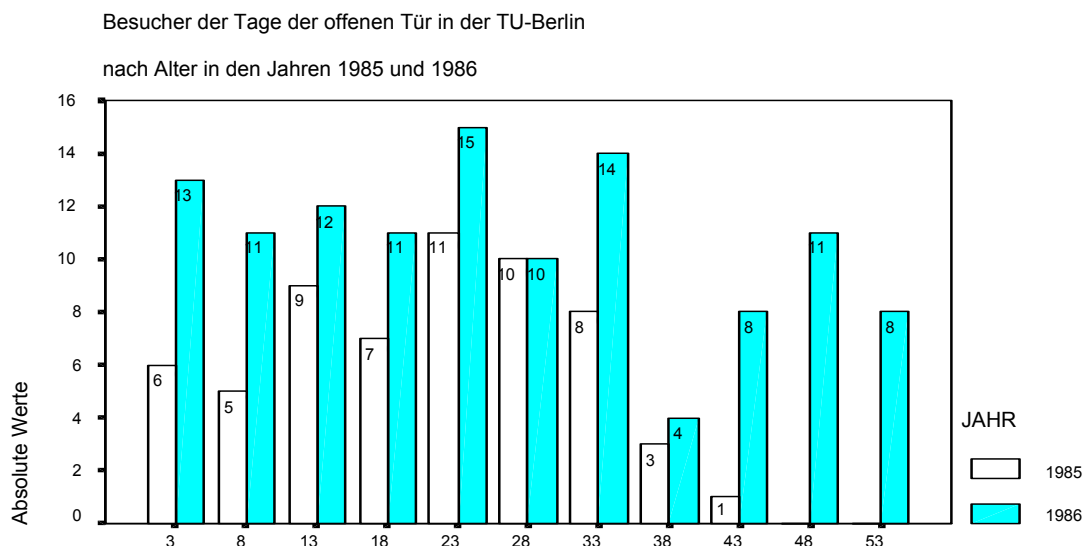
Um nun mit prozentualen Häufigkeiten vergleichend umzugehen, stellen wir unsere Daten vom Tag der offenen Tür aus Jahr 1985 anderen Daten aus dem Jahr 1986 gegenüber, die ebenfalls am Tag der offenen Tür erhoben wurden:

Geordnete Angaben von Besuchern des Tages der offenen Tür über ihr Lebensalter vom 1.6.1986:

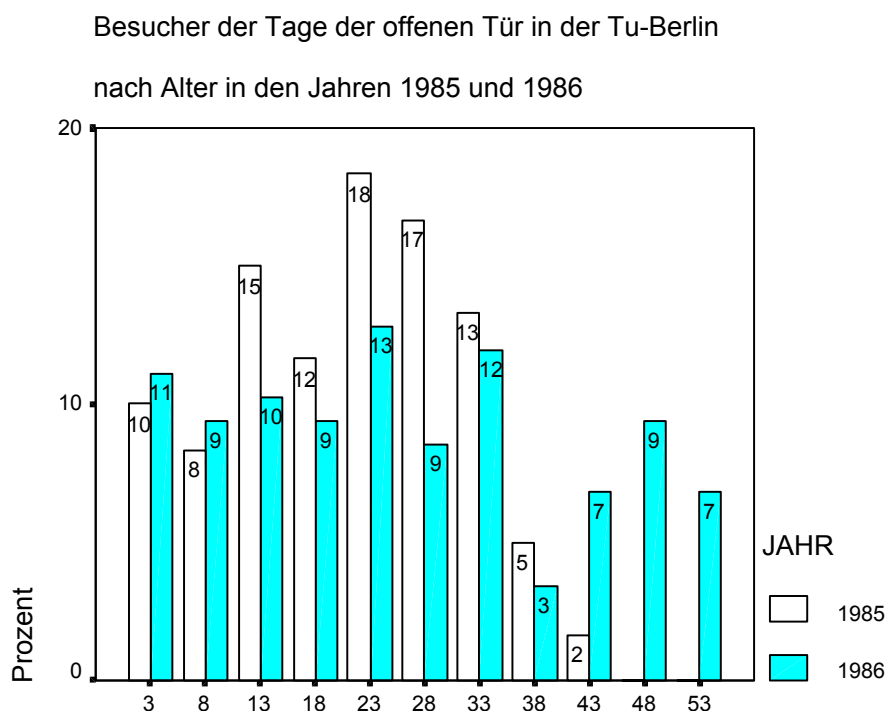
0,0,0,0,1,1,1,2,2,3,3,4,4,5,5,6,7,8,8,8,8,8,9,10,10,11,11,12,12,13,13,13,13,14,14,15,15,17,17,17,17,17,18,18,19,19,20,20,20,21,21,21,22,22,22,22,24,24,24,24,24,25,25,26,26,28,28,29,29,29,29,31,32,32,32,32,33,33,33,34,34,34,34,34,36,39,39,39,40,40,40,40,41,42,44,44,45,45,45,45,45,45,46,49,49,49,49,50,51,51,52,52,53,53,54

Zunächst erfolgt der Vergleich auf der Grundlage absoluter Zahlen:

Vergleich der Altersverteilungen der Besucher am Tage der offenen Tür 1985 und 1986
in der Technischen Universität (aufgrund von klassifizierten Werten mit der Klassenbreite 5):



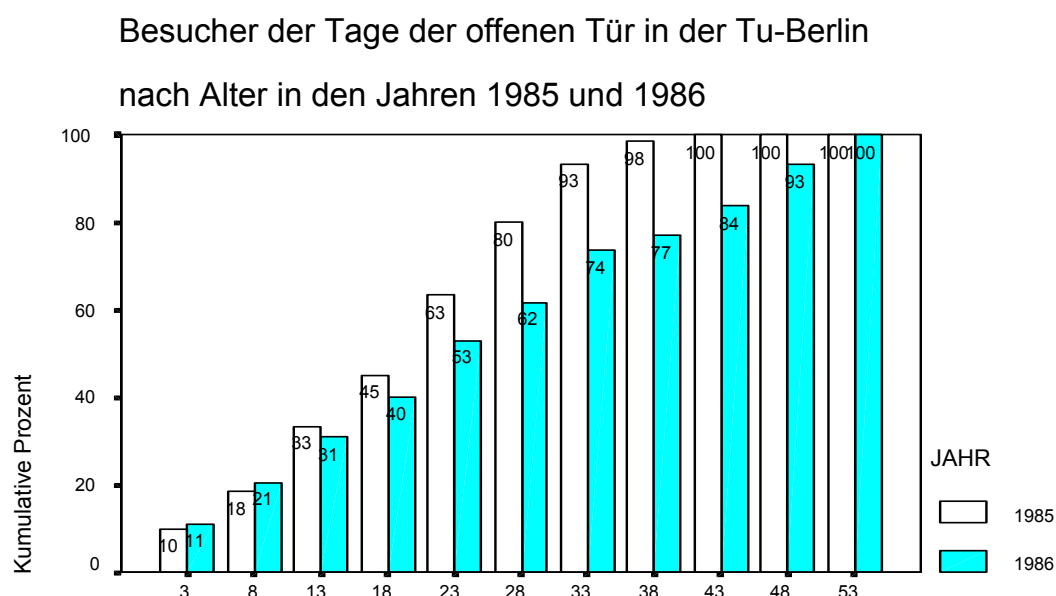
Wie man sieht, ist der Vergleich der absoluten Häufigkeiten von beiden Aufstellungen der Jahre 1985 und 1986 nicht sehr fruchtbar, weil wir z.B. zwar feststellen können, daß in der 5. Klasse (der ca. 23 - jährigen) 1985 11 Personen kamen, 1986 waren es 15. Wir können aber aus dieser Feststellung noch nicht schließen, daß 1986 die Häufigkeit in der Altersgruppe „23“ angestiegen ist, denn diese scheinbar geringe Zunahme wird durch die Information zweifelhaft, daß 1986 insgesamt fast doppelt soviel wie 1985 kamen, nämlich 117 Personen statt 60: Vergleicht man die Prozente, dann wird aus der scheinbaren Zunahme ein dramatischer Rückgang der Besucherzahl in dieser Altersgruppe. Vergleichen wir deshalb einmal die Prozentzahlen:



Es zeigt sich, daß in der Gruppe der ca. 23-jährigen 1986 nur noch 13% der Besucher waren, 1985 waren es immerhin 18%.

Zugleich zeigt der Vergleich auch, daß sich die Besucher 1986 auf eine größere Spannweite (Range) verteilen. Der älteste Besucher 1985 war 40, 1986 war er über 50. Die größere Streuung im Alter 1986 führt dazu, daß bei den älteren Gruppen eine gewaltige Zunahme zu verzeichnen ist, bei vielen jüngeren Gruppen eine Abnahme.

Eine weitere Möglichkeit, empirische Verteilungen zu vergleichen, ist der Vergleich der kumulierten prozentualen Häufigkeiten. Hierbei wird, wie oben schon ausgeführt, bei jeder der Verteilungen die prozentuale Häufigkeit der 2. Klasse zur prozentualen Häufigkeit der 1. Klasse hinzuaddiert, die der 3. Klasse zu dieser Summe und so fort. Dieses Verfahren nennt man "Aufwärtskumulation". Es endet beim Wert "100%". Das folgende graphische Beispiel ist durch Aufwärtskumulation der prozentuierten Häufigkeiten der Besucher in den Jahren 1985 und 1986 beim Tag der offenen Tür gewonnen, wobei wieder die Klassenbreite 5 zugrundeliegt.



Die Interpretation kumulierter Häufigkeiten ist folgende (die folgenden Beispiele sind der obigen Graphik entnommen): Wenn man davon ausgeht, daß die Klassengrenzen von über 0 bis 5, von über 5 bis 10 etc. verlaufen, und daß es sich in der obigen Graphik um aufgerundete Klassenmitten handelt, dann kann man aufgrund der Graphikwerte sagen:

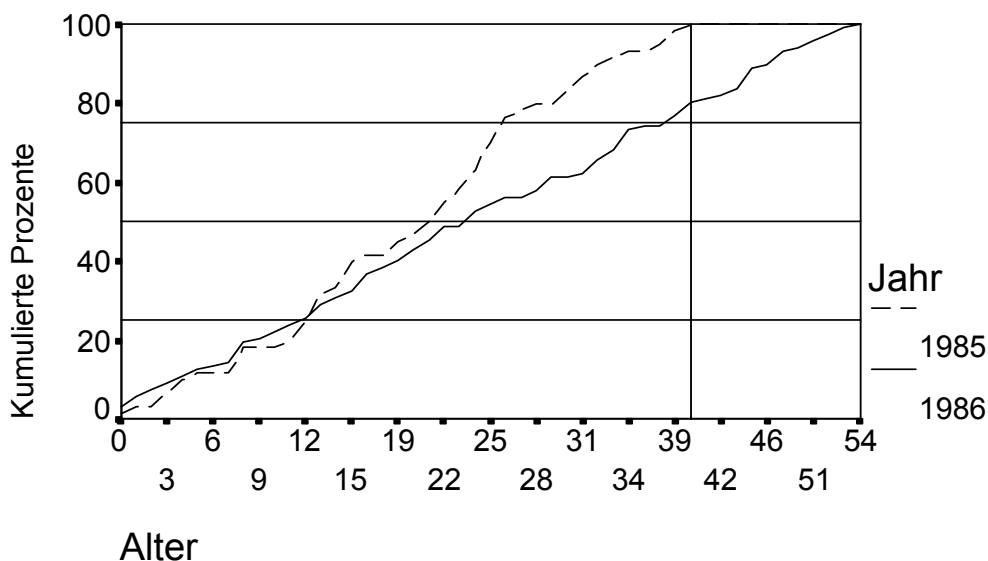
1985 waren 80 % der Besucher bis zu 30 Jahre alt. 1986 waren es nur 62 %.

1985 waren 63 % der Besucher bis zu 25 Jahre alt, 1986 waren es nur 53 %

1985 waren 7 % älter als 35 Jahre, 1986 waren 26 % älter als 35 Jahre

1985 waren 2 % älter als 40 Jahre, 1986 waren 23 % älter als 40 Jahre.

Für die Darstellung kumulierter Verteilungen benützt man in der Regel **Summenpolygone**, die bei Aufwärtskumulieren durch Streckenzüge von der Untergrenze der kleinsten Klasse zur Obergrenze der höchsten Klasse nach Maßgabe der addierten Werte gewonnen werden. Die addierten Prozente beginnen bei 0 und enden bei 100 %, wenn die Obergrenze der höchsten Klasse erreicht ist. Die beiden Altersverteilungen von den Tagen der offenen Tür in der TUB 1985 und 1986 werden in der folgenden Liniengrafik verglichen, die die gleiche Information enthält wie die oben wiedergegebene Balkengrafik mit dem Vergleich der kumulierten, prozentualen Häufigkeiten. Unten ist allerdings die Klassenbreite 1 gewählt worden:



Die waagerechten Linien in der Grafik entsprechen den Quartilen. Die Grafik zeigt die Unterschiede der Verteilungen sehr deutlich: Besonders beim dritten Quartil ergeben sich große Differenzen des Alters.

3.3.1.8. Zur unterschiedlichen Leistung der verschiedenen Mittelwerte.

Jedes Maß der zentralen Tendenz – so wurde oben ausgeführt – versucht, alle Werte einer Verteilung durch einen einzigen Wert „ersetzen“. Dieser Ersatz ist immer ein unvollkommener Ersatz. Daher ist es sinnvoll die Fehler zu systematisieren, die man bei den verschiedenen Maßen der zentralen Tendenz macht, wenn man sie auf empirische Verteilungen anwendet. Dadurch werden die Maße der zentralen Tendenz in gewisser Weise in ihrer Brauchbarkeit und Unbrauchbarkeit einschätzbar.

Wie Diehl und Kohn in ihrem Lehrbuch zur deskriptiven Statistik (a.a.O. S. 76 ff) ausführen, wird für jede Art des Mittelwertes ein bestimmter Fehler, den man mit der Gleichsetzung aller Meßwerte mit einem beliebigen Wert „a“ macht, minimiert.

Wenn man z. B. behauptet, in der Verteilung der Meßwerte 1,3,6,9,11 käme nur der Meßwert 6 (hier das arithmetische Mittel) vor, dann hätte man ein gutes Maß für die damit verursachten Fehler in der Summe der Abweichungsquadrate der Meßwerte vom arithmetischen Mittel. (Die Summe der Abweichungen zwischen arithmetischem Mittel und Meßwerten (unquadriert) allein wäre kein gutes Maß, weil sie für das arithmetische Mittel Null ergibt und für Werte, die größer als das arithmetische Mittel sind, negative Summen ergibt).

Die Summe der Abweichungsquadrate (SAQ) ist:

$$SAQ = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{Wir berechnen: } (1-6)^2 + (3-6)^2 + (6-6)^2 + (9-6)^2 + (11-6)^2 = 68. \text{ Diese Zahl ist kleiner als}$$

alle Summen, die wir erzielen, wenn wir in die Formel der Abweichungsquadrate einen größeren oder kleineren Wert anstelle von \bar{x} eingesetzt hätten. Wir probieren das aus:

$$(1-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (11-7)^2 = 73; \quad 73 > 68$$

$$(1-4)^2 + (3-4)^2 + (6-4)^2 + (9-4)^2 + (11-4)^2 = 88; \quad 88 > 68$$

Daher gilt:

Für das arithmetische Mittel ist der „minimierte Fehler“

$$\sum_{i=1}^n (x_i - a)^2; \text{ d.h. wenn } a = \bar{x}, \text{ ist der nebenstehende Ausdruck ein Minimum.}$$

Für den Median ist der „minimierte Fehler“

$\sum_{i=1}^n |x_i - a|$; d.h. wenn $a = Md$ ist, ist der nebenstehende Ausdruck ein Minimum.

Für den Modus ist der „minimierte Fehler“

$n - f_a$; d. h., wenn „a“ der Index der Modalklasse ist, ist der nebenstehende Ausdruck ein Minimum

(vorausgesetzt, es gibt nur einen Modus).

3.3.2. Streuungsmaße: Range, Quartilabstand, durchschnittliche, absolute und quadrierte Abweichungen vom Mittelwert, Varianz und Standardabweichung

Während die bisherigen Beobachtungen sich darauf bezogen, welche Meßwertpunkte zur Charakterisierung von empirischen Verteilungen herangezogen werden konnten, wird jetzt die Betrachtung auf die Heranziehung von Intervallen zur Charakterisierung von Verteilungen ausgeweitet. Wir gehen insofern von den **Lagemaßen** zu den Streuungsmerkmalen oder Streuungsmaßen über. (Man spricht in diesem Zusammenhang auch von Lokalisationsmaßen und Dispersionsmaßen, das klingt gelehrter, ist aber nur eine lateinische Bezeichnung für den gleichen Sachverhalt, nämlich für Lage- und Streuungsmaße).

Der **Range** („Eingriffsspielraum; Spannweite“) stellt die Differenz zwischen dem größten und dem kleinsten Meßwert dar. Er gibt an, wie groß dieser Bereich ist. Er ist ein primitives Streuungsmaß, das nur bei sehr großen Stichproben herangezogen werden sollte, und dann nur in Verbindung mit anderen Streuungsmaßen.

Ein weiteres einfaches Streuungsmaß ist der "Quartilabstand", also die Differenz vom 1. zum 3. Quartil.

Er ist für die Daten vom Tag der offenen Tür von 1985 gleich $26 - 12,25 = 13,75$; (Halber Quartil-Abstand = 6,875). (Berechnet mit SPSS.10).

Da der Quartilabstand nur die Lage der mittleren 50% aller Meßwerte charakterisiert, die Verteilung der äußeren 50% aber ignoriert, suchte man in der Vergangenheit nach anderen Charakteristika für die Streuung, die alle Meßwerte einbeziehen. Hierbei fand man zwei Maße, von denen sich eines schließlich durchsetzte.

Zunächst versuchte man, das **arithmetische Mittel der Differenz aller Meßwerte vom arithmetischen Mittel** als Streuungsmaß zu benutzen. Der Nachteil hierbei war, daß die Differenz aller Meßwerte vom arithmetischen Mittel gleich Null ist (siehe oben). Nimmt man aber von den einzelnen Differenzen nur den (immer positiven) „Betrag“, dann erhält man ein Streuungsmaß, daß als AD-Streuung in der Literatur auftaucht:

$$AD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Die senkrechten Striche in der Formel sind die sogenannten „Betragsstriche“, die jede der hier ausgerechneten Differenzen in einen positiven Betrag verwandeln.

Für unsere Daten vom Tag der offenen Tür betragen die so gebildeten durchschnittlichen Abweichungen von \bar{x} :

$$AD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = 8,5767 \quad (\text{für 1985})$$

Wir können nun in diese Formel statt des arithmetischen Mittels auch versuchsweise den Median anstelle von \bar{x} einsetzen und berechnen, welches Ausmaß die dann resultierende Streuung hat. Das Ergebnis ist:

$$AD_{Md} = \frac{\sum_{i=1}^n |x_i - M_d|}{n} = 8,5$$

Es zeigt sich, daß dieses Ergebnis kleiner ist als das der mit dem arithmetischen Mittel erzielte.

Hiernach wäre es zweckmäßig (wenn die Daten besser zur Verwendung des Medians statt des arithmetischen Mittels bei Berechnung eines Mittelwertes geeignet erscheinen) als Maß der Streuung die durchschnittliche absolute Abweichung vom Median zu wählen, was allerdings in der Praxis selten geschieht, da diese Berechnung in Computerprogrammen für Statistik nicht als Routine vorgesehen ist..

Sucht man nun nach einem Streuungsmaß, das für \bar{x} eine analoge Minimierungsbedingung erfüllt, dann findet man es in der Varianz bzw. in der Standardabweichung (siehe hierzu unsere Ausführungen zum minimierten Fehler).

Die Varianz von Grundgesamtheiten ist definiert als arithmetisches Mittel der **quadrirten** Abweichungen der Meßwerte x_i vom arithmetischen Mittel der Grundgesamtheit μ . (N ist die Zahl der Elemente der Grundgesamtheit.)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Für Stichproben ist aus Gründen einer möglichst erwartungstreuen "Parameter"-Schätzung die Formel abzuwandeln in

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Parameter sind - in der Terminologie bestimmter Lehrbücher der Statistik - Bezeichnungen für Lokalisations- und Streuungsmaße, wenn diese sich auf die **Grundgesamtheit** beziehen. Sie werden mit kleinen griechischen Buchstaben bezeichnet. Ist die **Stichprobe** gemeint, wird oft von **"Kennwerten"** gesprochen, die mit lateinischen Buchstaben belegt werden.

In der Stichprobe ist n die Stichprobengröße. **Im Nenner der Stichprobenvarianz steht n-1 statt n.** Durch n-1 im Nenner wird die Stichprobenvarianz etwas größer. Die Gefahr der **Unterschätzung** der Grundgesamtheitsvarianz soll sich hierdurch verringern.

Da nun die Beträge der Varianz wegen der Quadrierung sehr groß sind

$$s^2 = 105,85 \quad (\text{für die Daten vom Tag der offenen Tür 1985})$$

verwendet man als optisch anschauliches Streuungsmaß die Quadratwurzel der Varianz, die auch **"Standardabweichung"** genannt wird. Die Stichprobenstandardabweichung hat die Formel:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Diese Formel ist zwar einfach und noch relativ anschaulich, in der Praxis jedoch umständlich zu berechnen, wenn man sie mit Taschenrechner oder auf dem Papier berechnet (siehe unten), weil man vorher das arithmetische Mittel ausrechnen muß und dann erst die Standardabweichung ermitteln kann. Die Formel entspricht dem unten demonstrierten „Rechengang 1“.

Es ergibt sich für unsere Daten vom Tage der offenen Tür 1985: $s = 10,289$

3.3.2.1. Berechnungsweise für die Varianz und die Standardabweichung nach einer vereinfachenden Formel

Versuchen wir die Varianz für irgendeinen Datensatz "zu Fuß" (Rechengang 1) zu berechnen, dann bemerken wir, daß

- (1) zunächst \bar{x} bestimmt werden muß
- (2) die quadrierten Differenzen von x_i und \bar{x} sehr große Werte erbringen,

so daß sehr viel Mühe entsteht. Hieraus entwickelt sich das Bedürfnis nach vereinfachter Berechnungsweise (Rechengang 2). Man formte die Formel für s^2 (und s) so um, daß nur noch folgende Ausdrücke berechnet werden müssen, um s^2 (bzw. s) zu ermitteln:

- (1) n
- (2) $\sum x_i$
- (3) $\sum x_i^2$

Hat man diese Ausdrücke ermittelt, dann kann man sie leicht in die vereinfachte Formel (des Rechengangs 2) einsetzen:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left[\sum_{i=1}^n x_i \right]^2}{n - 1}$$

$$= \frac{n \sum_{i=1}^n x_i^2 - \left[\sum_{i=1}^n x_i \right]^2}{n(n-1)}$$

Um beide Rechengänge zu demonstrieren, sollen im folgenden die Varianz und die Standardabweichung (für die ersten 10 Werte vom Tage der offenen Tür in der TU 1985) berechnet werden:

Rechenbeispiel für die zwei Rechengänge der Ermittlung von Varianz und Standardabweichung

Rechengang 1	Rechengang 2
$x_1 = 26.00$ $x_2 = 39.00$ $x_3 = 5.00$ $x_4 = 40.00$ $x_5 = 21.00$ $x_6 = 22.00$ $x_7 = 31.00$ $x_8 = 15.00$ $x_9 = 38.00$ $x_{10} = 30.00$ Berechne \bar{x} $\sum x_i = 267.00$ $n = 10.00$ $\bar{x} = \sum x_i / n = 26,70$ Berechne $x_i - \bar{x}$ $26.00 - 26.70 = -0,70$ $39.00 - 26.70 = 12,30$ $5.00 - 26.70 = -21,70$ $40.00 - 26.70 = 13,30$ $21.00 - 26.70 = -5,70$ $22.00 - 26.70 = -4,70$ $31.00 - 26.70 = 4,30$ $15.00 - 26.70 = -11,70$ $38.00 - 26.70 = 11,30$ $30.00 - 26.70 = 3,30$ $(x_1 - \bar{x})^2 = 0,49$ $(x_2 - \bar{x})^2 = 151,29$ $(x_3 - \bar{x})^2 = 470,89$ $(x_4 - \bar{x})^2 = 176,89$ $(x_5 - \bar{x})^2 = 32,49$ $(x_6 - \bar{x})^2 = 22,09$ $(x_7 - \bar{x})^2 = 18,49$ $(x_8 - \bar{x})^2 = 136,89$ $(x_9 - \bar{x})^2 = 127,69$ $(x_{10} - \bar{x})^2 = 10,89$	Es werden die folgenden Ausdrücke gebildet: $\sum x_i = 267.00$ $\sum x_i^2 = 8277,00$ $n = 10.00$ Aus diesen wird der Ausdruck s^2 gebildet: $s^2 = \frac{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}{n * (n - 1)}$ $= \frac{10 \times 8277 - 71289}{10 * (9)}$ $= 127,57$ Die Wurzel daraus ist s $s = \sqrt{127,57} = 11,3$

$$\sum (X_i - \bar{x})^2 = 1148,10$$

VARIANZ =

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = 127,57$$

Standardabweichung=S=Wurzel aus Varianz= 11.3

Mit Hilfe von Taschenrechnern, die eine Summenfunktion haben, ist nun heutzutage die Standardabweichungsberechnung stark vereinfacht. Man gibt z.B. beim HP-Rechner fortlaufend die Werte für x_i einfach mit dem Befehl $\Sigma+$ ein, und nach jeder dieser Eingaben bildet der Rechner automatisch auf den sogenannten "Statistik-Registern" die Ausdrücke: $\sum f_k$ ($=n$), $\sum x_i$ und $\sum x_i^2$, wie bei Rechengang 2. Dieser 2. Rechengang ist daher der wesentlich ökonomischere, wenngleich unanschaulichere.

Bei klassifizierten Häufigkeitsverteilungen **multipliziert man die jeweiligen Meßwerte (bzw. Meßwertquadrate) immer mit der Häufigkeit jener Klasse „k“, in der sie als Klassenmittelpunkte auftauchen:**

$$s^2 = \frac{\sum_{i=1}^n f_k x_i^2 - \frac{1}{n} \left[\sum_{i=1}^n f_k x_i \right]^2}{n-1}$$

Beispiel: Wir klassifizieren unsere Daten vom Tag der offenen Tür 1985 mit der Klassenbreite 10 und erhalten:

OFFTUE10

	Häufigkeit
Gültig 5,00	11
15,00	16
25,00	21
35,00	11
45,00	1
Gesamt	60

Wir berechnen:

$$n = 60$$

$$\sum_{k=1}^l f_k x_i = 11*5 + 16*15 + 21*25 + 11*35 + 1*45 = 1250$$

$$\sum_{k=1}^I f_k x_i^2 = 11 \cdot 5^2 + 16 \cdot 15^2 + 21 \cdot 25^2 + 11 \cdot 35^2 + 1 \cdot 45^2 = 32500$$

und erhalten

Statistiken

OFFTUE10		
N	Gültig	60
	Fehlend	0
Mittelwert		20,8333
Standardabweichung		10,4625
Varianz		109,4633

3.3.3. Daten auf dem Ordinal- und Nominal-Niveau. Lage- und Streuungsmaße

Von den oben beschriebenen Lage- und Streuungsmaßen fallen beim Ordinalniveau alle die fort, die die Information über Abstände voraussetzen. Es können bei **Ordinalskalen** nicht mehr berechnet werden:

\bar{x} (arithmetisches Mittel)
 s^2 (Varianz)
 S (Standardabweichung)

Noch berechnet werden können: Absolute und relative sowie prozentuale Häufigkeiten, Prozentränge, der Median und N-tile. Allerdings wäre die Berechnung des mittleren Quartil-**Abstands** nicht sinnvoll, da hier ja keine Abstands-Informationen vorliegen. Der Modus ist ebenfalls berechenbar.

Bei Nominalskalen bleiben absolute, relative und prozentuale Häufigkeiten sowie der Modus erhalten. Median, N-tile und Prozentränge fallen fort, da die Anordnung der Kategorien beliebig ist, auf die sich die Häufigkeiten beziehen.

4. Wahrscheinlichkeit und theoretische Verteilungen

Unabhängig von inhaltlichen Fragen der Deutung liefert die theoretische "schließende" Statistik zahlreiche Probleme der formalen Deutung von empirischen Stichprobendaten, deren Lösung nur mit Hilfe von wahrscheinlichkeitstheoretischen Überlegungen auf der Grundlage von theoretischen Verteilungen gefunden werden kann.

4.1. Begriff Wahrscheinlichkeit

Bei der Erörterung der relativen Häufigkeit hatten wir schon den Bruch kennengelernt, der die relative Häufigkeit definiert, nämlich

Anzahl der Fälle einer bestimmten Art
 Anzahl aller möglichen Fälle

Beispiel: In unseren Daten aus 1985 waren 7 Personen bis unter 6 Jahre alt. Insgesamt erhielten wir 1985 60 Personen-Altersangaben. Die relative Häufigkeit in der genannten Altersklasse beträgt damit

$$\frac{7}{60} = 0,117 \text{ aufgerundet}$$

Würde man nun aus der Menge der 60 Personen zufällig eine Person herausgreifen, dann wäre die Wahrscheinlichkeit, daß sie im Alter von 0 bis unter 6 Jahren ist, gleich sieben Sechzigstel. Da im gleichen

Datensatz 12 der Personen ein Alter bis unter 12 Jahre haben, wäre die Wahrscheinlichkeit, eine Person mit unter 12 Jahren zufällig auszuwählen, gleich $12/60 = 0,20$.

Da die älteste Person 40 ist, wäre die Wahrscheinlichkeit, eine Person mit "bis unter 42 Jahren" Lebensalter zufällig auszuwählen, gleich 1.

Wir können noch viele weitere solcher Betrachtungen anstellen, sollten diese aber erst vornehmen, wenn wir die grundsätzlichen Regeln des Rechnens mit Wahrscheinlichkeiten kennengelernt haben.

Diese lauten:

1. Die Wahrscheinlichkeit eines Ereignisses A kann höchstens 1 sein.
2. Ist sie gleich 1, so ist das Ereignis "sicher". Ist sie gleich 0, ist das Ereignis unmöglich.
3. Einander ausschließende Ereignisse haben Wahrscheinlichkeiten, die "additiv" sind.
4. Einander überlappende Ereignisse haben Wahrscheinlichkeiten, die „additiv“ sind, minus der Wahrscheinlichkeit des Überlappungsbereichs.
5. Voneinander unabhängige Ereignisse haben eine Wahrscheinlichkeit, gleichzeitig aufzutreten, die ihrem Produkt entspricht.
6. Sind zwei Ereignisse A und B voneinander abhängig, dann ist die Wahrscheinlichkeit, daß B unter der Voraussetzung eintritt, daß A zuvor eingetreten ist, gleich dem Quotienten aus der Wahrscheinlichkeit, daß A und B eintritt, geteilt durch die Wahrscheinlichkeit, daß A eingetreten ist !

Mathematisch ausgedrückt, sehen die obigen Regeln wie folgt aus:

1. $0 \leq P(A) \leq 1$

2. Wenn $P(A) = 1$, ist das Ereignis „sicher“. Wenn $P(A) = 0$, ist das Ereignis „unmöglich“. So ist es sicher, daß beim Würfeln irgendeine der Zahlen oben liegt, unmöglich ist, daß keine der Zahlen oben liegt.

3. Sind A und B einander ausschließende Ereignisse, so ist die Wahrscheinlichkeit, daß das eine oder das andere Ereignis eintritt gleich $P(A+B) = P(A) + P(B)$.

Zur dritten Regel ist am Beispiel unserer Daten aus 1985 festzustellen, daß die relative Häufigkeit im Alter bis zu 12 gleich 0,20 war. Die relative Häufigkeit im Alter von 36 bis unter 42 war 0,0667. Beide Altersklassen schließen sich gegenseitig aus, denn niemand kann gleichzeitig unter zwölf und älter als 36 sein. Infolgedessen ist die Wahrscheinlichkeit, eine Person, die **entweder** unter 12 **oder** 36 und älter ist, zufällig auszuwählen, gleich $0,20 + 0,0667 = 0,2667$.

4. Sind A und B einander nicht ausschließende Ereignisse, dann ist die Wahrscheinlichkeit, daß A oder B eintritt, gleich $P(A+B) = P(A) + P(B) - P(A \cap B)$.

Zur der vierten Regel kann man anhand der Daten aus 1985 fragen, wie wahrscheinlich es ist, daß eine Person entweder unter 17 Jahre alt ist, oder zwischen 10 und (einschließlich) 25 Jahre alt. Die Wahrscheinlichkeit, daß eine Person unter 17 ist, beträgt für die Daten aus 1985 0,4 (oder 40 %). Die Wahrscheinlichkeit, daß eine Person zwischen 10 und 25 Jahre alt ist, beträgt 0,517 (oder 51,7%). Die beiden Bereiche überlappen sich im Intervall von 10 bis 16. Die Wahrscheinlichkeit daß eine Person in dieses Intervall fällt, beträgt 0,217. Wir können daher die Wahrscheinlichkeit, daß eine Person entweder unter 17 Jahre alt ist oder zwischen 10 und (einschließlich) 25 Jahre durch die Rechnung $0,4 + 0,517 - 0,217 = 0,7$ beantworten. Die gesuchte Wahrscheinlichkeit beträgt 70 %.

5. Sind A und B zwei voneinander unabhängige Ereignisse, dann ist die Wahrscheinlichkeit, daß sie gemeinsam eintreffen, gleich $P(AB) = P(A) * P(B)$.

Zur fünften Regel könnte man fragen, wie wahrscheinlich es ist, daß beim Tag der offenen Tür zweimal hintereinander jemand kommt, der männlich ist. Wenn die Annahme zutrifft, daß in der Grundgesamtheit 50%

der Menschen männlich waren, und wäre das Geschlecht der jeweils nachfolgenden Person unabhängig vom Geschlecht der vorhergehenden Person, dann könnte man die gesuchte Wahrscheinlichkeit mit $p_{m,m} = 0,5^2$ bestimmen, also dem Produkt aus $0,5 \cdot 0,5$.

6. Ist das Eintreten des Ereignisses B davon abhängig, daß das Ereignis A vorher eingetreten ist (bedingte Wahrscheinlichkeit), dann gilt:

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad \text{Ein Beispiel hierfür wäre das Ziehen von Kugeln aus einer Urne mit 6}$$

durchnummerierten Kugeln mit den Zahlen 1 bis 6 ohne Zurücklegen.

Wenn ich hier wissen will, wie wahrscheinlich es ist, zweimal eine 6 zu ziehen, dann ist die Lösung nicht mehr 6^2 sondern Null, da die Wahrscheinlichkeit, eine zweite 6 zu bekommen, gleich Null wird, wenn die 6 schon beim ersten Male gezogen wurde.

Eine nochmalige Übersicht über die Formeln und grundlegenden Definitionen der Wahrscheinlichkeitsrechnung findet man in diesem Skript in einer weiteren Tabelle in Kapitel 4.4.2.5., wo auch noch weitere Beispiele geliefert werden.

4.2 Theoretische Verteilungen

4.2.1 Allgemeines

Wie bei den empirischen Verteilungen müssen wir bei den theoretischen Verteilungen unterscheiden, welche Voraussetzungen gegeben sein müssen, um eine solche Verteilung zu definieren. Zu diesen Voraussetzungen zählen u.a. Eigenschaften von Variablen, die wir als „diskret“ oder „stetig“ bezeichnen.

4.2.2 Diskrete Ereignisse

Wir hatten oben schon zwischen stetigen und diskreten Merkmalen unterschieden. Dabei waren die diskreten Merkmale solche, für die nur bestimmte, aber nicht alle Werte einer Skala auftreten können. Wären in einer Statistik zum Beispiel die Tage der offenen Tür in einer Woche als "Montag", "Dienstag" ... etc. definiert, würden die Besuchszeiten nur hiernach erfaßt, dann würde man die Besuchszeiten nicht mit einer stetigen Funktion abbilden können.

4.2.3 Stetige Ereignisse

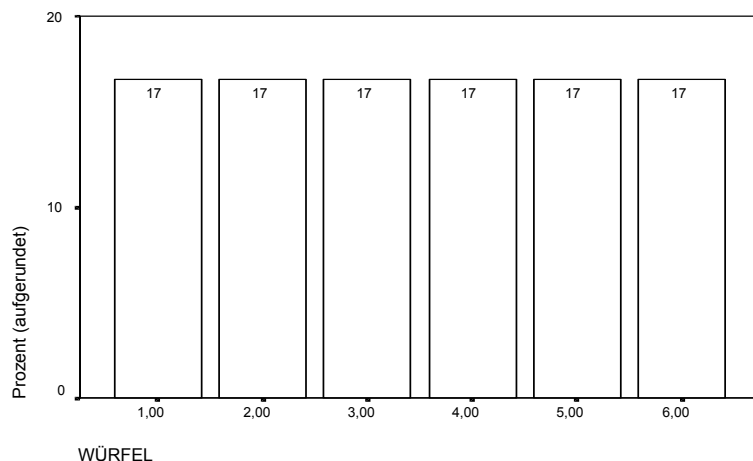
Das Lebensalter und das Einkommen, mit denen wir uns oben beschäftigt haben, sind im Prinzip als stetige Variablen definierbar, wenn auch in der Praxis nur bestimmte Meßpunkte genannt werden. Um zu stetigen Ereignissen zu kommen, ist mindestens Intervallniveau der Messung erforderlich. Für stetige Funktionen gilt zusätzlich, daß sie differenzierbar sein müssen (Steigung muß feststellbar sein).

4.3. Beispiele für theoretische Verteilungen

4.3.1 Gleichverteilung - diskret

Wir greifen auf das eingangs referierte Würfelbeispiel zurück. Ein Würfel habe sechs Flächen und sei vom Gewicht her gleichmäßig. Jede der Flächen hat die "theoretische" Wahrscheinlichkeit von 1:6, nach einem Wurf 'oben zu liegen'.

Wir können - wie oben schon getan - für diesen Fall die Wahrscheinlichkeiten graphisch darstellen:

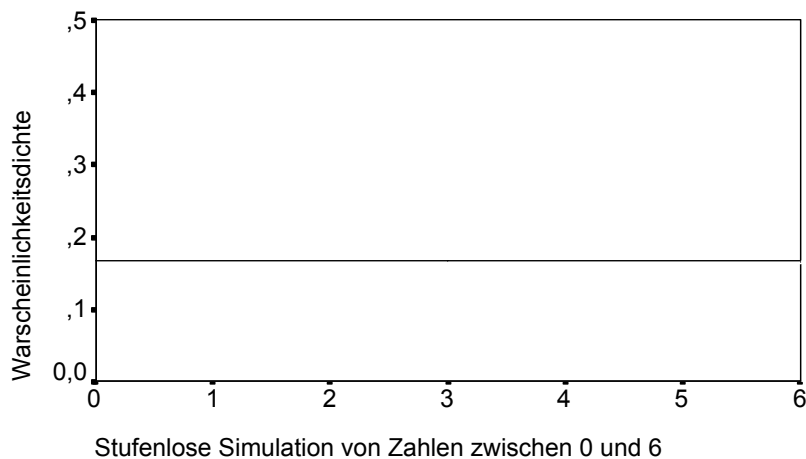


Jedem einzelnen der möglichen Ereignisse wird eine Strecke mit der Länge "ein Sechstel" ($\approx 17\%$) zugeordnet. Die Summe dieser Strecken ist gleich 1.

4.3.2 Gleichverteilung - stetig

Hätten wir nun anstelle des Würfels einen Zufallszahlengenerator, der uns im Bereich 0 bis einschließlich 6 Zahlen produziert, die beliebig fein differenziert werden können, also beliebig viele Stellen nach dem Komma haben können, dann müßten wir die o.a. Darstellung umwandeln in folgende Graphik:

Gleichverteilung in sechs Intervallen auf einer Ratioskala:



Das Besondere dieser Darstellung liegt darin, daß nicht die Strecken, sondern die Fläche unter der Kurve $f(x)$ als Produkt von

$$6 * \frac{1}{6} = 1$$

die Zahl 1 ergibt.

Die Kurve $f(x)$ (hier die Parallele zur x-Achse) heißt deshalb auch nicht "Wahrscheinlichkeit", sondern **Wahrscheinlichkeits-Dichte**. Den einzelnen Zufallszahlen in dem Intervall von 0 bis 6 kann keine einzelne Wahrscheinlichkeit mehr zugeordnet werden, sondern nur noch Intervallen wie z.B. dem Intervall

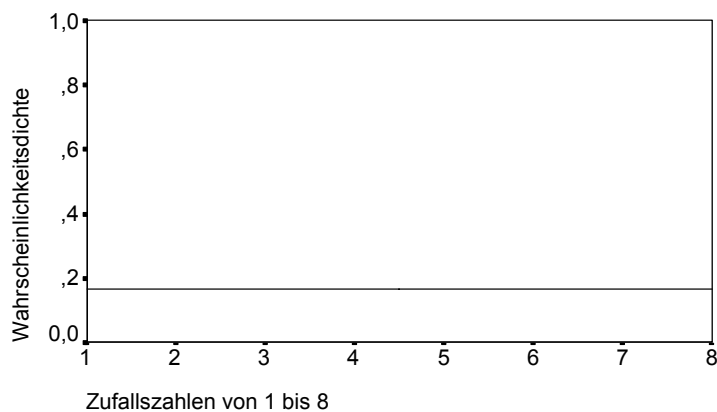
4 bis 5,5

Da die Länge dieses Intervalls gleich 1,5 ist, ist die Wahrscheinlichkeit, einen Zufallswert in diesem Intervall zu bekommen, gleich

$$1,5 * \frac{1}{6} = \frac{1,5}{6} = 0,25$$

Für den Fall, daß unser Intervall, für das die Gleichverteilung definiert ist, nicht von Null begrenzt wird, führen wir als Grenzen allgemein die Untergrenze a und die Obergrenze b ein. In der folgenden Graphik sind

$a = 1$ und
 $b = 8$.



Eine allgemeine Gleichung für die "Gleichverteilung" lautet dann:

$$f(x) = \frac{1}{b-a} \quad \text{für } a < x < b$$

Der Mittelwert ist für diese Gleichverteilung mit

$$\mu = \frac{a+b}{2}$$

gegeben. Die Varianz kann nach der Formel

$$\sigma^2 = \frac{(b-a)^2}{12}$$

berechnet werden. Entsprechend ist die Standardabweichung mit

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

gegeben. (Vgl. Bosch, Wahrscheinlichkeitsrechnung, S. 128 f)

4.4. Kombinatorik als Voraussetzung zur Herleitung von theoretischen Verteilungen

Zunächst müssen wir den Begriff des Ereignisraums einführen. Ein **Ereignisraum** ist die Menge aller theoretisch möglichen Ergebnisse (Elementarereignisse) eines Zufallsexperiments, zum Beispiel die Menge aller Möglichkeiten, mit 2 Würfeln eine Zahlenkombination zu werfen.

Solche Ereignisräume werden als **Permutationen** oder **Kombinationen** beschrieben.

Es soll im folgenden zunächst eine zusammenfassende Darstellung der Definitionen und ein Überblick über die Formeln gegeben werden. Danach werden die einzelnen Fälle noch einmal anhand von Beispielen erläutert.

(1) PERMUTATIONEN

(1.1) **Ohne Wiederholung**: Menge der möglichen Veränderungen der Reihenfolge aller Elemente einer Menge von **verschiedenen Elementen**.

Formel: Permutationsmenge = $n!$

(1.2) **Mit Wiederholung**: Menge aller möglichen Veränderungen der Reihenfolge aller Elemente einer Menge von **teilweise identischen Elementen** in den Teilmengen n_1, n_2, \dots, n_k

Formel: Permutationsmenge_{id} =
$$\frac{n!}{n_1! n_2! \dots n_k!}$$

(2) KOMBINATIONEN

(2.1) Ohne Zurücklegen

(2.1.1) ohne Berücksichtigung der Anordnung:

Mengen von k verschiedenen Elementen aus der Menge n **verschiedener** Elemente
ohne Unterscheidung der Anordnung (Kombinationen k-ter Ordnung)

$$\text{Komb}_{\text{oWoA}} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

(Den letztgenannten Ausdruck liest man „n über k“. Hier fehlt kein Bruchstrich, die Zahlen stehen in den Klammern übereinander. Es handelt sich hier um den sog. Binomialkoeffizienten).

(2.1.2) mit Berücksichtigung der Anordnung:

Mengen von k verschiedenen
Elementen aus der Menge n **verschiedener** Elemente
mit Unterscheidung der Anordnung (wird auch
VARIATION k-ter Ordnung ohne Wiederholung
genannt)

$$\text{Formel} \quad \text{Komb}_{\text{oWmA}} = \frac{n!}{(n-k)!}$$

(2.2) mit Zurücklegen

(2.2.1) ohne Berücksichtigung der Anordnung:

Mengen von k Elementen aus n
verschiedenen Elementen, die unbeschränkt oft
wiederholt auftauchen können, ohne
Berücksichtigung der Anordnung

$$\text{Formel} \quad \text{Komb}_{\text{mWoA}} = \binom{n+k-1}{k}$$

(2.2.2) mit Berücksichtigung der Anordnung

Mengen von k Elementen aus n verschiedenen Elementen, die unbeschränkt oft wiederholt - mit Berücksichtigung der Anordnung - auftauchen können (=VARIATION MIT WIEDERHOLUNG).

$$\text{Komb}_{\text{mWmA}} = n^k$$

Zur Systematik der verschiedenen Fälle der Kombinatorik können wir uns merken, daß wir zur Auffindung der relevanten Formel folgende Fragen stellen müssen:

- Geht es um Teilmengen aus einer größeren Menge (**Kombinationen**) oder um die größere Menge allein sowie um die Reihenfolge ihrer Elemente (**Permutationen**) ?
- Enthält die größere Menge n nur verschiedene Elemente ("ohne Wiederholung/ohne Zurücklegen") oder auch teilweise identische Elemente ("mit Wiederholung, mit Zurücklegen") ?
- Kommt es bei den Teilmengen auch auf die Anordnung an ("mit Berücksichtigung der Reihenfolge") oder nicht ("ohne Berücksichtigung der Reihenfolge") ?

Nun zu konkreten Beispielen:

4.4.1. Permutationen

Permutationen: Veränderungen der Anordnung einer Menge von n verschiedenen Elementen einer Menge:

$$\text{Permutationsmenge} = n!$$

Beispiel: In einer Urne sind drei Kugeln, eine grüne, eine rote, eine blaue. In wieviel verschiedenen Reihenfolgen können sie gezogen werden, wenn man nach dem Ziehen keine zurücklegt ?

Lösung: GRB, GBR, RGB, RBG, BGR, BRG

Das sind 6 Möglichkeiten, die sich aus $3 \cdot 2 \cdot 1$ errechnen lassen, also aus $3! = 6$

Sind in der Urne nun 3 gelbe Kugeln, 3 rote Kugeln und zwei blaue Kugeln, dann wäre - ohne Zurücklegen - die

$$\text{Permutationsmenge}_{\text{id}} = \frac{n!}{n_1! n_2! \wedge n_k!}$$

Die Menge der möglichen Reihenfolgen ließe sich berechnen aus

$$8!/(3!*3!*2!) = 560.$$

4.4.2 Kombinationen

4.4.2.1. Kombinationen ohne Zurücklegen mit Unterscheidung der Anordnung

Greift man aus einer Menge verschiedener Elemente einen Teil der Elemente (ohne Zurücklegen) heraus und fragt, wieviele verschiedene KOMBINATIONEN (im Unterschied zu den Permutationen) dabei entstehen können, die sich **auch durch die Anordnung** unterscheiden, dann ergibt sich, daß die fragliche Menge der Kombinationen nach der Formel

$$\text{Komb}_{\text{oWmA}} = \frac{n!}{(n-k)!}$$

ermittelt werden kann.

Beispiel: Ein Lotto-Unternehmer bringt ein neues Spiel auf den Markt, in dem aus 29 Zahlen 5 Zahlen auszuwählen sind, ohne Zurücklegen – und es soll bei der getippten Lösung auf die Reihenfolge (Anordnung) der Ziehung ankommen !

Wie groß ist die Chance, 5 richtige in der richtigen Reihenfolge zu bekommen ? Die Menge aller möglichen Reihenfolgen beträgt:

$$29! / 24! = 14.250.600$$

Folglich ist die Chance, 5 richtige in der richtigen Reihenfolge zu bekommen,

$$p(5) = 1 / 14.250.600.$$

4.4.2.2. Kombinationen ohne Zurücklegen ohne Unterscheidung der Anordnung

Unterscheidet man im letztgenannten Fall die Ergebnisse **nicht** danach, **in welcher Reihenfolge** (Anordnung) die Zahlen gezogen werden, dann lautet die Formel:

$$\text{Komb}_{\text{oWoA}} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Für unseren o.a. Fall würden nur noch

$$29! / (5! * 24!) = 118755$$

verschiedene Möglichkeiten existieren. Das heißt: die Menge der Möglichkeiten wird kleiner.

Hinweis:

Der Ausdruck $\binom{n}{k}$

entspricht den Binomial-Koeffizienten, auf die wir weiter unten noch zu sprechen kommen.

4.4.2.3 Kombinationen mit Zurücklegen unter Berücksichtigung der Anordnung

Wenn das Lottospiel nun so konstruiert ist, daß aus 29 nummerierten Kugeln 5 ausgewählt werden, wobei die gezogene Kugel immer zurückgelegt wird, und unterscheidet man die sich ergebenden Kombinationen auch nach der Anordnung, dann errechnet man die Menge der sich ergebenden Kombinationen aus

$$n^k = 29^5 = 20.511.149$$

4.4.2.4 Kombinationen mit Zurücklegen ohne Berücksichtigung der Anordnung

Berücksichtigen wir die Anordnung, also die Reihenfolge, in der die 5 aus den 29 Kugeln gezogen werden, **nicht**, dann reduziert sich die Menge der Kombinationen (bei Zurücklegen) erheblich:

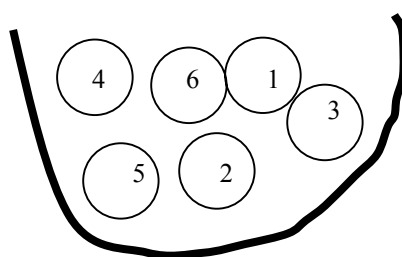
Die Formel lautet: $\binom{n+k-1}{k}$ (lies: n + k - 1 über k)

Setzt man die Werte aus obigem Beispiel ein, dann ergibt sich

$$Komb_{mWoA} = \binom{n+k-1}{k} = (29+5-1)! / (5! (29+5-1-5)!) = 33! / (5! 28!) = 273.336$$

4.4.2.5. Anwendung dieser Formeln auf ein Beispiel: Urne mit 6 Kugeln

Das folgende soll eine Urne mit 6 nummerierten Kugeln sein:



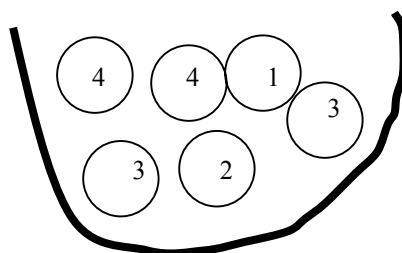
Anhand dieser Urne wollen wir uns **die 6 Fälle der Kombinatorik erarbeiten**.

Wenn wir eine Kugel blind, d.h. zufällig, entnehmen, können wir vorher nicht wissen, welche Kugel das sein wird. Entnehmen wir alle 6, dann kennen wir die Reihenfolge nicht, in der wir sie ziehen werden. Fragen wir uns, wieviele mögliche Reihenfolgen, die alle verschieden sind, wir beim Ziehen aller sechs Kugeln bekommen können, dann lautet die Antwort:

$$n! = 6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720.$$

Man nennt dieses Problem: **Permutation ohne Wiederholung (alle Kugeln sind verschieden)**.

Das folgende soll eine Urne mit 6 nummerierten Kugeln sein, aber die Nummern sind nicht alle verschieden, die 3 und die vier tauchen zweimal auf:



Fragen wir uns, wieviele mögliche unterscheidbare Reihenfolgen wir beim Ziehen aller sechs Kugeln bekommen können, dann lautet die Antwort:

$$\frac{n!}{n_1! \cdot n_2!} = \frac{720}{2! \cdot 2!} = 180$$

Man nennt diese Problem: **Permutation mit Wiederholung** (alle Kugeln sind nicht alle verschieden, sondern die 3 kommt zweimal vor (n1 ist im Nenner deshalb gleich 2), und die 4 kommt auch zweimal vor (n2 ist im Nenner deshalb auch gleich 2).

Stellen wir uns vor, wir ziehen aus einer Urne mit 6 von 1 bis 6 durch-numerierten Kugeln zweimal zufällig eine Kugel, die wir jeweils nach dem Ziehen **zurücklegen**, und werten wir die Ergebnisse **auch unter Berücksichtigung der Reihenfolge** der gezogenen Zahlen aus, dann können wir die in der folgenden Tabelle aufgeführten Ergebnisse bekommen. Die beiden Zahlen in den Zellen der Tabelle zeigen die gezogene Kombination nach der Durchführung der 1. und zweiten Ziehung.

In der grau unterlegten Vorspalte stehen die möglichen Ergebnisse der ersten Ziehung, in der grau unterlegten Kopfzeile stehen die Ergebnisse der zweiten Ziehung, in den Zellen der Tabelle stehen die möglichen Kombinationen aus erster und zweiter Ziehung – wie gesagt: bei Zurücklegen der Kugel und Berücksichtigung der Reihenfolge der gezogenen Kugeln.

Die Menge der Möglichkeiten beträgt:

$$n^k = 6^2 = 36$$

Erste Ziehung

Zweite Ziehung

	1	2	3	4	5	6
1	11	12	13	14	15	16
2	21	22	23	24	25	26
3	31	32	33	34	35	36
4	41	42	43	44	45	46
5	51	52	53	54	55	56
6	61	62	63	64	65	66

Dieses Problem heißt: **Menge der Kombinationen aus k Elementen, die aus einer Menge von n Elementen mit Zurücklegen gezogen werden, unter Berücksichtigung der Reihenfolge.**

Würden wir obige Ergebnisse **nicht nach der Reihenfolge** unterscheiden, die bei der Ziehung entsteht, dann wären z.B. die Kombinationen 21 und 12 identisch, so daß wir einige Zellen aus der Tabelle streichen könnten. Es bleiben nur 21 gefüllte Zellen übrig. Dieses Ergebnis, 21, errechnen wir mit der oben mitgeteilten Formel:

$$\binom{n+k-1}{k} = \binom{6+2-1}{2} = \frac{7!}{2!(7-2)!} = \binom{7}{2} =$$

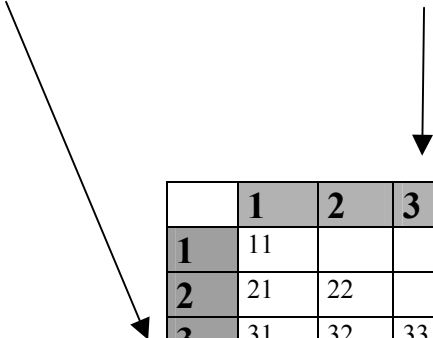
$$\frac{7 \cdot 6}{2} = 21$$

Zu beachten ist, daß die Zellen der Hauptdiagonale mit den Kombinationen 11, 22, 33 etc. nicht wegfallen, denn sie haben ja keine Doppelgänger, die gestrichen werden, wenn nicht mehr nach der Reihenfolge unterschieden wird.

Hier das Ergebnis nach Streichung der „Doppelgänger“

Erste
Ziehung

Zweite Ziehung



	1	2	3	4	5	6
1	11					
2	21	22				
3	31	32	33			
4	41	42	43	44		
5	51	52	53	54	55	
6	61	62	63	64	65	66

Dieses Problem heißt: **Menge der Kombinationen aus k Elementen, die aus einer Menge von n Elementen mit Zurücklegen gezogen werden, ohne Berücksichtigung der Reihenfolge.**

Bleibt man aber bei der **Berücksichtigung der Reihenfolge**, verzichtet aber beim Ziehen auf das **Zurücklegen der Kugeln**, fällt jetzt in der Tabelle die Hauptdiagonale weg, es bleiben aber alle übrigen Zelleninhalte erhalten. Der Wegfall der Hauptdiagonale geschieht deshalb, weil ja z.B. nach Ziehung der 1 die 1 nicht noch einmal vorkommen kann.

Die Formel zur Berechnung der verbleibenden Kombinationen lautet (s.o.)

$$\frac{n!}{(n-k)!} = 6! / 4! = 6 \cdot 5 = 30.$$

Die Ergebnisse sind in der folgenden Tabelle festgehalten:

Erste
Ziehung

Zweite Ziehung

	1	2	3	4	5	6
1		12	13	14	15	16
2	21		23	24	25	26
3	31	32		34	35	36
4	41	42	43		45	46
5	51	52	53	54		56
6	61	62	63	64	65	

Dieses Problem heißt: Menge der Kombinationen aus k Elementen, die aus einer Menge von n Elementen ohne Zurücklegen gezogen werden, mit Berücksichtigung der Reihenfolge.

Berücksichtigt man die Reihenfolge der gezogenen Kugeln nicht und verzichtet beim Ziehen auf das Zurücklegen, dann fällt in der folgenden Tabelle nicht nur die Hauptdiagonale weg, sondern auch alle „Doppelgängerfelder“ (z. B. für die Kombination „43“ der Doppelgänger „34“). Es verbleiben dann nur noch die Zelleninhalte der Zellen unterhalb der Hauptdiagonale (sofern man bei Zweierkombinationen immer die größere Zahl voranstellt):

Die Berechnung der Menge der verbleibenden Zellen (Menge der möglichen Kombinationen) geschieht nach der Formel für den Binomialkoeffizienten (siehe mathematischer Anhang im Dieckmann – Text):

$$\frac{n!}{k!(n-k)!} = \binom{n}{k} = \frac{6!}{2!(6-2)!} = \binom{6}{2} = 15$$

Erste
Ziehung

Zweite Ziehung

	1	2	3	4	5	6
1						
2	21					
3	31	32				
4	41	42	43			
5	51	52	53	54		
6	61	62	63	64	65	

Dieses Problem heißt: Menge der Kombinationen aus k Elementen, die aus einer Menge von n Elementen ohne Zurücklegen gezogen werden, ohne Berücksichtigung der Reihenfolge.

Alle unsere Beispiele setzten $k = 2$ und $n = 6$ voraus. Natürlich kann n auch jede andere ganze positive Zahl annehmen und k kann alle Zahlen von 0 bis n annehmen.

Allgemein gesprochen haben wir nun zwei Fälle der Permutation kennengelernt und vier Fälle der Kombination. Bei der **Permutation** wurden alle Kugeln entnommen und gefragt wurde nach der Menge der möglichen Reihenfolgen. Im ersten Fall waren alle Kugeln verschieden, im zweiten teilweise identisch.

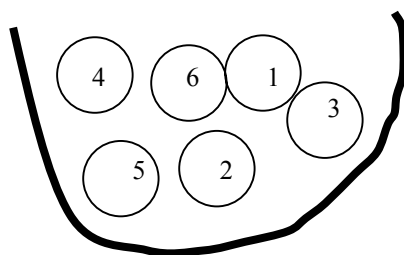
Die vier Fälle von **Kombinationen** unterschieden sich darin, daß die Kugeln zurückgelegt wurden, oder nicht, und daß es bei den entnommenen Kugeln auf die Reihenfolge ankam oder nicht.

Wenden wir uns nun noch einmal der Terminologie der Wahrscheinlichkeitsrechnung zu.

Dort spräche man, wenn man solche Dinge tut, wie Kugeln aus einer Urne zu ziehen, von einem Zufallsexperiment. Das Ziehen von k Kugeln aus n Kugeln hätte dann den Namen „Zusammengesetztes Ereignis“ und – weil zu diesem Ereignis meistens mehrere einzelne Kugeln gehören, bekommen diese den Namen „Elementar-Ereignis“. Die in der Wahrscheinlichkeitsrechnung interessierenden Ereignisse können zusammengesetzte oder Elementarereignisse sein.

Die Kombinatorik gibt uns die Antwort auf die Frage, wie groß der Ereignisraum für ein spezifisches Zufallsexperiment ist, das entweder Elementarereignisse oder zusammengesetzte Ereignisse betrifft.

Betrachten wir wiederum unsere Urne mit den 6 verschiedenen Kugeln:



Ziehen wir nur eine der Kugeln heraus, dann fragen wir uns, wie wahrscheinlich es ist, z. B. die 4 zu ziehen. Dabei benutzen wir den Begriff der Wahrscheinlichkeit. Dieser kann, als „klassischer“ Wahrscheinlichkeitsbegriff, wie folgt definiert werden:

$$P(A) = \frac{\text{Zahl der interessierenden Fälle}}{\text{Zahl aller gleichmöglichen Fälle}}$$

Diesem klassischen Begriff der Wahrscheinlichkeit steht ein statistischer Begriff der Wahrscheinlichkeit gegenüber, bei dem $P(A)$ eine Zahl ist, die bei langen Versuchsreihen immer besser angenähert wird: Wenn ich aus unserer Urne erst 10, dann 100, dann 1000 mal eine Kugel ziehe und dann jedesmal die Menge der gezogenen Sechsen durch die Menge aller Experimente teile, dann nähert sich das Ergebnis immer mehr dem Wert $1/6$ an. $1/6$ wäre hier ein Grenzwert. Daher kann man schreiben:

$$P(A) = \lim h_n/N$$

wobei h_n die Menge der geworfenen Sechsen wäre, N die Menge der Experimente.

In der **axiomatischen Fassung des Wahrscheinlichkeitsbegriffs** geht man nun von folgenden Festlegungen aus:

Satz	Formel	Beispiel	Kommentar
Die Wahrscheinlichkeit des Ereignisses A kann durch eine Zahl $P(A)$ ausgedrückt werden, die mindestens 0, maximal 1 beträgt.	$0 \leq P(A) \leq 1$	Die Wahrscheinlichkeit, aus einer Urne mit 6 verschieden nummerierten Kugeln eine bestimmte Zahl zu ziehen, ist $1/6$ liegt also zwischen 0 und 1	
Ist das Ereignis „sicher“, dann ist $P(A) = 1$. S ist das „sichere“ Ereignis.	$P(A) = 1 = P(S)$.	Die Wahrscheinlichkeit, eine beliebige von den vorhandenen Zahlen zu ziehen, ist gleich 1	
Additionssatz für einander ausschließende Ereignisse:	$P(A \cup B) = P(A) + P(B)$	Die Wahrscheinlichkeit, eine 3 oder eine 6 zu ziehen ist gleich $2/6$	Schließen sich zwei Ereignisse im Experiment gegenseitig aus, dann ist die Wahrscheinlichkeit, daß entweder das <u>eine oder</u> das andere auftritt, gleich der Summe ihrer einzelnen Wahrscheinlichkeiten.
Additionssatz für einander nicht ausschließende, beliebige Ereignisse:	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Man zieht aus 52 Karten eine Karte und fragt, wie wahrscheinlich es ist, daß diese ein As oder Karo ist. Die Wahrscheinlichkeit, ein As zu ziehen, ist $4/52$, die Wahrscheinlichkeit, Karo zu ziehen, ist $13/52$. Die Wahrscheinlichkeit, das Karo As zu ziehen ist $1/52$. Daher ergibt sich: $P(A \cup B) = 4/52 + 13/52 - 1/52$.	Schließen sich zwei Ereignisse im Experiment nicht gegenseitig aus, dann ist die Wahrscheinlichkeit, daß entweder das eine oder das andere auftritt, gleich der Summe ihrer einzelnen Wahrscheinlichkeiten, minus des Inhalts ihrer Schnittmenge
Die Wahrscheinlichkeit des unmöglichen Ereignisses U ist gleich Null	$P(\emptyset) = 1 - P(S) = 0$	Die Wahrscheinlichkeit, aus unserer Urne eine 7 zu ziehen, ist Null	
Die Wahrscheinlichkeit, daß von zwei komplementären Ereignissen wenigstens eines eintritt, ist gleich 1	$P(A) \cup P(\text{Nicht } A) = 1$	Die Wahrscheinlichkeit, entweder eine Zahl im Bereich von 1 bis 3 oder eine Zahl im Bereich von 4 bis 6 zu ziehen, ist gleich 1.	Komplementäre Ereignisse können nur zwei Ereignisse sein.
Die Wahrscheinlichkeit des gleichzeitigen Auftretens von zwei sich gegenseitig	$P(A \cap B) = 0$	Die Wahrscheinlichkeit, daß die gezogene Zahl gleichzeitig dem Bereich 1	Einander ausschließende Ereignisse können

ausschließenden Ereignissen ist gleich Null.		bis 3 und dem Bereich 4 bis 6 angehört, ist gleich Null.	mehr als 2 sein.
Multiplikationssatz für voneinander unabhängige Ereignisse:	$P(A \cap B) = P(A) * P(B)$	Wahrscheinlichkeit, zweimal die 1 mit Zurücklegen zu ziehen, ist $1/6 * 1/6 = 1/36$.	Durch das Zurücklegen ist das zweite Ereignis vom ersten unabhängig
Multiplikationssatz für voneinander abhängige Ereignisse	$P(A \cap B) = P(A) * P(B A)$	Die Wahrscheinlichkeit für das gleichzeitige Auftreten zweier Ereignisse A und B ist gleich dem Produkt der Wahrscheinlichkeit des ersten Ereignisses A mit der bedingten Wahrscheinlichkeit des Ereignisses B, die unter der Voraussetzung berechnet wurde, daß das Ereignis A schon eingetreten war. Beispiel: Die Wahrscheinlichkeit, zweimal die 1 ohne Zurücklegen zu ziehen ist gleich $1/6 * 0 = 0$	Durch das „Entnehmen ohne Zurücklegen“ verändern sich die Wahrscheinlichkeiten für die zweite Ziehung

Exkurs zur Berechnung der Summe von Zahlenfolgen und der Häufigkeit von Würfelsummen; Würfelbeispiele zur Kombinatorik

Die Menge der Zellen unter- oder oberhalb der Hauptdiagonale ist auch durch die Formel $(n * (n-1)) / 2 = (6 * 5) / 2 = 15$ zu errechnen. Das Ergebnis ist gleichzeitig die Summe der Zahlen von 1 bis n-1, also für unser Beispiel : $1 + 2 + 3 + 4 + 5 = 15$.

Die Summe der Zahlen von 1 bis n errechnet sich aus einer ähnlich einfachen Formel, nämlich

$$(n * (n+1)) / 2 = 1 + 2 + 3 + 4 + 5 + 6 = 21.$$

Diese Formeln sind von großem Nutzen, wenn wir die Summe großer Zahlenbereiche bilden wollen, wie z. B. die Summe aller Zahlen von 1 bis 10000:

$$(10000 * (10000 + 1)) / 2 = 50\,000\,5000.$$

Wollen wir die Summe der Zahlen von 2001 bis 5000 bestimmen, dann rechnen wir nur die Summe der Zahlen von 1 bis 5000 aus und subtrahieren die Summe der Zahlen von 1 bis 2000:

$$(5000 * (5001) / 2) - (2000 * (2001) / 2) = 10\,501\,500.$$

Das folgende ist ein Rechenverfahren zur Ermittlung der **Häufigkeit** von Würfelsummen mit Würfeln beliebiger Flächenzahl. Wir beginnen mit Würfeln, die nur zwei Flächen haben, auf denen die Zahlen 1 und 2 aufgetragen sind. Diese Würfel, Münzen vergleichbar, erzeugen binomial-verteilte **Häufigkeiten** ihrer Summe, wenn angenommen wird, daß zuerst mit einem Würfel, dann mit zwei Würfeln, dann mit drei Würfeln, dann mit vier Würfeln etc. geworfen wird. Das Rechenverfahren ist folgendes: Wir tragen die Häufigkeiten der möglichen Würfelsummen senkrecht in die folgende Tabelle ein. Wenn beim Werfen mit einem Würfel die Zahlen 1 und 2 jeweils nur einmal vorkommen können, sind das die Häufigkeiten 1 und 1. Diese sind links oben in der Tabelle senkrecht untereinander eingetragen. Nun tragen wir sie in die nächste Spalte noch einmal ein, aber um eine Zeile nach unten versetzt. Jetzt addieren wir zeilenweise diese versetzt eingetragenen Häufigkeiten und erhalten eine neue Spalte mit den Häufigkeiten für die Würfelsummen mit zwei Würfeln: 1 2 1. (Die Würfelsummen selber tauchen in der Tabelle nicht auf, es wären hier die Summen 2, 3 und 4). Die erhaltenen Häufigkeiten tragen wir wieder – um eine Zeile nach unten versetzt – in die nächste Spalte ein und addieren. Wir erhalten die Häufigkeiten für die Würfelsummen von drei Würfeln: 1 3 3 1 (für die Summen 3 4 5 6). Bei einem Würfel mit zwei Seiten erhalten wir mit diesem Verfahren die Binomialkoeffizienten, wie auch in der Kopfzeile der folgenden Tabelle angegeben wurde.

[illegible]

Was ist aber zu tun, wenn wir die Flächenzahl des Würfels variieren ? Die Anweisung lautet dann: Die Häufigkeiten werden nicht nur zweimal, um eine Zeile versetzt in die nächste Spalte geschrieben, sondern entsprechend der Flächenzahl dreimal: Beispiel für einen Würfel mit drei Flächen:

[illegible]

Exkurs mit Würfelbeispielen zur Kombinatorik

Man habe mit dem obigen Verfahren ausgerechnet, wie häufig mit 6 Würfeln folgende Würfelsummen geworfen werden können und noch einiges mehr, was im folgenden abgedruckt ist:

Würfelsumme mit 6 Würfeln

Würfelsumme	Häufigkeit	Gültige Prozente	Kumulierte Prozente
6,00	1	,0	,0
7,00	6	,0	,0
8,00	21	,0	,1
9,00	56	,1	,2
10,00	126	,3	,5
11,00	252	,5	1,0
12,00	456	1,0	2,0
13,00	756	1,6	3,6
14,00	1161	2,5	6,1
15,00	1666	3,6	9,6
16,00	2247	4,8	14,5
17,00	2856	6,1	20,6
18,00	3431	7,4	27,9
19,00	3906	8,4	36,3
20,00	4221	9,0	45,4
21,00	4332	9,3	54,6
22,00	4221	9,0	63,7
23,00	3906	8,4	72,1
24,00	3431	7,4	79,4
25,00	2856	6,1	85,5
26,00	2247	4,8	90,4
27,00	1666	3,6	93,9
28,00	1161	2,5	96,4
29,00	756	1,6	98,0
30,00	456	1,0	99,0
31,00	252	,5	99,5
32,00	126	,3	99,8
33,00	56	,1	99,9
34,00	21	,0	100,0
35,00	6	,0	100,0
36,00	1	,0	100,0
Gesamt	46656	100,0	

Würfelsumme bei sechs Würfeln - Kennwerte

N	Gültig	46656
Mittelwert	21,0000	
Standardfehler des Mittelwertes	1,937E-02	
Median	21,0000	
Modus	21,00	
Standardabweichung	4,1833	
Varianz	17,5004	
Spannweite	30,00	
Summe	979776,00	
Perzentile	25	18,0939
	50	21,0000
	75	23,9061

Der Spieler möchte vor dem Hintergrund dieser Daten folgendes wissen:

Wie groß sind die Chancen für

- 6 Einsen ? (1 / 46656)
- 6 Sechsen ? (1 / 46656)
- die Würfelsumme bis einschließlich 17 ? (20,6 %)
- die Würfelsumme größer als 17 ? (79,4%)
- zwei Dreien und vier Vieren ? Es könnten 6 Würfel 15 mal ($n! / n_1! * n_2!$) die Kombination von zwei Dreien und vier Vieren haben. Die Wahrscheinlichkeit ist demnach 15/46656.
- zweimal drei gleiche Zahlen ? Es gibt 15 Kombinationen zweier nicht identischer Zahlen aus 6 Zahlen. Diese können bei 6 Zahlen jeweils in 20 verschiedenen Reihenfolgen auftreten ($n! / n_1! * n_2!$). 15 mal 20 = 300. Lösung daher: 300/46656.
- einen Pasch ? (= 6 identische Zahlen) ? (6/46656)
- eine Kombination aus den Zahlen 1 bis 6 (1, 2, 3, 4, 5, 6) ? $n!/n^k = 6!/46656 = (720/46656)$
- Wie müßte man ein Würfelspiel mit 6 Würfeln, die gleichzeitig geworfen werden, konstruieren, in dem ein Spieler und sein potentieller Gegner gleiche Chancen haben, wenn sie sich den Bereich der Würfelsummen in Bereiche aufteilen, in denen der eine gewinnt und der andere verliert bzw. keiner gewinnt ? Wieviele Lösungen gibt es für dieses Problem ?

Wie man der obigen Tabelle entnehmen kann, wäre eine faire Lösung folgende: Ein Spieler gewinnt immer dann, wenn die Würfelsumme kleiner als 21 ist, der andere immer dann, wenn die Würfelsumme größer als 21 ist, und wenn die 21 kommt, gewinnt keiner. Man könnte aber auch noch weitere Lösungen anbieten: Zum Beispiel: Einer gewinnt immer, wenn die Würfelsumme kleiner als 18 ist, der andere, wenn die Würfelsumme größer als 24 ist und wenn die Würfelsumme im Restintervall liegt (18 bis 24) gewinnt keiner. Die Menge solcher Lösungen beträgt hier 15.

Exkurs Ende.

4.5. Die Binomialverteilung

Nehmen wir folgenden Fall an: Es gibt Neuwahlen in Berlin. Wir beobachten die Wahlbeteiligung von Männern und Frauen. Wir wollen wissen, wie wahrscheinlich es ist, daß unter 10 Wählern, die das Wahllokal betreten,

0 Frauen sind
 1 Frau ist
 2 Frauen sind
 3 " "
 4 " "
 5 " "
 6 " "
 7 " "
 8 " "
 9 " "
 10 " ".....

Wir nehmen an, daß die Menge der Wahlberechtigten **sehr groß ist**, und daß sich die Geschlechter gleich verteilen. Wir bezeichnen die Wahrscheinlichkeit, daß eine **Frau** das Wahllokal betritt, mit p und nehmen an, diese sei konstant $p = 0,5$; wir bezeichnen die Wahrscheinlichkeit, daß ein Mann das Wahllokal betritt, mit $q (= 1-p)$ und nehmen an, diese sei konstant $1-p = q = 0,5$. Außerdem nehmen wir an, daß die Wahrscheinlichkeit, daß

das Geschlecht desjenigen, der jeweils als nächster kommt, unabhängig vom Geschlecht des Vorgängers ist (d.h. die Ehepaare kommen nicht zusammen).

Unter diesen Bedingungen wäre z.B. die Wahrscheinlichkeit, zuerst 3 Frauen und danach 7 Männer bei 10 Leuten, die das Lokal betreten, zu bekommen, gleich

$$\text{pppqqqqqq} = p^3 q^7 = 0,5^3 * 0,5^7 = 0,000976563$$

Da nun diese Wahrscheinlichkeit auch gegeben wäre, wenn eine andere Reihenfolge, zum Beispiel diese

$$\text{pqqqpqqqpq}$$

vorhanden wäre, müssen wir $p^3 q^7$ noch mit der Anzahl multiplizieren, in der es Kombinationen gibt, die drei Frauen und 7 Männer enthalten.

Hierzu stellen wir folgende Überlegung an: Wir numerieren die Fälle, in denen ein Mensch das Wahllokal betritt, durch (bei 10 Menschen von 1 bis 10). Dann hätten wir die Zahlen 1 bis 10. Aus diesen greifen wir 3 Zahlen heraus und sagen, daß diesen Zahlen der Fall entspricht, daß eine Frau das Wahllokal betreten hat. Um die Menge der Kombinationen von 3 aus 10 verschiedenen Zahlen (ohne daß es auf die Reihenfolge ankommt) zu ermitteln, benutzen wir den oben dargestellten Binomialkoeffizienten.

Bei 10 Personen und 3 Frauen ergibt unser Binomialkoeffizient $10!/(3!7!) = 120$. Multiplizieren wir diesen mit $p^k q^{n-k}$,

dann ergibt sich $P_{n,k} = \binom{n}{k} p^k q^{n-k}$; hierbei ist $P_{n,k}$ die gesuchte Wahrscheinlichkeit,

daß unter n Menschen k Frauen sind.

Es ist zu beachten, daß das große P und das kleine p unterschiedliche Bedeutung haben! Das kleine p hatten wir ja für die Wahrscheinlichkeit „reserviert“, daß eine Frau das Wahllokal betritt.

Wir kommen zu folgendem Ergebnis, wenn wir das Beispiel für alle möglichen k (von 0 bis 10) durchrechnen:

Die Wahrscheinlichkeit, daß bei 10 Wählern, die das Wahllokal betreten,

$$\text{keine Frau ist, ist } \binom{10}{0} p^0 q^{10} = 1 * 0,000976563 = 0,000976563$$

$$1 \text{ Frau ist, ist } \binom{10}{1} p^1 q^9 = 10 * 0,000976563 = 0,00976563$$

$$2 \text{ Frauen sind, ist } \binom{10}{2} p^2 q^8 = 45 * 0,000976563 = 0,043945313$$

$$3 \text{ Frauen sind, ist } \binom{10}{3} p^3 q^7 = 120 * 0,000976563 = 0,1171875$$

$$4 \text{ Frauen sind, ist } \binom{10}{4} p^4 q^6 = 210 * 0,000976563 = 0,205078125$$

$$5 \text{ Frauen sind, ist } \binom{10}{5} p^5 q^5 = 252 * 0,000976563 = 0,24609375$$

6 Frauen sind, ist $\binom{10}{6} p^6 q^4 = 210 * 0,000976563 = 0,205078125$

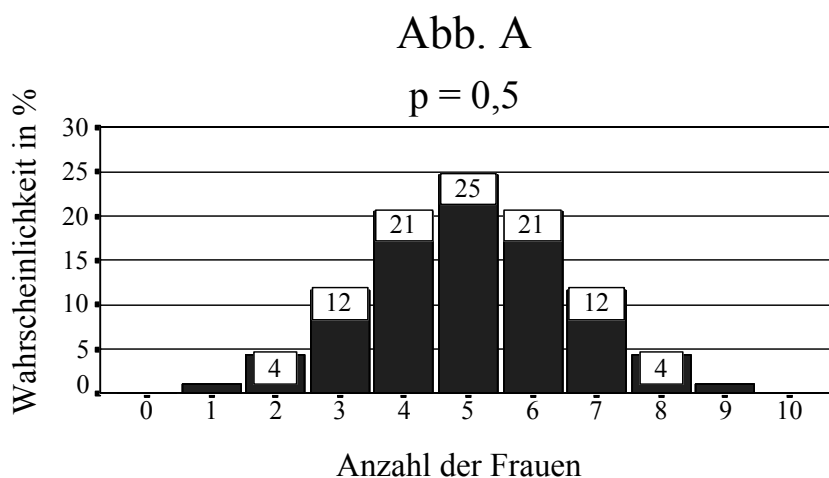
7 Frauen sind, ist $\binom{10}{7} p^7 q^3 = 120 * 0,000976563 = 0,1171875$

8 Frauen sind, ist $\binom{10}{8} p^8 q^2 = 45 * 0,000976563 = 0,043945313$

9 Frauen sind, ist $\binom{10}{9} p^9 q^1 = 10 * 0,000976563 = 0,00976563$

10 Frauen sind, ist $\binom{10}{10} p^{10} q^0 = 1 * 0,000976563 = 0,000976563$

Stellen wir diese Wahrscheinlichkeiten graphisch dar: Wir bemerken, daß sich eine symmetrische Verteilung ergibt, die ihr Maximum in der Mitte hat: Abbildung A (die in den Stäben in Prozent ausgedruckten Wahrscheinlichkeiten sind gerundet):



In den folgenden Abbildungen B, und C sind unterschiedliche Binomialverteilungen wiedergegeben, die sich nur durch Veränderung von p (und damit auch von q, da $q = 1-p$) ergeben. Sie werden berechnet, indem man für ein gegebenes p nacheinander - wie oben geschehen - den Wert k in der Formel von 0 bis N anwachsen läßt.

In den Graphiken ist unten die dem jeweiligen k entsprechende Wahrscheinlichkeit p als gerundeter Prozentsatz ausgedruckt.

Uns fällt auf, daß diese Verteilungen asymmetrisch sind. Diese Asymmetrie folgt aus der Tatsache, daß p und q bei diesen Verteilungen nicht gleich 0,5 sind, sondern gleich 0,1 bzw. 0,9.

Abb. B

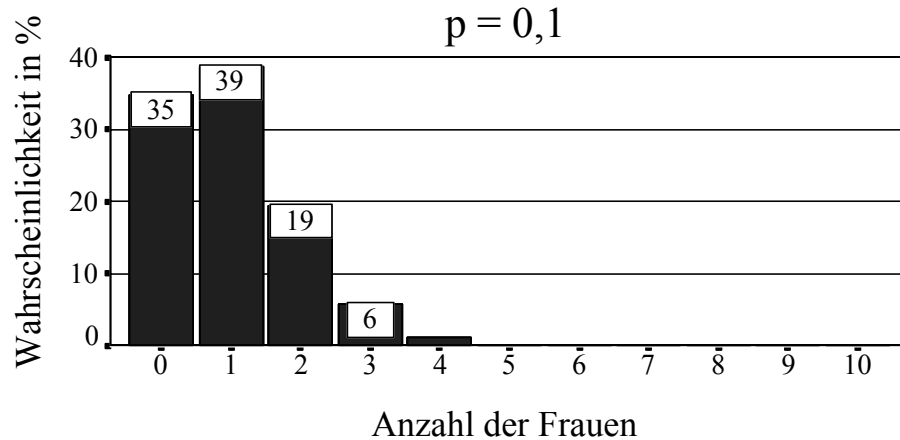
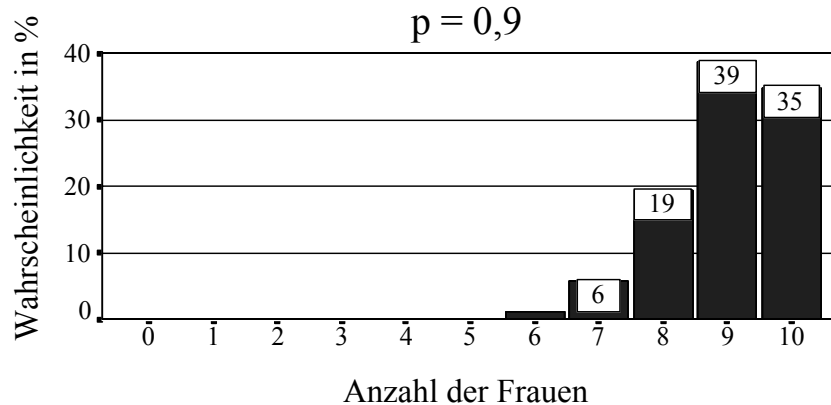
 $p = 0,1$ 

Abb. C

 $p = 0,9$ 

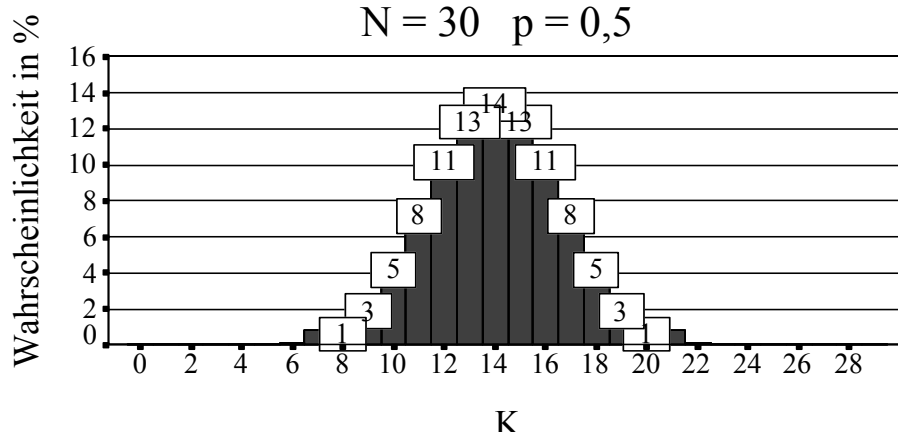
Es fällt auf, daß für $p = 0,1$ und $p = 0,9$ die Binomialverteilungen ihr Maximum bei $k = 1$ bzw. bei $k = 9$ haben, d.h. diese Verteilungen sind spiegelbildliche Abbilder voneinander.

Für die Abbildungen A bis C ist $N = 10$. Würde $N = 30$ sein, und wäre $p = 0,5$, dann sähe die entsprechende Binomialverteilung so aus (Abb. D):

Abb. D.

Binomialverteilung

$$N = 30 \quad p = 0,5$$



Rechnen wir nun - versuchsweise - die Wahrscheinlichkeiten für die einzelnen k in Häufigkeiten um und tun wir so, als wäre k ein Meßwert in einer empirischen Verteilung, dann können wir für die Binomialverteilung einen Mittelwert berechnen und auch eine Standardabweichung. Hierzu multiplizieren wir die Wahrscheinlichkeiten mit einem beliebigen, aber hinreichend großen n und erhalten dann für die einzelnen Ausprägungen von k 'empirische' Häufigkeiten. Für diese werden dann \bar{x} und s berechnet.

Berechnen wir hierfür den Mittelwert \bar{x} , dann erhalten wir $\bar{x} = 15$, die Standardabweichung beträgt $s = 2,7386$.

Für die Binomialverteilung können wir uns diese Berechnung insofern vereinfachen, als es für sie besondere, einfache Formeln zur Ermittlung von \bar{x} und s gibt.

Der Mittelwert von Binomialverteilungen errechnet sich aus

$$\mu = p * n.$$

Die Standardabweichung errechnet sich aus

$$S = \sqrt{p * n * q}$$

Für die oben aufgeführte Binomialverteilung errechnen wir:

$$p = 0,5; n = 30; p * n = 15$$

$$s = \sqrt{p * q * n} = \sqrt{0,5 * 0,5 * 30} = 2,7386$$

Die Binomialverteilung ist allerdings nur auf unser oben genanntes Problem, "Menge der Frauen unter 10 Wählern" anwendbar, wenn sich die Wahrscheinlichkeiten, daß Männer oder Frauen das Wahllokal betreten, nicht im Verlaufe des Experiments dadurch verändern, daß schon eine bestimmte Anzahl von Personen gewählt hat.

Beim Experiment mit einer Los-Trommel würde man also nach jedem Ziehen das gezogene Los **zurücklegen** müssen, wenn man die Binomialverteilung anwenden will.

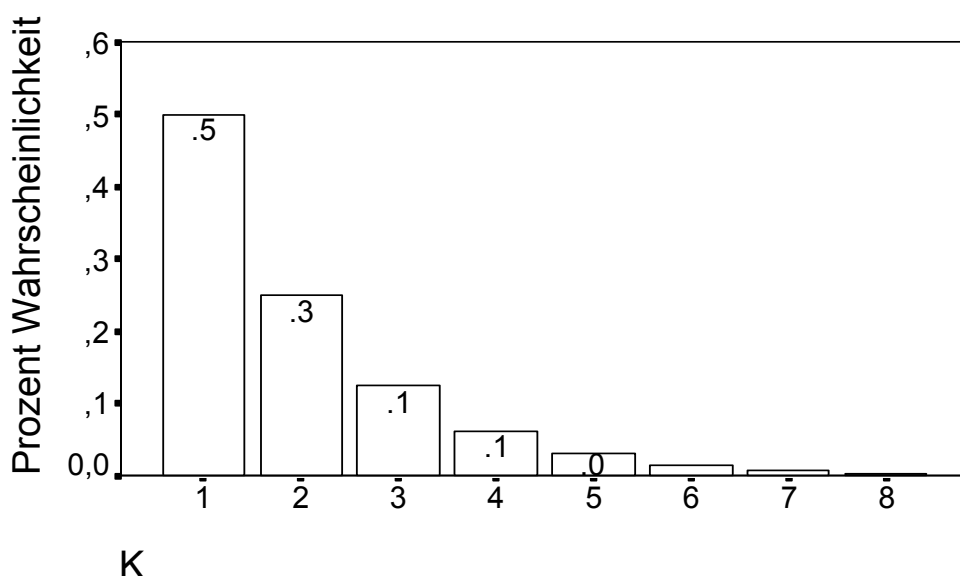
4.6. Die geometrische Verteilung

Fragen wir etwas einfacher in unserem Beispiel danach, wie groß die Wahrscheinlichkeit ist, daß der erste Wählende eine Frau ist; der erste und der zweite; der erste, der zweite und der dritte; etc., dann ist ein Anwendungsfall für die geometrische Verteilung gegeben. Wir setzen wieder voraus, daß 50% der Wahlberechtigten Frauen sind. Wir hatten oben schon die geometrische Reihe kennengelernt als Reihe von fortfolgenden Multiplikationen mit einem konstanten Faktor. So etwas geschieht auch hier. Die Wahrscheinlichkeit, daß der 1. Wählende eine Frau ist, der erste und der zweite Wählende, der erste, der zweite und der dritte Wählende etc. wird nach folgender Formel berechnet:

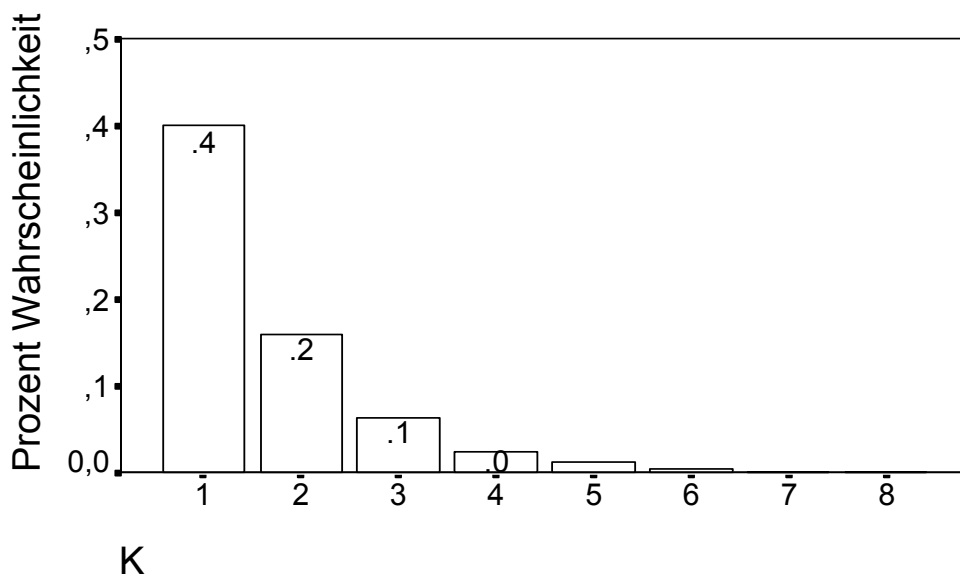
$$P = p(1-p)^{k-1} \quad (k = 1, 2 \dots n)$$

Hierbei ist P die gesuchte Wahrscheinlichkeit, p die relative Häufigkeit von Frauen, k ist die laufende Nummer der Beobachtung.

Wenn die relative Häufigkeit von Frauen unter den Wahlberechtigten gleich 0,5 ist, errechnen wir folgende Wahrscheinlichkeiten für unser Beispiel:



Wäre die relative Häufigkeit der weiblichen Wahlberechtigten nur $p = 0,4$ dann ergäbe sich folgendes Bild:



Man kann auch für die geometrische Verteilung einen Mittelwert und eine Standardabweichung angeben:

$$\mu = \frac{1}{p}$$

$$\sigma = \sqrt{\frac{q}{p^2}}$$

Für $p = 0,4$ wären $\mu = 2,5$
 $\sigma = 1,936$

4.7. Die hypergeometrische Verteilung

Wir nehmen wiederum unser Beispiel zum Ausgangspunkt, das wir schon von der Binomialverteilung her kennen. Wählen insgesamt z.B. **nur 100 Personen**, also nicht viele, die natürlich nicht ins Wahllokal zurückkehren, und sind davon 50 Männer und 50 Frauen, dann muß auf die gleiche Problematik die **hypergeometrische Verteilung** angewendet werden: Werden die 100 Personen mit N bezeichnet, die 50 Frauen mit M , die Häufigkeit des Auftretens von Frauen in einer Gruppe von n Personen mit x und die Zahl dieser "Personen" mit n , dann ergibt sich nach der Formel

$$P = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$

P ist die Wahrscheinlichkeit, unter $n = 10$ Personen, die das Wahllokal verlassen x ($x = 0; 1; 2; \dots; 10$) Frauen zu beobachten.

Für die o.a. hypergeometrische Verteilung müßten wir daher in die Formel einsetzen:

$N = 100$

$M = 50$

$x = 0, 1, 2 \dots 10$

$n = 10$

$$\text{d.h. } f(x) = \frac{\binom{50}{x} \binom{100-50}{10-x}}{\binom{100}{10}}$$

Es ergeben sich folgende Wahrscheinlichkeiten für $x = 0, x = 1; \dots; x = 10$:

x = 0 Männer, Wahrscheinlichkeit:		0,000593420
1 Mann	"	0,007236825
2 Männer	"	0,037993333
3 Männer	"	0,113096432
4 Männer	"	0,211413217
5 Männer	"	0,259333546
6 Männer	"	0,211413217
7 Männer	"	0,113096432
8 Männer	"	0,037993333
9 Männer	"	0,007236825
10 Männer	"	0,000593420

Das Beispiel zeigt, daß Unterschiede zur Binomialverteilung in den Wahrscheinlichkeiten erst in der zweiten Stelle nach dem Komma auftreten. Die Binomialverteilung ist der hypergeometrischen Verteilung insofern oft nah verwandt, besonders, wenn N, M oder N-M groß sind und wenn n im Vergleich dazu klein ist.

Neu bei der hypergeometrischen Verteilung ist - gegenüber der Binomialverteilung - die Einführung der Gesamtmenge von Ereignissen, auf die sich die Verteilung jeweils bezieht.

Demzufolge berechnet sich bei der hypergeometrischen Verteilung der Mittelwert - anders als bei der Binomialverteilung - nach folgender Formel:

$$\mu = n * \frac{M}{N}$$

Die Standardabweichung ist auch anders als bei der Binomialverteilung zu berechnen, denn sie ist in der Regel kleiner:

$$S = \sqrt{\frac{nM(N-M)(N-n)}{N^2(N-1)}}$$

wobei sich M/N und p aus der Binomialverteilung entsprechen.

(Weiteres und Formelherleitung: siehe Bosch, Karl: Einführung in die elementare Wahrscheinlichkeitsrechnung, Vieweg, Braunschweig 1984⁴, S. 84 ff, sowie Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen, Vandenhoeck 1973, 4. Aufl, S. 121)

4.8 Die Poissonverteilung

Für einen Spezialfall der Binomialverteilung hat man noch eine weitere Verteilungsfunktion ermittelt:

Ist die Zahl der "Personen" n in unserem Wahlbeispiel sehr groß und die Wahrscheinlichkeit des einzelnen Ereignisses p sehr klein, dann ist es angeraten, statt der Binomialverteilung die sogenannte Poissonverteilung zu wählen. Diese ist für große n eine gute Annäherung der Binomialverteilung, aber weniger umständlich zu berechnen.

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}$$

Hierbei ist $\mu = np$

ein Ausdruck, den wir als Mittelwert der Binomialverteilung schon kennengelernt haben. Beispiel:

Wir wissen aus früheren Untersuchungen, daß nur wenige Frauen rechtsradikal wählen. Wir sind daran interessiert zu erfahren, wie wahrscheinlich es ist, unter 100 Frauen 0, 1, 2, ... 100 Frauen zu treffen, die rechtsradikal wählen. Die Wahrscheinlichkeit, daß eine Frau rechtsradikal wählt, sei z.B. 1,8%. Dann wäre hiernach

$$\mu = 1,8.$$

Rechnet man die o.a. Formel durch, dann ergibt sich, daß die Wahrscheinlichkeit, bei 100 Frauen

0	keine solche zu treffen	0,17	ist
1	solche zu treffen	0,30	"
2	" " "	0,27	"
3	" " "	0,16	"
4	" " "	0,07	"
5	" " "	0,03	"
6	" " "	0,01	"
....			
10	" " "	0,000016264	ist.

Die Poisson-Verteilung ist kein Ersatz für die Binomial-Verteilung in allen Fällen, in denen letztere schwer zu berechnen ist. Wenn nämlich große Werte für $x!$ auftauchen, kann es technisch schwierig sein, $x!$ zu berechnen.

4.9. Die Normalverteilung und die Prüfung einer Verteilung auf Normalität

Waren die Binomialverteilung, die geometrische, die hypergeometrische und die Poisson-Verteilung auf diskrete Merkmale bezogen, so ist die Normalverteilung auf stetige Merkmale bezogen. Wir können die Normalverteilung als Spezialfall der Binomialverteilung betrachten, wenn dort N bzw. K gegen unendlich gehen und $p = 1-q = 0,5$ ist.

Nach dem "lokalen Grenzwertsatz" von Moivre-Laplace gilt für jedes p , das zwischen 0 und 1 liegt, folgende Beziehung:

$$\binom{n}{k} p^k q^{n-k} = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} * [1 + R_n(k)]$$

für $k = 0, 1, 2, \dots, n$

Hierbei ist $R_n(k)$ ein sogenanntes "Restglied" in der Formel, das für wachsende k gegen 0 strebt, so daß für sehr große k das Restglied fortfällt.

Ersetzen wir jetzt (Übergang zur stetigen Variable!)

$(k-np)$ durch x und setzen wir $npq = 1$,

dann erhalten wir die Formel

$$z(x) = \frac{e^{-\left(\frac{x^2}{2}\right)}}{\sqrt{2\pi}}$$

Dies ist die Formel der sogenannten "**Standardnormalverteilung**".

$z(x)$ hat die Standardabweichung 1 und den Mittelwert $x = 0$.

Um Verwechslungen mit den Meßwerten x_i auszuschließen, bezeichnen wir künftig in der Standard-Normalverteilung die x_i als z -Werte (z_i).

Da die Verteilung stetig ist, können wir die Ordinaten der Kurve der Standardnormalverteilung nicht als **Wahrscheinlichkeiten** p interpretieren, sondern nur als **Wahrscheinlichkeitsdichte**.

Dementsprechend können konkrete Wahrscheinlichkeiten auch nur bestimmten Intervallen dieser Funktion zugeordnet werden, nicht aber einzelnen x_i . Dies bedeutet, daß - im Unterschied zu allen diskreten Wahrscheinlichkeitsverteilungen - die Aufsummierung aller auf einzelne z_i bezogenen Wahrscheinlichkeitsdichten nicht 1 ergibt (sondern unendlich), während die Addition der einzelnen Wahrscheinlichkeiten für die Binomialverteilung, die geometrische Verteilung, die hypergeometrische Verteilung ect. immer die Summe 1 ergab.

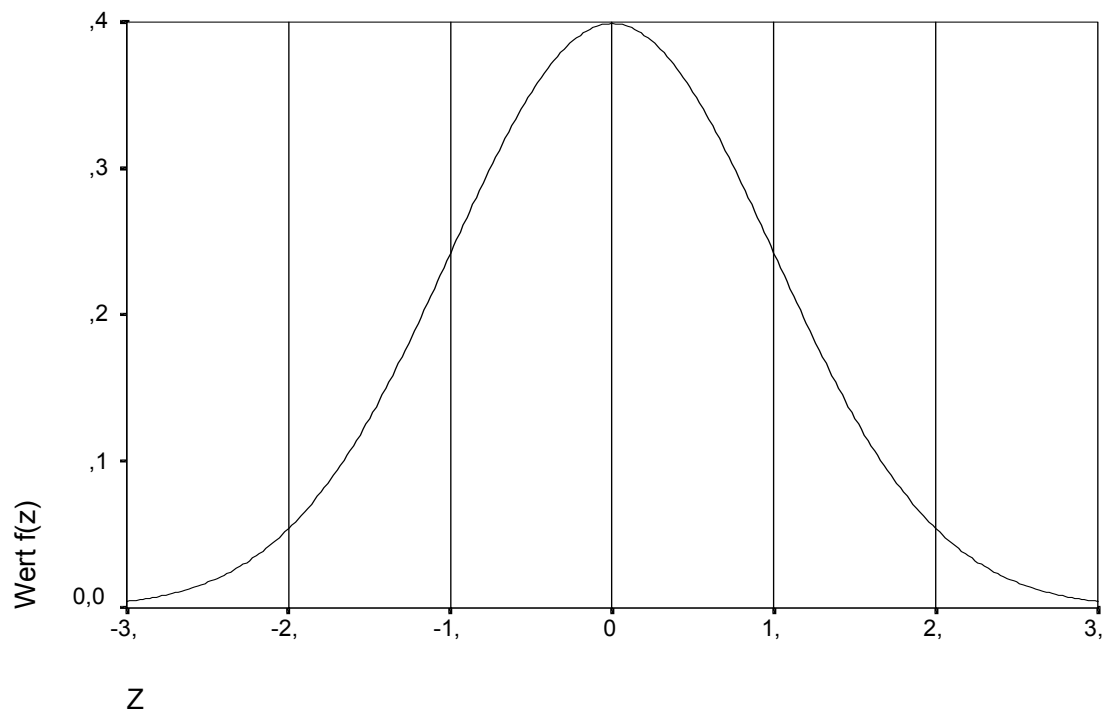
Die Summe 1 als Summe aller Einzelwahrscheinlichkeiten resultiert in allen Fällen genannten Verteilungen daraus, daß das sichere Ereignis die Wahrscheinlichkeit 1 hat. Die Wahrscheinlichkeit, daß ein mögliches Ereignis (für das jeweilige Experiment) im Wertebereich der angegebenen Verteilung auftaucht, ist immer gleich 1, weil dieses Ereignis sicher ist; z.B. beim Münzwurf kommt immer Kopf **oder** Zahl.

Wie ermitteln wir aber bei der Standard-Normalverteilung die Einzelwahrscheinlichkeiten, deren Summe 1 ergibt? Wir müssen dazu die **Flächen** (Integrale) berechnen, die bestimmten Intervallen auf der z -Achse entsprechen. Wer dies nicht mit dem (programmierbaren) Taschenrechner durchführen kann, ist praktisch darauf angewiesen, in Tabellen der Statistik-Lehrbücher diese Flächenanteile nachzuschlagen. Dabei ist zu beachten, daß solche Tabellen von Lehrbuch zu Lehrbuch anders aufgebaut sind, und man muß die Leseanweisung für solche Tabellen sorgfältig beachten.

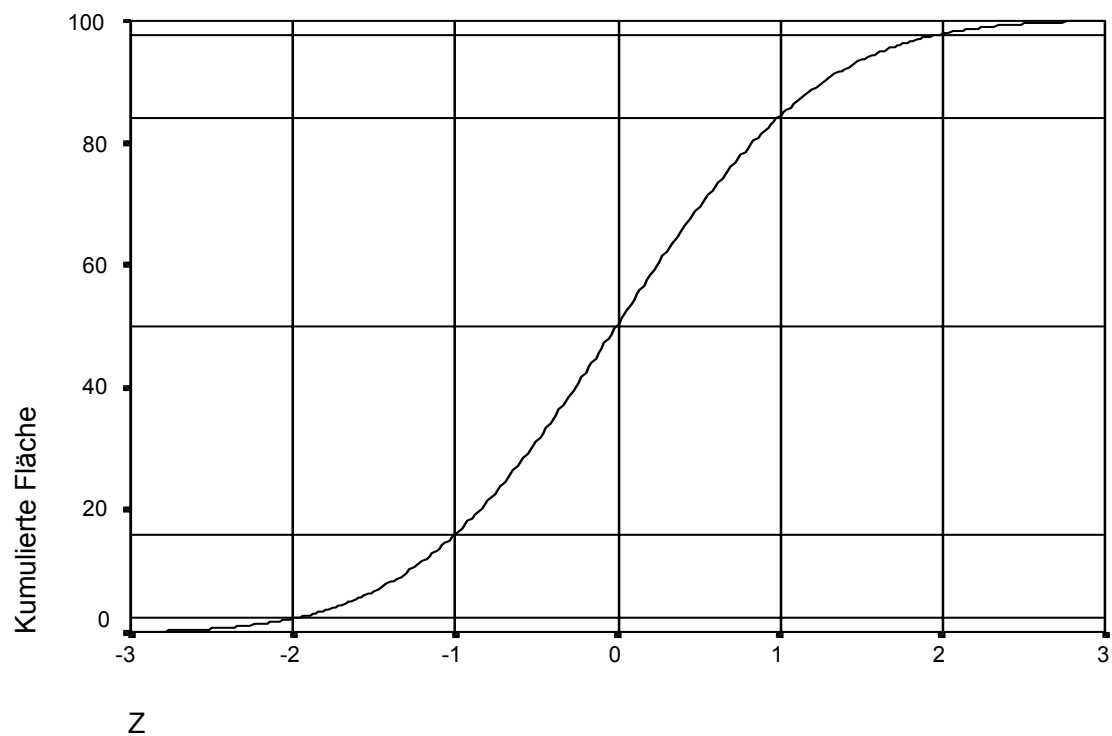
Untersuchen wir, welche Werte das Integral der Standard-Normalverteilung annimmt, wenn wir, bei minus unendlich beginnend, über null bis plus unendlich die Fläche zusammensetzen, die unter der Funktion der Wahrscheinlichkeitsdichte liegt:

Wir wissen im voraus, daß die Summe der Flächenstücke gleich 1 ist, weil ja die Standardnormalverteilung nur ein Grenzfall der Binomialverteilung ist, bei der - wie bei allen untersuchten Wahrscheinlichkeitsverteilungen die Summe der Einzelwahrscheinlichkeiten immer gleich 1 war. Hiervon ausgehend können wir sagen, daß bei die Fläche links (oder rechts) vom Mittelwert (0) gleich 0,5 sein muß. Ermitteln wir die Fläche im Intervall von -1 bis 0, dann ergibt sich der Wert 0,341. Folglich beträgt die Fläche links von -1 „0,5-0,341 = 0,159“. Da die Standardnormalverteilung symmetrisch zu 0 ist, kann man alle Überlegungen aus dem Bereich links von 0 analog in den Bereich rechts von 0 übertragen, durch Spiegelung an der 0-Achse.

Verlauf der Standardnormalverteilung:



Summiert man die Flächenstücke unter der Standardnormalverteilung von links nach rechts auf - wie bei der Kumulierung von Prozents - dann ergibt sich das Integral (die Summenfunktion) der Verteilung als von 0 bis 1 ansteigende Kurve. Die oben gezeichnete Grafik enthält die Standardnormalverteilung, unten finden wir die dazugehörige Summenfunktion. Die Ordinate wird hier in Prozent dargestellt:



Zu berücksichtigen ist, daß die beiden Kurven sich auf unterschiedliche Maßstäbe in der Senkrechten beziehen: auf die Wahrscheinlichkeitsdichte in der Standardnormalverteilung oben, auf das Integral (als prozentualer Anteil der Gesamtfläche ausgedrückt), unten.

Praktisch werden die Standardnormalverteilung und die ihr zugeordnete Integralfunktion zur Schätzung von Häufigkeiten in bestimmten Intervallen empirischer Verteilungen benutzt. Wenn ich zum Beispiel weiß, daß eine empirische Verteilung **das arithmetische Mittel 20** hat und die **Standardabweichung von 6**, dann kann ich leicht die Wahrscheinlichkeit dafür errechnen, daß ein Meßwert im Bereich von 14 bis 20 auftritt.

Wir ziehen hierzu von jedem empirischen Meßwert das empirische arithmetische Mittel ab und teilen die Differenz durch die empirische Standardabweichung. Dieses Verfahren nennt man „**z-Standardisierung**“.

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Für ein Intervall von 14 bis 20 heißt dies:

$$Z_1 = \frac{14 - 20}{6} = \frac{-6}{6} = -1$$

$$Z_2 = \frac{20 - 20}{6} = \frac{0}{6} = 0$$

Für das Intervall z_1 bis z_2 , also von minus 1 bis 0, errechnen wir jetzt die "Fläche unter der Standard-Normalverteilung" und erhalten, wie oben schon ermittelt, den Wert 0,341. Damit kann vermutet werden, daß für die empirische Verteilung mit $\bar{x} = 20$ und $S = 6$ im Intervall von 14 bis 20 34,1% der Meßwerte liegen.

Die Berechnung von z-Werten aus empirischen Meßwerten ist jedoch nur sinnvoll, wenn die empirische Verteilung annähernd normal verteilt ist. Ist dies nicht der Fall, dann ergeben sich ganz andere Wahrscheinlichkeitsverteilungen.

Wie kann man aber prüfen, ob eine Verteilung annähernd „normalverteilt“ ist ? Die **Prüfung auf Normalität** kann viele Wege gehen. Ein einfaches Verfahren ist die Berechnung von Maßen der "Schiefe" (englisch: Skewness) und des "Exzesses" (englisch: Kurtosis), worunter wir uns Abweichungen einer empirischen Verteilung von der Form der Normalverteilung vorzustellen haben. Solche Abweichungen können in der Richtung der Meßwertachse (Problem der Schiefe) vorliegen oder in Richtung der Häufigkeitsachse (Problem des Exzesses oder der Kurtosis, s.u.).

Das Schiefemaß lautet:

$$Sch = \frac{\sum_{i=1}^n z_i^3}{n}$$

Bei Vorliegen einer „Normalverteilung“ ist die Schiefe = 0. Ist die Verteilung „schief“, d.h. ist der Schwerpunkt nach links oder rechts verschoben, dann nimmt das Schiefemaß bei Schwerpunktverschiebung nach links größere, bei Schwerpunktverschiebung nach rechts kleinere Werte als 0 an.

Das „Exzess“-Maß lautet:

$$Ex = \frac{\sum_{i=1}^n z_i^4}{n}$$

Bei Vorliegen einer Normalverteilung nimmt es den Wert 3 an. Deshalb wird der Exzess auch gelegentlich als

$$Ex = \frac{\sum_{i=1}^n z_i^4}{n} - 3$$

definiert. Dann ergibt sich bei Vorliegen einer Normalverteilung der Wert 0. Unter Exzess wird die Verschiebung der Verteilung in der Senkrechten verstanden (im Gegensatz zur Schiefe, bei der es um waagerechte Abweichungen von der Normalverteilung geht). Liegen in der Mitte der Verteilung die Häufigkeiten oberhalb der Normalverteilung, am Rande aber unterhalb, dann wächst das Exzessmaß auf Werte oberhalb von Null, ist es umgekehrt, fällt das Exzessmaß auf Werte unter Null.

Es ist darauf hinzuweisen, daß SPSS10 bei Schiefe und Exzess mit anderen Formeln arbeitet. Die Schiefe und der Exzess sind in SPSS.10 wie folgt definiert:

$$Schiefe = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{s^3 (n-1)(n-2)}$$

$$Exzess = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3(n-1) \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}{(n-1)(n-2)(n-3)s^4}$$

Diese Formeln führen zu leichten Abweichungen der Ergebnisse gegenüber den oben genannten Formeln für Schiefe und Exzess. Im Übrigen wird der Exzess in SPSS.10 mit „Kurtosis“ bezeichnet.

Ob eine konkrete Abweichung der Schiefe und des Exzesses von ihren „Normwerten“ erheblich (signifikant) ist, kann mit Hilfe des unten besprochenen Chiquadrat-Tests oder mit Hilfe des Kolmogoroff-Smirnov-Tests ermittelt werden.

Eine weitere Möglichkeit, die Normalität einer Verteilung festzustellen, existiert mit Hilfe von Schiefe und Exzess wie folgt:

$$Chi^2 = \frac{Sch^2}{\frac{6}{n}} + \frac{(Ex-3)^2}{\frac{24}{n}}; df = 2$$

Der Wert Chi^2 ist chiquadratverteilt (s.u.).

5. Prüfverteilungen

Verschiedene, mit der Normalverteilung verwandte, spezielle Verteilungen werden in der Statistik - neben der Binomial- und der Normalverteilung - dazu benutzt, statistische Hypothesen zu überprüfen. (Was statistische Hypothesen sind, wird weiter unten erörtert). Hierzu gehören die Chi-Quadrat-Verteilung, die t-Verteilung und die F-Verteilung. In diesem Kapitel wird zunächst nur die Form der Verteilung eingeführt und ihre Berechnungsweise. Wie man mit ihr praktisch umgeht, wird nur kurz angedeutet, später werden die Anwendungsfälle genauer ausgeführt. Tabellen dieser Verteilungen findet man im Anhang, wobei ausführlichere Fassungen solcher Tabellen in der Lehrbuchliteratur für Statistik aufzusuchen sind.

5.1. Die Chi-Quadrat-Verteilung

Wenn wir mehrere voneinander unabhängige Zufallsgrößen (Variablen) haben, dann kann deren Summe - innerhalb bestimmter Grenzen - jeden beliebigen Wert annehmen. Das gilt auch für die Summe der quadrierten Werte, nur daß diese Summe immer positiv ist, d.h. kein negatives Vorzeichen haben kann.

Die Summe einer Menge von quadrierten Daten aus standardisierten Normalverteilungen

$z_1^2 + z_2^2 + z_3^2 + z_4^2 + \dots + z_n^2$ kann wie folgt ausgedrückt werden:

$$CHI^2 = \sum_{i=1}^n z_i^2$$

Der Menge der Summanden, n, entspricht in diesem Fall die Menge der **Freiheitsgrade**.

Kurzer Exkurs zum Begriff „Freiheitsgrad“, der uns noch oft begegnen wird: Die Schätzung von Parametern ist eng verbunden, mit der jeweils zur Verfügung stehenden Information. Die formale Anzahl von Informationen (z.B. die Meßwerte von N Befragten), die zur Schätzung eines Parameters herangezogen werden, ist der Ausgangspunkt für die Festlegung der zur Verfügung stehenden Freiheitsgraden. Im allgemeinen ist die Anzahl der Freiheitsgrade gleich der formalen Anzahl unabhängiger Einzelinformationen minus der Anzahl der in die Berechnung des jeweiligen Parameters eingehenden zusätzlichen Parameter. Wird z.B. die Varianz (σ^2) einer Verteilung mit N Werten geschätzt, dann ist die Anzahl der Freiheitsgrade N-1, da die Formel zur Berechnung von Varianz den Mittelwert μ als weiteren Parameter enthält.

Die folgende Tabelle enthält in den ersten 6 Spalten 6 *quadrierte* z-Verteilungen (Normalverteilungen mit dem arithmetischen Mittel 0 und der Standardabweichung 1), und zwar pro Verteilung 20 Werte. Diese unsortierten quadrierten Werte werden Zeile für Zeile addiert und ergeben in der letzten Spalte Ausprägungen der Variable **Chi²_{df6}**:

z_1^2	z_2^2	z_3^2	z_4^2	z_5^2	z_6^2	Chi ² _{df6}
,00	,00	,24	,58	1,60	,23	2,65
,77	,16	1,12	,17	3,75	2,38	8,36
2,30	,00	,94	,69	1,93	,22	6,08
,07	,17	1,24	,75	,49	,05	2,77
,00	,02	3,10	4,18	1,64	1,58	10,52
1,86	,10	,23	,23	,19	,27	2,87
,05	,68	,44	3,85	,06	,18	5,26
,10	,00	,01	2,96	1,02	,05	4,15
,48	,00	1,29	,14	1,12	,07	3,12
,96	,09	1,07	,50	,03	,28	2,95
,00	1,42	4,52	,32	,31	5,64	12,21
1,12	7,51	2,92	,15	2,11	,06	13,86
1,92	,63	,20	2,84	,47	,13	6,18
,08	1,97	,21	,01	,00	,06	2,33
,61	1,14	,01	,55	,71	,01	3,03
,10	,17	1,97	1,78	2,32	,76	7,09
1,30	,94	,76	1,07	,06	2,44	6,57
1,32	,20	,23	1,07	2,67	,03	5,51
1,84	5,74	,82	,05	,34	,08	8,86
1,49	,45	3,95	,07	,62	,16	6,76

Die Möglichkeiten des Verlaufs dieser Variable „Chi²_{df}“ sind theoretisch durch eine Verteilungsfunktion (Wahrscheinlichkeitsdichtefunktion) zu beschreiben, die folgende Formel hat

$$f[\text{Chi}^2] = \frac{e^{-\left(\frac{x^2}{2}\right)} * x^{\frac{df}{2}-1}}{2^{\frac{df}{2}} * \Gamma\left(\frac{df}{2}\right)}$$

In der Formel für Chi-Quadrat bedeuten

e = Eulersche Konstante = 2,71828 ...

Chi² = der konkrete Wert Chiquadrat = $\sum_{i=1}^n z_i^2$

df = Menge der sog. Freiheitsgrade der Funktion (df > 0).

Γ = Im Nenner befindet sich das Zeichen „ Γ “. Es bezeichnet die sogenannte Gammafunktion, die wir als stetige Erweiterung der (diskreten) Fakultät für nichtgradzahlige Werte auffassen können, wenn wir beachten: $\Gamma(n) = (n-1)!$ für alle natürlichen Zahlen. Eine ausführliche Erläuterung der Gamma-Funktion und auch eine entsprechende Tabelle liefert Erwin Kreyszig, Statistische Methoden und ihre Anwendungen, Göttingen 1965, S. 158 u. S. 391.

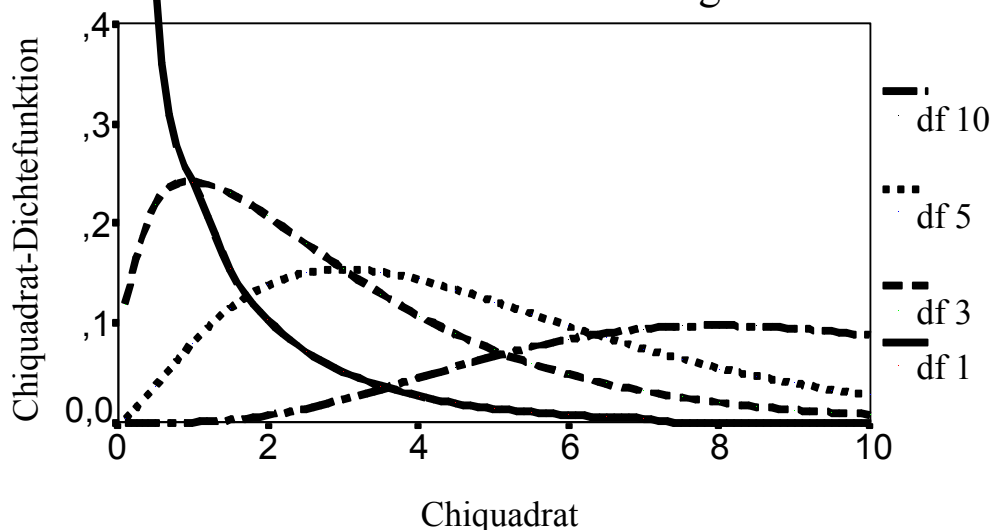
Wie man in der Formel erkennt, gibt es so viele Chiquadrat-Funktionen wie es Freiheitsgrade (df) gibt.

Für alle diese Funktionen gilt:

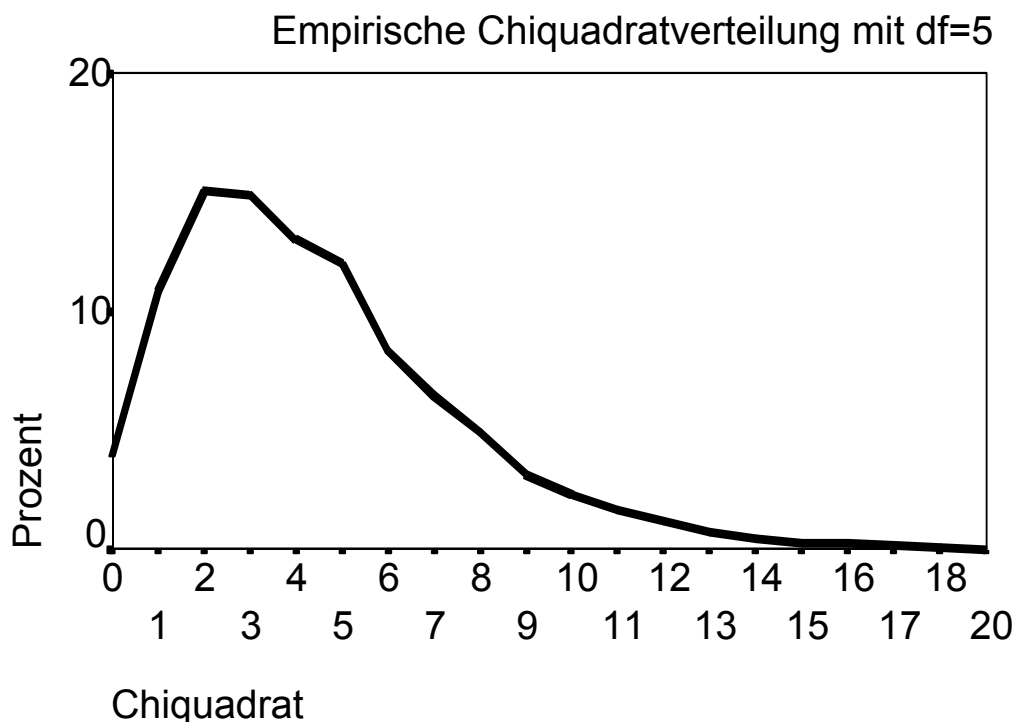
- Die Chi²-Verteilung hat ein arithmetisches Mittel von df und eine Varianz von 2df.
- Alle Chiquadrat-Funktionen können als Wahrscheinlichkeitsdichtefunktionen mit dem Gesamt-Integral gleich 1 interpretiert werden.

Die Funktion verläuft bei einem Freiheitsgrad monoton fallend, bei 2 Freiheitsgraden ebenfalls, bei 3 Freiheitsgraden steigt sie zunächst an und fällt dann stetig. Ab 10 Freiheitsgraden wird sie der Normalverteilung immer ähnlicher.

Chiquadrat-Dichtefunktionen für unterschiedliche Freiheitsgrade

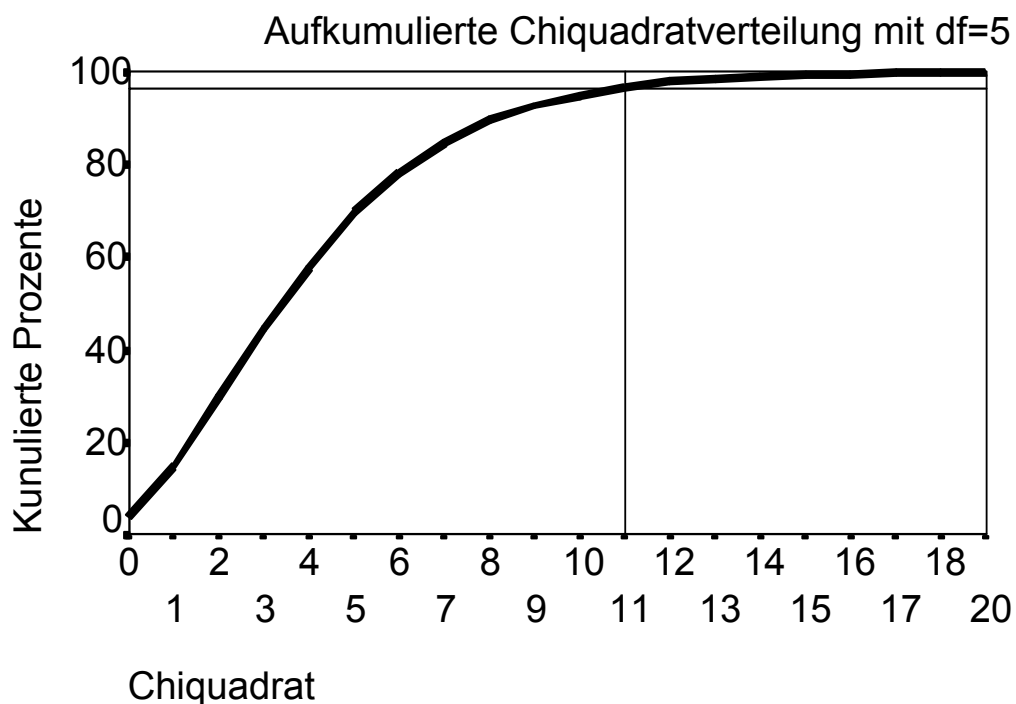


„Empirische“ Chiquadratfunktionen lassen sich durch das oben erwähnte Additionsverfahren erzeugen und haben in ihrer äußeren Gestalt große Ähnlichkeit mit den in der Graphik der Dichtefunktionen dargestellten Funktionsverläufen, wie das folgende Beispiel einer empirischen Chiquadratverteilungen mit 5 Freiheitsgraden zeigt:



Der Unterschied der empirischen Chiquadratverteilung zur theoretischen Chiquadratverteilung besteht hier in folgendem: Während die Ordinate der empirischen Verteilung mit Prozent (Werte von 1 bis 100) bezeichnet werden kann, wird die Ordinate der theoretischen Verteilung mit Wahrscheinlichkeitsdichte (Werte von 0 bis 1) bezeichnet. Die theoretische Verteilung, die nach der o.a. Formel berechnet wurde, ist glatt, die empirische Verteilung hat Unregelmäßigkeiten.

Praktisch wird die Chiquadrat-Verteilung dazu benützt, abzuschätzen, wie wahrscheinlich es ist, daß bestimmte chiquadrat-verteilte Rechenergebnisse in bestimmten Intervallen der Chiquadratverteilung als Resultat empirischer Untersuchungen auftreten. Hierbei bedient man sich der Integrale (Flächen unter) der Chiquadratdichtefunktion. Diese Integrale der Chiquadrat-Dichtefunktionen liegen in Tabellenform vor, so daß aus ihnen entnommen werden kann, in welchen Intervallen der Chiquadratverteilung Chiquadratwerte mit welcher Wahrscheinlichkeit auftreten können. So zeigt die folgende Graphik der aufkumulierten Flächen einer Chiquadrat-Verteilung ($df = 5$), daß oberhalb des Chiquadratwertes 11 noch etwa 5 % der Werte von Chiquadrat auftreten können.



Die Wahrscheinlichkeit, aus Zufallsprozessen in chiquadratverteilten Variablen mit 5 Freiheitsgraden Werte von größer als 11 zu erhalten, wäre also gleich 0,05, wie nicht nur aus dieser empirischen Verteilung, sondern auch aus den Tabellen zu den Flächen unter der Chiquadrat-Verteilung abgelesen werden kann.

5.2 Die t-Verteilung

Diese Verteilung ist ein „Zwitter“, da sie als Quotient aus einer Standard-Normalverteilung und einer Chi²-Verteilung mit n Freiheitsgraden wie folgt konstruiert wird:

$$T_{df} = \frac{z}{\sqrt{\frac{\chi_{df}^2}{df}}}$$

Die Variable z ist dabei „standard-normalverteilt“ mit dem Mittel 0 und der Standardabweichung 1; die Variable χ_{df}^2 ist chiquadrat-verteilt, mit df Freiheitsgraden.

Die Entstehung einer t-Verteilung kann an der folgenden Tabelle nachvollzogen werden:

Z_1^2	Z_2^2	Z_3^2	Z_4^2	Z_5^2	Z_6	$\chi^2_{df5}/5$	t_{df5}
,83	,00	,00	,24	,58	-1,26	,33	-2,20
,00	,77	,16	1,12	,17	1,94	,45	2,90
,08	2,30	,00	,94	,69	1,39	,80	1,55
,13	,07	,17	1,24	,75	-,70	,47	-1,02
3,44	,00	,02	3,10	4,18	-1,28	2,15	-,87
3,12	1,86	,10	,23	,23	,43	1,11	,41
,10	,05	,68	,44	3,85	,24	1,02	,24
2,67	,10	,00	,01	2,96	-1,01	1,15	-,94
,03	,48	,00	1,29	,14	1,06	,39	1,70
,10	,96	,09	1,07	,50	-,18	,55	-,24
,12	,00	1,42	4,52	,32	-,56	1,28	-,50
1,35	1,12	7,51	2,92	,15	1,45	2,61	,90
2,04	1,92	,63	,20	2,84	,68	1,52	,55
,76	,08	1,97	,21	,01	,02	,61	,03
,20	,61	1,14	,01	,55	,84	,50	1,19
3,19	,10	,17	1,97	1,78	1,52	1,44	1,27
2,32	1,30	,94	,76	1,07	-,24	1,28	-,21
1,03	1,32	,20	,23	1,07	-1,63	,77	-1,87
,54	1,84	5,74	,82	,05	-,58	1,80	-,43
,34	1,49	,45	3,95	,07	,79	1,26	,70

Durch Addition entsteht aus 5 quadrierten z-Verteilungen eine Chiquadratverteilung mit 5 Freiheitsgraden. Teilt man die Variable z_6 , die nicht quadriert wurde, durch die Chiquadratverteilung, deren einzelne Werte zuvor noch durch 5 (die Menge ihrer Freiheitsgrade) geteilt und aus denen noch die 2. Wurzel gezogen wurde, dann erhält man die gewünschte t-Verteilung mit 5 Freiheitsgraden.

Die Dichtefunktion der t-Verteilung lautet

$$h_{df}(t) = \frac{\Gamma \frac{df+1}{2}}{\sqrt{df\pi} \Gamma \frac{df}{2}} \left[\left(1 + \frac{t^2}{df} \right)^{-\frac{df+1}{2}} \right]$$

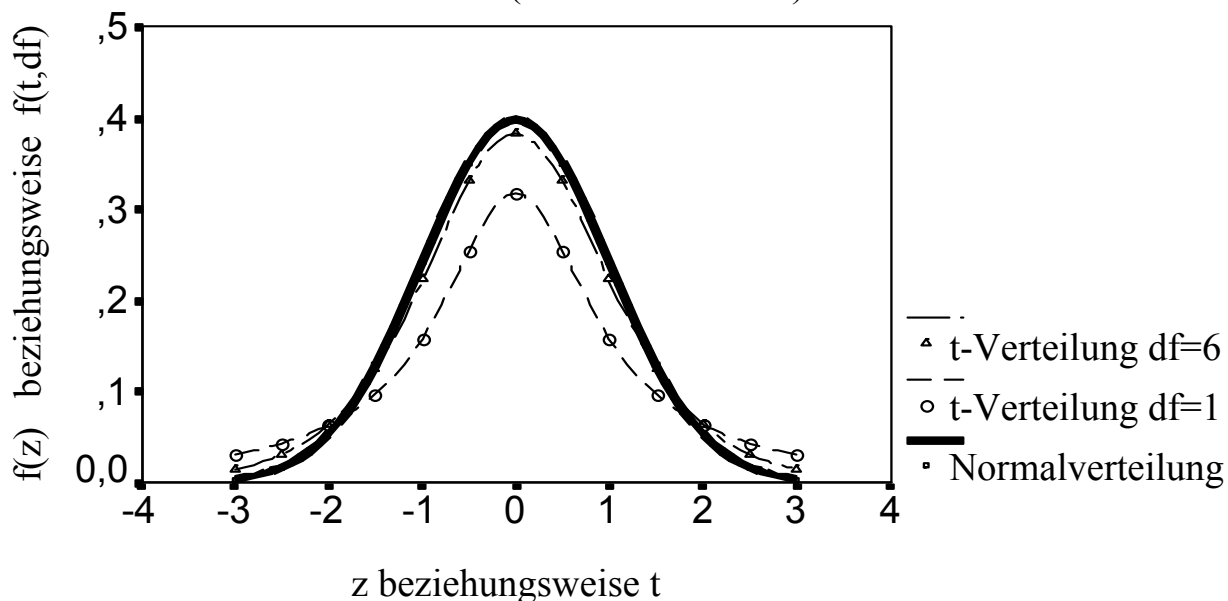
mit df als Freiheitsgraden.

Die t-Verteilung hat, wenn die Freiheitsgrade 3 und größer sind, einen Erwartungswert (Mittel) von 0 und eine Varianz von

$$s^2 = \frac{df}{df - 2}$$

Deshalb wird mit wachsenden Freiheitsgraden wird die t-Verteilung der Normalverteilung ähnlich, weil die Varianz der Standardnormalverteilung 1 ist und bei wachsenden Freiheitsgraden der Ausdruck für die Varianz der t-Verteilung sich 1 annähert.

Vergleich von Normalverteilung und t-Verteilungen mit $df = 1$ und $df = 6$ (Dichtefunktionen)



Praktisch bedeutsam wird die t-Verteilung z.B. im Zusammenhang der Prüfung von Hypothesen über Mittelwertsunterschiede zweier Stichproben und der Prüfung der Signifikanz von Korrelationskoeffizienten.

5.3 Die F-Verteilung

Die F-Verteilung entsteht aus dem Quotienten zweier chiquadrat-verteilter Zufallsvariablen, die verschieden viele Freiheitsgrade haben können (m und n), und die jeweils durch die Menge ihrer Freiheitsgrade geteilt werden:

$$F_{(m,n)} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

Demonstration der Formel:

Die folgende Tabelle enthält 4 quadrierte z-Verteilungen z_m^2 bis z_m^2 , deren Summe, durch 4 geteilt, $\text{chiquad}_m / m$ bildet; weiterhin enthält die Tabelle 6 quadrierte z-Verteilungen z_n^2 bis z_n^2 , deren Summe, durch 6 geteilt, $\text{chiquad}_n / n$ bildet. Der Quotient $(\text{chiquad}_m / m) / (\text{chiquad}_n / n)$ wird gebildet und man erhält die F-verteilte Variable

$f_{4;6}$

mit 4 Freiheitsgraden im Zähler und 6 Freiheitsgraden im Nenner:

zm^2_1	zm^2_2	zm^2_3	zm^2_4	zn^2_1	zn^2_2	zn^2_3	zn^2_4	zn^2_5	zn^2_6	$\frac{\chi^2_m}{m}$	$\frac{\chi^2_n}{n}$	$f_{4;6}$
,83	,12	,00	,00	,00	1,42	,24	4,52	,58	,32	,24	1,18	,20
,00	1,35	,77	1,12	,16	7,51	1,12	2,92	,17	,15	,81	2,01	,40
,08	2,04	2,30	1,92	,00	,63	,94	,20	,69	2,84	1,58	,88	1,79
,13	,76	,07	,08	,17	1,97	1,24	,21	,75	,01	,26	,72	,36
3,44	,20	,00	,61	,02	1,14	3,10	,01	4,18	,55	1,06	1,50	,71
3,12	3,19	1,86	,10	,10	,17	,23	1,97	,23	1,78	2,07	,75	2,77
,10	2,32	,05	1,30	,68	,94	,44	,76	3,85	1,07	,94	1,29	,73
2,67	1,03	,10	1,32	,00	,20	,01	,23	2,96	1,07	1,28	,75	1,71
,03	,54	,48	1,84	,00	5,74	1,29	,82	,14	,05	,72	1,34	,54
,10	,34	,96	1,49	,09	,45	1,07	3,95	,50	,07	,72	1,02	,71

Die Dichte-Funktion lautet (für $x \geq 0$)

$$g(m,n) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{(\chi)^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n} \chi\right)^{\frac{m+n}{2}}}$$

Eine Zufallsvariable, die F-verteilt ist, hat den Erwartungswert (arithmetisches Mittel)

$$\bar{x} = \frac{n}{n-1}$$

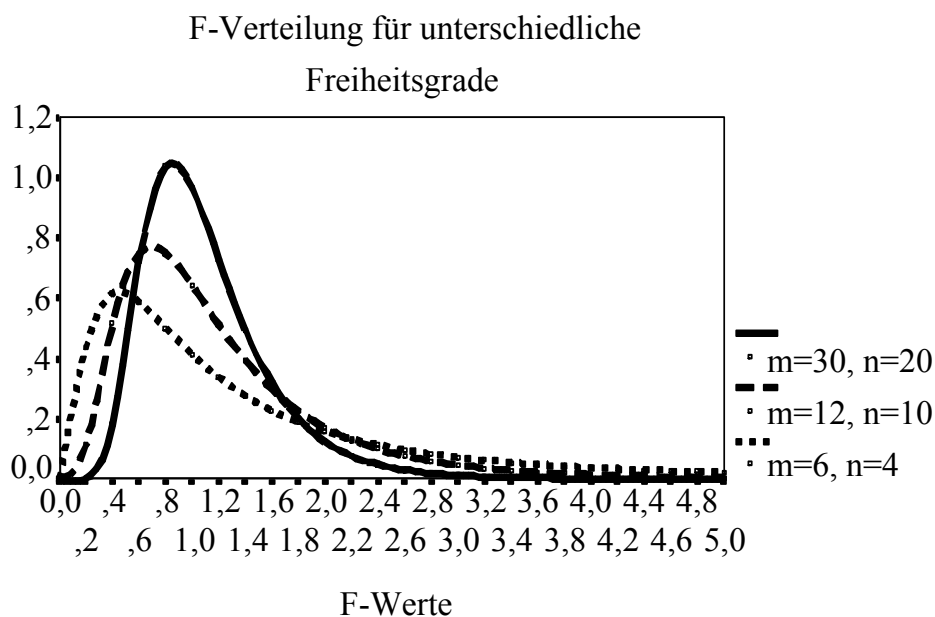
für $n > 3$

und die Varianz

$$s^2 = \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2} \text{ für } m > 5$$

(Vgl. Bamberg, Bauer: Statistik, 3. Auflage, Oldenburg, S. 11).

Die Integrale dieser Dichtefunktionen liegen tabelliert vor, wobei zu beachten ist, daß jede F-Verteilung zwei verschiedene Angaben über die Freiheitsgrade m und n im Zähler und im Nenner auf sich vereinigt.



Praktisch bedeutsam wird die F-Verteilung z.B. bei der Untersuchung der Frage, ob mehrere Stichproben mit verschieden großen Varianzen aus der gleichen - normalverteilten - Grundgesamtheit gezogen worden sein können. Siehe hierzu das Kapitel über den F-Test unten.

6. Der Standardfehler von Stichprobenkennwerten

6.1. Allgemeines.

Terminologisch muß folgendes nochmals betont werden: Für **Stichproben** werden die Lage- und Streuungsmaße „**Kennwerte**“ genannt. Sie werden mit lateinischen Buchstaben (\bar{x}, s) bezeichnet. Berechnet man Lage- und Streuungsmaße von **Grundgesamtheiten**, dann bezeichnen wir diese mit griechischen Buchstaben (arithmetisches Mittel = μ ; Standardabweichung = σ) und nennen sie „**Parameter**“.

Stichprobenkennwerte streuen um die entsprechenden Parameter der Grundgesamtheit, wie leicht nachzuvollziehen ist, denn es wäre schon ein merkwürdiger Zufall, wenn man beim Stichprobenziehen immer Kennwerte bekäme, die genau den Parametern entsprächen. Interessant ist nun, ob die Streuung der Kennwerte bekannt ist und ob man hieraus statistisch Nutzen ziehen kann.

6.2. Der Standardfehler des arithmetischen Mittels

6.2.1 Demonstration des Standardfehlers des arithmetischen Mittels

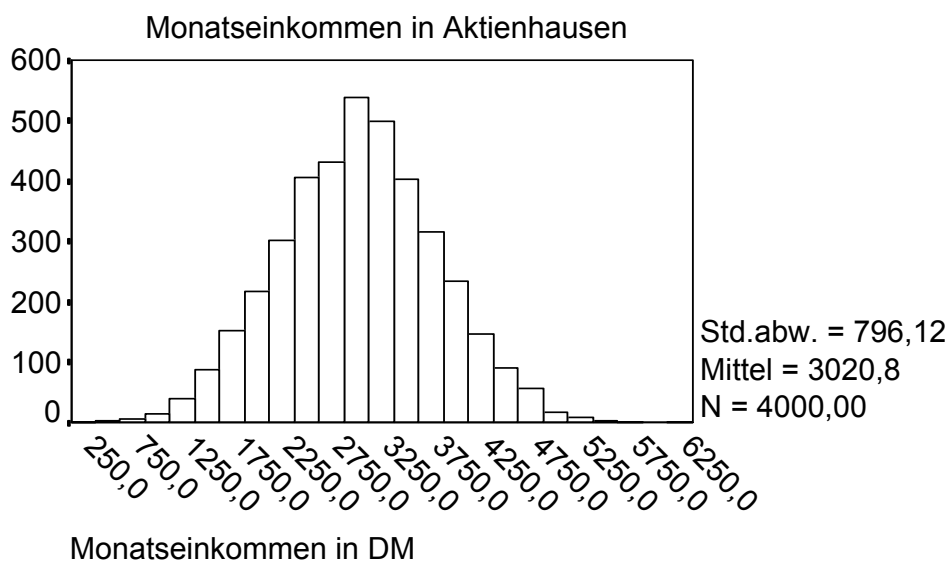
Wenn wir wissen wollen, in welchem Bereich wir den Mittelwert einer Stichprobe erwarten dürfen, wenn wir die Stichprobengröße und den wahren Mittelwert kennen, dann könnten wir hierzu ein Experiment durchführen. Wir ziehen aus einer großen Grundgesamtheit fortlaufend Stichproben eines bestimmten Umfangs und notieren jeweils den Mittelwert \bar{x} . Die notierten Mittelwerte stellen wir wiederum in einer Häufigkeitsverteilung dar und berechnen für diese die Standardabweichung.

Beispiel: Aus einer Stichprobe mit 4000 Personen aus Aktienhausen werden 10 Stichproben (Einkommensangaben in DM) gezogen.

Die wahren Werte (Parameter der Grundgesamtheit) sind:

Arithmetisches Mittel	$\mu = 3020,80$
Standardabweichung	$\sigma = 796,12$
Minimum	323,39
Maximum	6323,96
n	= 4000

Graphik der Häufigkeitsverteilung in der Grundgesamtheit:



10 Stichprobenkennwertgruppen, die bei ca. 5-prozentiger Auswahl zustandekommen, sind:

Variable	ar. Mittel	Std-Abw.	Minimum	Maximum	n	
STICHPROBE NR. 1	3050,37	821,37	1099,84	4691,61	174	
STICHPROBE NR. 2	3000,93	745,35	558,16	4741,84	186	
STICHPROBE NR. 3	3068,58	816,77	1000,76	6323,96	223	****
STICHPROBE NR. 4	2947,53	776,78	718,65	5329,43	211	****
STICHPROBE NR. 5	3056,28	854,96	958,02	6323,96	182	
STICHPROBE NR. 6	3006,44	817,24	459,24	5416,04	213	
STICHPROBE NR. 7	2994,61	849,97	447,67	5298,42	200	
STICHPROBE NR. 8	2948,32	754,04	1345,91	5036,31	189	****
STICHPROBE NR. 9	2970,61	776,48	323,39	4821,20	186	****
STICHPROBE NR. 10	3019,45	807,99	1060,44	5549,10	202	

Berechnen wir aus den 10 Stichprobenmittelwerten eine Standardabweichung, dann erhalten wir folgendes Ergebnis:

Standardabweichung der Stichprobenmittelwerte = 43,14

Diese Standardabweichung der Stichprobenmittelwerte wird nun **Standardfehler** genannt. Wir bemerken, daß die 10 Mittelwerte der Stichproben nicht identisch mit dem wahren Wert sind. Wir können aber feststellen, daß 6 von 10, also 60 % unserer Stichprobenmittel, im Intervall von $\mu \pm$ ein Standardfehler liegen. Die Werte sind oben fett gedruckt. Alle Stichprobenmittelwerte liegen in einem Intervall von \pm zwei Standardfehlern.

Überprüfe dies an folgenden Angaben:

Wahrer Mittelwert + 1 empirischer Standardfehler: $3020,8 + 43,14 = 3063,94$

Wahrer Mittelwert - 1 empirischer Standardfehler: $3020,8 - 43,14 = 2977,66$

Wahrer Mittelwert + 2 empirische Standardfehler: $3020,8 + 86,28 = 3107,08$

Wahrer Mittelwert - 2 empirische Standardfehler: $3020,8 - 86,28 = 2934,52$

Wenn man den Standardfehler nun mit wenig Aufwand schätzen könnte, dann hätte man eine gute Methode, um den Bereich abzuschätzen, in dem der wahre Wert μ mit beispielsweise 60 % Wahrscheinlichkeit liegt, wenn man eines der Stichprobenmittel kennt. Hierzu machen wir uns folgende Gesetzmäßigkeit zunutze:

Der Standardfehler des Mittelwertes einer Verteilung

$$\text{beträgt } \sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$$

und ist gleich der Standardabweichung der Mittelwerte von Stichproben im Umfang n aus einer Population. Dabei ist σ^2 die Populationsvarianz.

Eine Häufigkeitsverteilung, die wir erhalten würden, wenn wir sehr viele Stichproben mit jeweils gleich großem Umfang aus einer bestimmten Grundgesamtheit ziehen und jeweils das arithmetische Mittel dieser Stichproben notieren würden, hätte eine Standardabweichung von $\sigma_{\bar{x}}$.

Ist die Populationsvarianz unbekannt, wird sie geschätzt durch die Stichprobenvarianz, die mit dem Faktor $n/(n-1)$ multipliziert wird.

$$\hat{\sigma}_{\bar{x}} = \sqrt{\hat{\sigma}^2} \cdot \frac{1}{n} = \sqrt{\frac{1}{n} \cdot \frac{\sum (x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}}$$

Der Standardfehler des Mittelwerts kann somit, wie leicht zu erkennen ist, auch geschrieben werden als

$$\sigma_{\bar{x}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{s}{\sqrt{n}};$$

Der „zentrale Grenzwertsatz“ besagt, daß die Mittelwerte von Stichproben des Umfangs n (für $n > 30$), die ein und derselben Grundgesamtheit entnommen werden, **normalverteilt sind, unabhängig davon, ob die Grundgesamtheit auch normalverteilt ist**, und zwar mit dem Mittelwert μ und der Standardabweichung von $\sigma_{\bar{x}}$.

6.2.2. Zur Bedeutung des Standardfehlers des arithmetischen Mittels

Der Standardfehler ist, wie wir wissen, die Standardabweichung des arithmetischen Mittels von Stichproben mit der Größe n . Er würde die Standardabweichung einer Verteilung darstellen, die bei zahlreichen Stichprobenziehungen aus einer gegebenen Grundgesamtheit für die Variable „Arithmetisches Mittel der Stichprobe im Umfang von n “ entstünde.

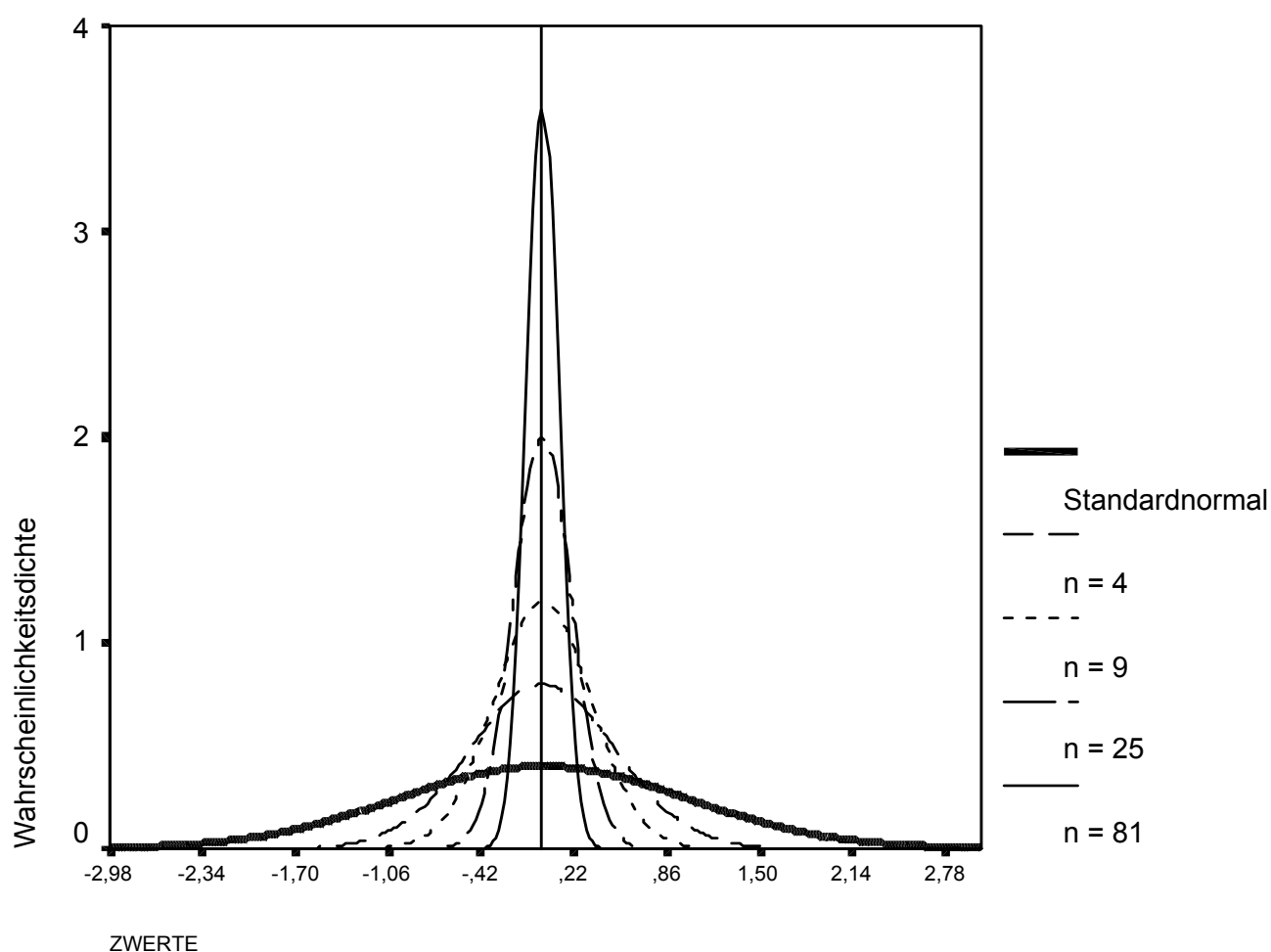
Er ist definiert durch (s. o.)

$$\sigma_{\bar{x}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{s}{\sqrt{n}};$$

Wird nun n variiert, dann ergeben sich für die Verteilung des Stichprobenmittels unterschiedliche Gestalten. Die Kurve der Wahrscheinlichkeitsdichte wird um so steiler, je größer n ist. Die folgende Graphik demonstriert dies für Stichproben mit

$n = 4$,
 $n = 9$,
 $n = 25$,
 $n = 81$,

die aus einer standard-normalverteilten Grundgesamtheit gezogen wurden:



Der Bereich, in dem das arithmetische Mittel einer Stichprobe auftauchen kann, wird daher mit wachsendem n der Stichprobe immer schmäler.

Für SPSS-Interessenten sei noch mitgeteilt, wie man mit Hilfe von SPSS eine solche Graphik herstellt:

1. Man definiert auf dem leeren Daten-Editor 301 Fälle mit dem Befehl:
 „Daten“ – „gehe zu Fall 301“ „Eingabe irgendeiner Zahl in ein Feld des Falles 301“.
2. Man weist diesen Fällen z-Werte mit folgendem Befehl zu:

```
COMPUTE zwerte = $casenum/50 - 3 .
EXECUTE .
```

3. Man weist den Fällen standardnormalverteilte Ordinaten mit folgendem Befehl zu:

```
COMPUTE normalv = (1/((1/1)*2.50663))*2.71828**(0-0.5*((zwerte-0)/(1/1))**2).
EXECUTE .
```

In diesem Befehl steckt die Wurzel aus n an den Stellen, die hier durch Pfeile bezeichnet sind. Bei der Standardnormalverteilung ist Wurzel aus n = 1.

4. Man erzeugt eine Graphik der Standardnormalverteilung mit folgendem Befehl:

```
GRAPH
/LINE(SIMPLE)=VALUE(normalv) BY zwerte .
```

4. Man erzeugt für die Verteilung der Standardfehler mit n = 4, n = 9, n = 25, n = 81 entsprechende Variablen durch folgende Befehlsketten, in die an den unterstrichenen Stellen jeweils die Wurzel aus n eingefügt ist:

n = 4

```
COMPUTE nor4 = (1/((1/2)*2.50663))*2.71828**(0-0.5*((zwerte-0)/(1/2))**2).
EXECUTE .
```

n = 9

```
COMPUTE nor9 = (1/((1/3)*2.50663))*2.71828**(0-0.5*((zwerte-0)/(1/3))**2).
EXECUTE .
```

n = 25

```
COMPUTE nor25 = (1/((1/5)*2.50663))*2.71828**(0-0.5*((zwerte-0)/(1/5))**2).
EXECUTE .
```

n = 81

```
COMPUTE nor81 = (1/((1/9)*2.50663))*2.71828**(0-0.5*((zwerte-0)/(1/9))**2).
EXECUTE .
```

5. Die Graphik erzeugt man durch den Befehl

```
GRAPH
/LINE(MULTIPLE)= VALUE(normalv nor4 nor9 nor25 nor81 ) BY zwerte .
```

Diese Graphik kann dann weiterbearbeitet werden.

Viel Vergnügen beim Nacharbeiten !

6. 3. Der Standardfehler von Prozentsätzen

Der Standardfehler von Wahrscheinlichkeiten läßt sich – in Anlehnung an die Binomialverteilung – ebenfalls bestimmen, sofern man die Frage auf Situationen begrenzt, wo in der Stichprobe zwei Eigenschaften vorliegen, die einander ausschließen (wie „krank“ und „gesund“). In diesem Fall kann der Standardfehler bestimmt werden aus der Gleichung:

$$\sigma_p^2 = \frac{p * q}{n}$$

p und q sind hier die einander ausschließenden Wahrscheinlichkeiten, in deren Umfang in der Stichprobe jemand krank oder gesund ist. Es muß allerdings gelten: $npq > 9$. Wenn z. B. $p = 0,3$ und $q = 0,7$ sind, dann müßte n mindestens 43 betragen, weil dann $npq = 9,03$ wäre.

Der Standardfehler für Wahrscheinlichkeiten wäre dann:

$$\sigma_p = \sqrt{\frac{p * q}{n}}$$

Beispiel: Wir ziehen eine Stichprobe von 200 Tomaten, von denen 20 % krank sind ($p = 0,2$) und 80 % gesund sind ($q = 0,8$). Der Standardfehler dieser Wahrscheinlichkeiten würde sich wie folgt errechnen:

$$\sigma_p = ((0,2 * 0,8)/200)^{0,5} = 0,0283$$

Will man hieraus einen Standardfehler für Prozentsätze machen, multipliziert man das Ergebnis mit 100. Es käme also ein Standardfehler für den Prozentsatz kranker Tomaten von 2,83 Prozent heraus.

7. Statistische Tests

Im Unterschied zu den Intervall-Schätzverfahren, bei denen für einen Parameter (z. B. μ) ein Bereich angegeben wird, in dem dieser unter bestimmten Voraussetzungen (z.B. Normalverteilung und Kenntnis der Varianz der Grundgesamtheit bei gegebenen Stichprobendaten) mit einer bestimmten Wahrscheinlichkeit liegt, werden beim Testen statistischer Hypothesen Bereiche einer Prüfverteilung bestimmt, innerhalb deren eine sogenannte „Nullhypothese“ mit einer bestimmten „Irrtumswahrscheinlichkeit“ beibehalten bzw. abgelehnt wird. Die Werte der Prüfverteilung werden mit einem Kritischen Wert c verglichen, der seinerseits einer bestimmten Irrtumswahrscheinlichkeit entspricht, deren Niveau vorab festgelegt wird. In der Regel ist dies die Wahrscheinlichkeit, eine Nullhypothese irrtümlich abzulehnen, obwohl sie richtig ist.

Beim Aufstellen statistischer Hypothesen formuliert man in der Regel eine Alternative in der Form: "Wir nehmen an, es ist **so** und nicht **so**." Die Aussage bezieht sich auf Eigenschaften der Grundgesamtheit, die Basis der Aussage sind, u.a. Stichprobendaten. Die Stichprobendaten und Überlegungen darüber, welches Risiko man eingehen will, sich bei der Annahme oder Verwerfung einer Hypothese zu irren, führen zur Entscheidung **für** die eine der Alternativen und **gegen** die andere.

7.1. Zur Terminologie

Die beiden Alternativen, um die es hierbei geht, werden in der Statistik als "Nullhypothese" und "Alternativ-Hypothese" bezeichnet.

Bamberg/Baur sprechen von der **zu untersuchenden** Hypothese immer als "**Nullhypothese**" (H_0). Die Gegenhypothese nennen sie immer **Alternativhypothese** (H_1). Noch eine weitere terminologische Festlegung spielt in diesem Zusammenhang eine Rolle:

Der Fehler 1. Art oder alpha-Fehler liegt vor, wenn ich die Alternativ-Hypothese H_1 annehme, obwohl H_0 richtig ist. Die Wahrscheinlichkeit für das Eintreten dieses Fehlers heißt α (alpha).

Der Fehler 2. Art oder beta-Fehler liegt vor, wenn ich die Null-Hypothese annehme, obwohl H_1 richtig ist. Die Wahrscheinlichkeit für das Eintreten dieses Fehlers heißt β (beta).

Die Größe α (alpha) wird als "Irrtumswahrscheinlichkeit" bezeichnet, die Größe $1 - \alpha$ als Sicherheitswahrscheinlichkeit oder auch als „Signifikanzniveau“ (Bosch, K.: Elementare Einführung in die angewandte Statistik, a.a.O., S. 75).

7.2. Zur Struktur der Statistischen Tests

(1) Für jede Fragestellung eines statistischen Tests gibt es eine Prüfverteilung (Testfunktion), deren Integral gleich 1 ist. Diese Prüfverteilung ist in Abhängigkeit davon zu wählen, wie das Verteilungsgesetz jener Kennwerte ist, mit deren Hilfe unsere Nullhypothese formuliert wird.

(2) Für den Wertebereich der Prüfverteilung wird bestimmt, wie groß α oder $(1 - \alpha)$ sein sollen und wo daher der **Annahmereich** und wo der **Ablehnungsbereich** für H_0 liegen soll. Die Grenze zwischen beiden Bereichen soll c , der „Kritische Wert“, markieren.

(3) Es wird geprüft, welcher Wert der Testfunktion aus den empirischen Daten berechnet werden kann und ob dieser empirische Wert in den Annahme- oder Ablehnungsbereich für H_0 fällt.

7.3. Beispiele

7.3.1. Z-Test und t-Test auf Abweichung der Stichprobe von der Grundgesamtheit

7.3.1.1. Z-Test

Beispiel für eine große Stichprobe mit beliebig verteilter Grundgesamtheit:

49 Studierende (Stichprobe) erlangen in einer Klausur im Durchschnitt 75 Punkte und eine Standardabweichung von 9 Punkten, in der zuvor eine große Gruppe (Population) Studierender im Durchschnitt nur 72 Punkte erreichten. Wir wollen wissen, ob es sich um eine zufällige Abweichung handelt, bei einer Irrtumswahrscheinlichkeit von $\alpha = 1\%$

Zunächst legen wir fest, daß unser Test „zweiseitig“ sein soll, das heißt, wir machen keine Vorannahme darüber, ob erwartet wird, daß unser empirisches arithmetisches Mittel der Stichprobe größer oder kleiner als das arithmetische Mittel der Grundgesamtheit ist.

Die Nullhypothese hätte hiernach die Formulierung: $H_0 : \mu = \bar{x}$

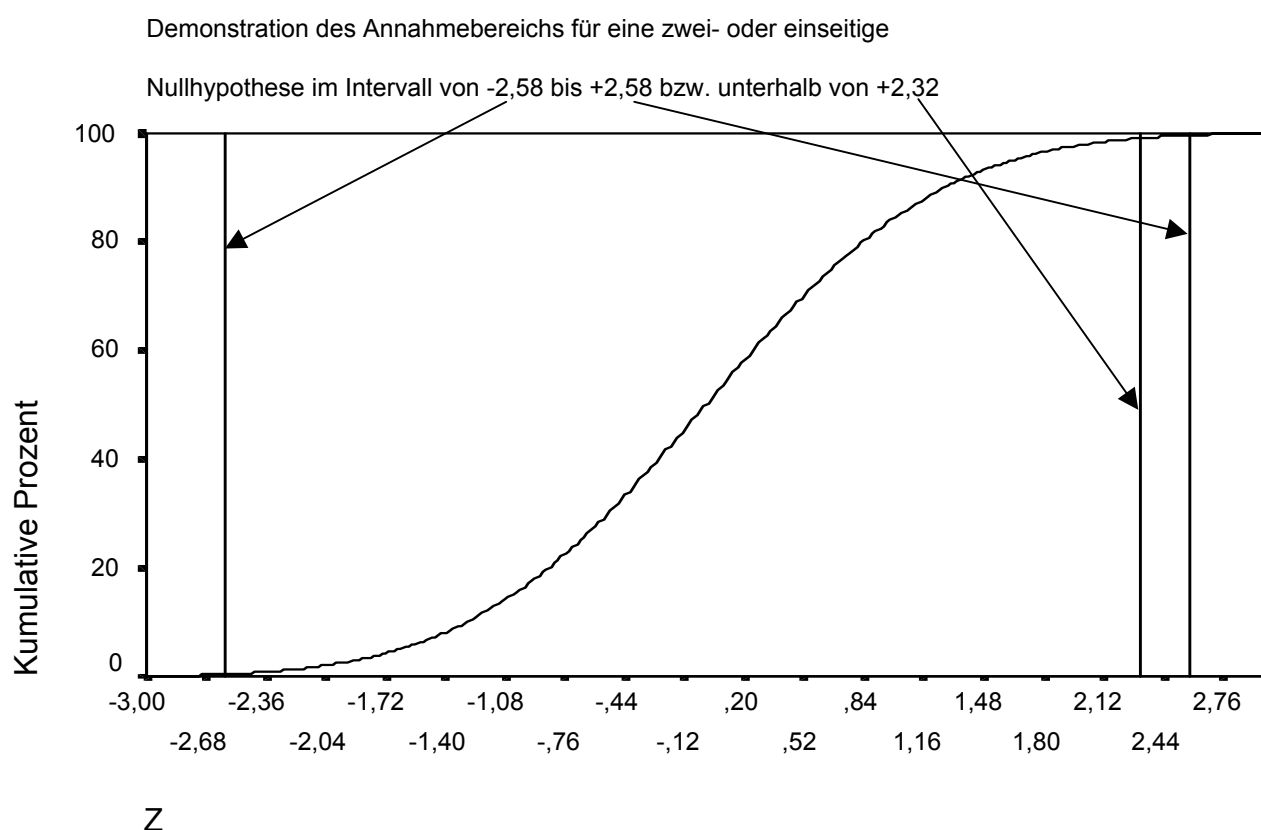
Die Alternativhypothese hätte hiernach die Formulierung: $H_1 : \mu \neq \bar{x}$

Hätten wir uns für einen „einseitigen“ Test der Nullhypothese entschieden, dann könnte die Alternativhypothese zwei unterschiedliche Formulierungen haben: entweder $H_1 : \mu < \bar{x}$ oder $H_1 : \mu > \bar{x}$.

Für einen zweiseitigen Test müßten wir nun einen kritischen Wert (c) ermitteln, der uns hilft, die Bedingung einzuhalten, daß die Irrtumswahrscheinlichkeit $\alpha = 1\%$ sein soll.

Wir wissen, daß die Mittelwerte aus hinreichend großen Zufallsstichproben *normalverteilt* sind, mit der Standardabweichung in der Größenordnung des Standardfehlers und dem arithmetischen Mittel μ . Hierzu stellen wir durch Analyse der Tabelle der *Standardnormalverteilung* fest, daß außerhalb des Intervalls $-2,58$ bis $+2,58$ etwa 1% der Fläche der Dichtefunktion der Standardnormalverteilung liegen, das heißt: oberhalb von $z = +2,58$ liegen 0,5 % der Fläche, unterhalb von $z = -2,58$ liegen ebenfalls 0,5 % der Fläche, was zusammen 1 % der Fläche sind.

Wie können wir diese Feststellung anhand der Tabelle der Standardnormalverteilung machen ? Dies hier zu erklären, würde voraussetzen, daß hier die entsprechende Tabelle hier abgedruckt wird, was aber nicht geschieht. Sie befindet sich, mit Gebrauchsanweisung im Anhang. Statt dessen wird versucht, das Prinzip graphisch zu erläutern:



Dem Tabellenwert in der Standardnormalverteilung $z = -2,58$ entspricht eine von links nach rechts aufkumulierte Wahrscheinlichkeit von $0,0049 \approx 0,005$. Das sind 0,5 % der Fläche der Standardnormalverteilung. Wegen der Symmetrie der Standardnormalverteilung ist das gleichzeitig die Fläche, die oberhalb des Tabellenwertes von $+2,58$ noch bis 100 % fehlt. Wir können daher den kritischen Wert c als jenen z -Wert bestimmen, der gleich $2,58$ ist.

Wir errechnen nun unser empirisches z :

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{75 - 72}{9 / \sqrt{49}} = 3 / 1,28 = 2,3333$$

Unser empirisches z ist kleiner als c . Daher behalten wir die Nullhypothese bei 1% Irrtumswahrscheinlichkeit und zweiseitigem Test bei.

Hätten wir aber erwartet, daß unser arithmetisches Mittel aus der Stichprobe auf jeden Fall größer ist als das Mittel der Grundgesamtheit (einseitiger Test), dann hätten wir einen kritischen Wert von $c=2,326347$ erhalten, da rechts von diesem Wert in der Dichtefunktion der Standardnormalverteilung ca. 1% ihrer Fläche liegt.

$Z=2,333$ ist geringfügig größer als dieses neu bestimmte c für den einseitigen Test. Unter diesen Bedingungen hätten wir die Nullhypothese abgelehnt und die Alternativhypothese angenommen.

7.3.1.2. t-Test

20 Studierende erlangen in einer Klausur im Durchschnitt 75 Punkte, in der zuvor 1000 Studierende im Durchschnitt nur 72 Punkte erreichten, mit einer Standardabweichung von 9 Punkten. Wir wollen wissen, ob es sich um eine zufällige Abweichung handelt, bei einer Irrtumswahrscheinlichkeit von $\alpha=1\%$. Wir führen einen einseitigen Test durch. Die Nullhypothese lautet wieder:

$$H_0 : \mu = \bar{x}$$

Die Alternativhypothese hätte die Formulierung:

$$H_1 : \mu < \bar{x}$$

Da wir wissen, daß das arithmetische Mittel *kleiner* Stichproben aus einer Grundgesamtheit *nicht normalverteilt* ist, sondern t-verteilt, errechnen wir jetzt keinen empirischen z-Wert – wie oben – sondern einen empirischen t-Wert (nach der gleichen Formel wie vorher den empirischen z-Wert).

$$t = \frac{75 - 72}{9 / \sqrt{20}} = 3 / 2,01 = 1,49$$

Woher bekommen wir nun einen „kritischen t-Wert“, den wir auch wieder c nennen können und mit dem wir unser empirisches t vergleichen können? Wir finden ihn in jener t-Verteilung, die $n-1$ Freiheitsgrade hat, also hier $20-1 = 19$ Freiheitsgrade. Für die t-Verteilung mit 19 Freiheitsgraden gilt, daß rechts vom Wert $t=2,539$ etwas weniger als 1% der Gesamtfläche der Verteilung liegt. Diese Feststellung wäre bei einem einseitigen Test mit 1% Irrtumswahrscheinlichkeit zu machen. Für diesen Test wäre also $c=2,539$. Wir müssen in diesem Fall die Nullhypothese beibehalten, daß unsere Stichprobe aus der Grundgesamtheit hätte gezogen werden können, denn $t=1,49$ ist kleiner als $c=2,539$.

Die beiden Beispiele zeigen, daß bei fast identischen Bedingungen, nur durch Variation der Stichprobengröße, veränderte Ergebnisse unserer Hypothesenprüfung entstehen, weil wir bei kleinen Stichproben die t-Verteilung benutzen.

7.3.2. t-Test für die Differenz zweier Mittelwerte aus unabhängigen Stichproben aus 2 varianzhomogenen Grundgesamtheiten

Es seien zwei unabhängige Stichproben gegeben, die aus normalverteilten Grundgesamtheiten stammen, deren Varianzen als annähernd gleich groß angenommen werden, die aber unbekannt sind. Für diese Stichproben kenne ich die arithmetischen Mittel \bar{x}_i die Standardabweichungen s_i und den Umfang n_i .

Für diesen Fall können wiederum verschiedene Formulierungen einer Alternativhypothese H_1 zu einer einheitlichen Nullhypothese H_0 (H_0 = die beiden Stichproben stammen aus einer gemeinsamen Grundgesamtheit) vorgenommen werden, wie folgt:

$H_0: \mu_1 = \mu_2$ (a) $H_1: \mu_1 \neq \mu_2$ („zweiseitiger Test“)

oder (b) $H_1: \mu_1 < \mu_2$ („einseitiger Test“)

oder (c) $H_1: \mu_1 > \mu_2$ („einseitiger Test“)

Wir wissen, daß für kleinere Stichproben, deren Elemente zusammen weniger als 50 ausmachen, die Differenz der Mittelwerte t-verteilt ist, mit $n_1 + n_2 - 2$ Freiheitsgraden.

Um hiervon Gebrauch machen zu können, müssen wir den Standardfehler der Mittelwertdifferenz schätzen.

Er errechnet sich nach folgender Schätzformel:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} * \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

Wenn die geschätzten Populationsvarianzen (die gleich den Stichprobenvarianzen sind) bekannt sind, gilt für den Standardfehler der Mittelwertsdifferenzen:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1) * \hat{\sigma}_1^2 + (n_2 - 1) * \hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)} * \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

Haben wir diesen ermittelt, dann können wir den zugehörigen t- bzw. z-Wert wie folgt ermitteln:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

d.h. wir setzen den Standardfehler in den Nenner des o.a. Bruchs.

Die Anwendung dieser Formel ist nur gestattet, wenn die Grundgesamtheiten normalverteilt sind, aus denen die Stichproben entnommen wurden und wenn die Varianzen dieser Grundgesamtheiten „homogen“ sind. Für den Fall inhomogener Varianzen müssen Verfahren mit Korrekturformeln benutzt werden. Ob homogene Varianzen vorliegen, wird entweder mit dem F-Test oder mit dem Levene-Test geprüft (den wir hier nicht besprechen).

Ein Beispiel für die Anwendung der o.a. Formeln sei folgendes: Eine Studentengruppe (I) erziele im Intelligenztest folgende Werte:

97	97	106	109	102	105	96	95	98	103	105	101	102	99	100
97	111	97	106	102	98	102	98	104	107	100	98	98	110	101
101	103	95	103											

Eine andere Gruppe (II) erzielte folgende Werte:

106	114	99	101	105	98	110	110	100	102	104	102	109	101	110	
106	103	109	99	110	101	105	110	112	106	113	102	102	102	107	105

Wir wollen wissen, ob beide Gruppen aus der gleichen Stichprobe entnommen worden sein können, d.h. die Nullhypothese würde lauten:

$$H_0: \mu_1 = \mu_2$$

Die Alternativhypothese würde- wenn wir einen zweiseitigen Test wählen – lauten:

$$H_1: \mu_1 \neq \mu_2$$

Wir wollen die Nullhypothese bei 1 % Irrtumswahrscheinlichkeit prüfen.

Wir errechnen für die Gruppe 1 und 2 folgende Kennwerte:

Gruppenstatistiken

GRUPPE	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
1,00	34	101,3529	4,2489	,7287
2,00	31	105,2581	4,4719	,8032

Durch Quadrierung der Standardabweichungen gewinnen wir die Varianzen:

Varianz in Gruppe 1 18,053

Varianz in Gruppe 2 19,998

Zur Schätzung des Standardfehlers der Mittelwertdifferenz setzen wir ein:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(34-1)*18,053 + (31-1)*19,99}{(34-1) + (31-1)} * \left(\frac{1}{34} + \frac{1}{31}\right)} = 1,0819$$

t läßt sich jetzt leicht ermitteln:

$$t = \frac{101,35 - 105,26}{1,0819} = -3,610$$

Dieser t-Wert müßte jetzt mit einem Kriteriumswert c verglichen werden. Wir wissen, daß die Mittelwertsdifferenz zweier *unabhängiger* Stichproben t-verteilt ist, und zwar mit $n_1 - 1 + n_2 - 1$ Freiheitsgraden.

Die Fläche außerhalb des Intervalls von - 2,660 bis + 2,660 beträgt etwa 1%. Unser gesuchtes c wäre daher 2,66. Daher kann man hier die Nullhypothese, daß kein Unterschied der Populationen bestünde, aus denen die Stichproben gezogen sein könnten, mit mindestens 1% Irrtumswahrscheinlichkeit verwerfen (zweiseitiger Test), denn $c=2,66 < t=3,610$.

Bei einem einseitigen Test müßten wir hier die Alternativhypothese anders formulieren, und zwar wie folgt:

$$H_1: \mu_1 < \mu_2$$

Wir müßten – wie leicht festgestellt werden kann - mit einem $c = 2,39$ rechnen und kämen zum gleichen Ergebnis: Ablehnung der Nullhypothese.

Dies alles gilt allerdings nur unter der Voraussetzung, daß die Varianzen der beiden Gruppen „homogen“ sind. Diese Homogenität kann mittels des F-Test überprüft werden. Heutzutage wird in SPSS die Varianzhomogenität mittels des Levene-Tests überprüft, der hier nicht besprochen wird.

7.3.3. Der F-Test

Der F-Test – der neuerdings bei Durchführung des t-Test in SPSS durch den „Levene-Test“ ersetzt wurde – ist u. a. geeignet, die Homogenität zweier Varianzen zu prüfen. Das bedeutet, es wird die Nullhypothese geprüft

$$H_0 \Rightarrow \sigma_1^2 = \sigma_2^2, \text{ und zwar fast immer einseitig gegen die Alternativhypothese}$$

$$H_1 \Rightarrow \sigma_1^2 > \sigma_2^2$$

Daher setzt man für σ_1^2 immer die größere Varianz ein, für σ_2^2 die kleinere.

Die Teststatistik, F wird wie folgt berechnet:

$$F = \frac{S_m^2}{S_n^2}$$

Hierbei sind m und n die Freiheitsgrade (Stichprobengröße minus 1) der jeweiligen Varianzen, wobei m der größeren, n der kleineren Varianz zugerechnet wird. Ob die Nullhypothese (s.o.) beibehalten werden kann, richtet sich nach dem Vergleich eines empirischen F-Wertes mit einem kritischen F-Wert. Diesen gewinnen wir aus der Tabelle der F-Verteilung, in der wir in der Regel folgende Informationen vorfinden:

Kopfzeile: Freiheitsgrade für die größere Varianz

Vorspalte: Freiheitsgrade für die kleinere Varianz

Zelle: Zwei Werte: a) kritischer F-Wert für eine Irrtumswahrscheinlichkeit von 5 %. b) kritischer Wert für eine Irrtumswahrscheinlichkeit von 1 %.

Wird der kritische Wert von dem empirischen unterschritten, behalten wir die Nullhypothese bei, und – bei Homogenitätsprüfung – wäre unsere Varianzhomogenität gegeben. Wird der kritische Wert von dem empirischen überschritten, wird die Nullhypothese abgelehnt und die Varianzhomogenität wäre nicht gegeben. Im Falle des t-Tests müßten wir daher mit einem Verfahren vorlieb nehmen, das diese Inhomogenität der Varianzen berücksichtigt.

7.3.4. t-Test für die Differenz zweier Mittelwerte aus varianzheterogenen Grundgesamtheiten

Muß man im konkreten Falle die Voraussetzung fallen zu lassen, daß die Varianzen gleich groß sind und bleibt die Voraussetzung erhalten, daß die Grundgesamtheiten normalverteilt sind, aus denen die Stichproben gezogen wurden, dann lautet für diesen Fall die Formel der Testfunktion (Korrekturformel nach Welch):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Dieses t bezieht sich auf eine t-Verteilung mit Freiheitsgraden (df), die nach folgendem Verfahren ermittelt werden:

$$df = \frac{1}{\frac{g^2}{n_1 - 1} + \frac{(1 - g)^2}{n_2 - 1}}$$

$$\text{für } g = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In diesen Formeln bedeuten

\bar{x}_1 = arithmetisches Mittel der Stichprobe 1

\bar{x}_2 = arithmetisches Mittel der Stichprobe 2

n_1 = Umfang der Stichprobe 1

n_2 = Umfang der Stichprobe 2

s_1 = Standardabweichung der Stichprobe 1

s_2 = Standardabweichung der Stichprobe 2

Für unseren oben durchgeführten Gruppenvergleich bedeutet die Benutzung der Formel nach dem Welch-Test, daß der t-Wert sich verändert und auch die Menge der Freiheitsgrade jener t-Verteilung, aus der man den kritischen Wert gewinnt.

Eine vergleichende Rechnung zeigt die unterschiedlichen Ergebnisse:

Test bei unabhängigen Stichproben

		Varianzen sind gleich	Varianzen sind nicht gleich
Levene-Test der Varianzgleichheit	F	,322	
	Signifikanz	,573	
T-Test für die Mittelwertgleichheit	T	-3,610	-3,601
	df	63	61,706
	Sig. (2-seitig)	,001	,001
	Mittlere Differenz	-3,9051	-3,9051
	Standardfehler der Differenz	1,0819	1,0845

Bei Annahme der Varianzenhomogenität ist $t = -3,61$, und es wird an einer t-Verteilung geprüft mit 63 Freiheitsgraden. Bei Annahme der Inhomogenität ist es absolut etwas kleiner: $t = -3,601$, und es wird an einer Verteilung mit nur 61,7 Freiheitsgraden geprüft. Auffällig ist, daß hier „gebrochene“ Freiheitsgrade auftreten. In der Praxis muß man hier also runden.

7.3.5. t-Test für Mittelwertsdifferenzen aus *abhängigen* Stichproben.

Die unter 7.3.2. abgehandelten t-Tests beziehen sich auf die Fälle „unabhängiger“ Stichproben. Untersucht man aber Fälle mit Meßwiederholung an den Fällen einer einzigen Stichprobe, dann gelten die Datensätze mit den Meßwerten der Meßwiederholung als Stichproben, die von der ersten Stichprobe mit den ersten Messungen **abhängig sind. Eine andere Terminologie spricht auch von Meßwertpaaren.**

In diesem Fall interessieren wir uns nicht nur dafür, wie die Differenzen der Mittelwerte der ersten und der zweiten Stichprobe (vor und nach dem Ereignis, an dessen Auswirkungen wir interessiert sind), sondern wir wollen alle einzelnen Fälle in ihrer Veränderung erfassen. Daher bilden wir den Ausdruck der Meßwertdifferenz:

$$d_{23} = x_{23;1} - x_{23;2}$$

d ist die Differenz der ersten und der zweiten Messung a z.B. einer Versuchsperson mit der Nummer 23. Fassen wir alle Meßwertdifferenzen in ihrem arithmetischen Mittel zusammen, dann können wir schreiben:

$$\bar{x}_d = \frac{\sum_{i=1}^n d_i}{n}$$

Für den Fall, daß wir annehmen, in der Grundgesamtheit zeige sich überhaupt kein Effekt eines interessierenden Ereignisses, könnten wir die Nullhypothese als

$$H_0 \Rightarrow \bar{x}_d = \mu = 0$$

formulieren, die (ungerichtete) Alternativhypothese (für einen zweiseitigen Test) entsprechend:

$$H_0 \Rightarrow \bar{x}_d \neq \mu = 0$$

In diesem Fall können wir unser empirisches t wie folgt bestimmen:

$$t = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}}$$

$\hat{\sigma}_{\bar{x}_d}$ ist hierbei der Standardfehler der Meßwertpaardifferenzen. Er wird errechnet aus:

$$\hat{\sigma}_{\bar{x}_d} = \frac{s_{x_d}}{\sqrt{n}}$$

Die Freiheitsgrade bestimmen sich hier durch $df=n-1$, wobei n gleich der Anzahl der Meßwertpaare ist. Ist $n < 31$, dann müssen die Meßwertdifferenzen in der Grundgesamtheit normalverteilt sein, wenn man den t-Test benutzen will. Ist $n > 50$, dann können hier statt der t-Tabelle auch Werte aus der Tabelle der Standardnormalverteilung benutzt werden.

Beispielrechnung:

In einem Gymnasium werden die Schüler auf Musikalität hin getestet. Es gibt einen Vortest, ein Training und einen Nachtest. Gefragt ist, ob sich die Werte im Nachtest signifikant von den Werten des Vortests unterscheiden. In diesem Fall könnte angenommen werden, daß das Training sich positiv ausgewirkt hat.

Vortestwerte	Nachtestwerte	Nachtest minus Vortest
118,00	98,00	-20,00
102,00	133,00	31,00
101,00	127,00	26,00
111,00	104,00	-7,00
95,00	97,00	2,00
104,00	116,00	12,00
115,00	114,00	-1,00
91,00	100,00	9,00
102,00	123,00	21,00
97,00	110,00	13,00
101,00	105,00	4,00
93,00	127,00	34,00
116,00	119,00	3,00
118,00	112,00	-6,00
92,00	121,00	29,00
104,00	128,00	24,00
108,00	109,00	1,00
105,00	94,00	-11,00
109,00	105,00	-4,00
106,00	120,00	14,00
101,00	106,00	5,00
113,00	95,00	-18,00
105,00	106,00	1,00
105,00	114,00	9,00
108,00	98,00	-10,00
Summe: 2620,00	Summe: 2781,00	Summe: 161,00

Es ergeben sich:

Das arithmetische Mittel der Meßwert-Paar-Differenzen:

$$\bar{x}_d = \frac{161}{25} = 6,44$$

Die Standardabweichung der Meßwert-Paar-Differenzen:

$$s_d = 14,9835$$

Der Standardfehler ergibt sich aus der Division dieser Standardabweichung durch Wurzel aus n, also durch 5:

$$s_{\bar{x}_d} = \frac{14,9835}{5} = 2,9967$$

Der empirische t-Wert errechnet sich somit aus:

$$t = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}} = \frac{6,44}{2,9967} = 2,149$$

Der kritische t-Wert kann einer t-Verteilung mit $df = n-1 = 24$ Freiheitsgraden entnommen werden. Er beträgt für eine Irrtumswahrscheinlichkeit α von 5 %: $c = 2,06$, wie wir der Tabelle der t-Verteilung entnehmen können (bei zweiseitigem Test).

Da $c < t$ ist, wird hier die Nullhypothese abgelehnt und die Alternativhypothese angenommen.

7.3.6. Der Binomialtest

Der Test setzt voraus, daß eine Zufallsstichprobe gegeben ist, in der ein Merkmal in einer gegebenen Häufigkeit vorkommt, das nur zwei Ausprägungen hat. Die (kleinere) Häufigkeit des Auftretens einer der Ausprägungen kann dann daraufhin untersucht werden, ob die Stichprobe aus einer Grundgesamtheit gezogen wurde, in der eine bestimmte Wahrscheinlichkeit eben dieses Merkmals gegeben ist.

Mit Hilfe des Binomialtests läßt sich daher die Nullhypothese prüfen, daß eine bestimmte Stichprobe, in der für zwei einander ausschließende Ereignisse zwei unterschiedliche Häufigkeiten vorliegen, aus einer bestimmten Grundgesamtheit stammt, in der diese Ereignisse mit gegebenen Wahrscheinlichkeiten p und q auftauchen. Die Nullhypothese würde dann folgende Form haben:

$$H_0 \Rightarrow p_s \geq p_g$$

Die Alternativhypothese hätte die Form

$$H_1 \Rightarrow p_s < p_g$$

Hierbei bedeuten

p_s = Prozentsatz eines interessierenden Ereignisses in einer Stichprobe

p_g = Prozentsatz desselben interessierenden Ereignisses in einer Grundgesamtheit.

Dem Binomialtest liegt die Binomialverteilung zugrunde, deren Formel wir oben schon kennengelernt haben:

$$P_{n,p,k} = \binom{n}{k} p^k q^{n-k}$$

Sie bestimmt die Wahrscheinlichkeit P , daß ein Ereignis, für dessen Eintretenswahrscheinlichkeit p steht, in einer Stichprobe vom Umfang n k -mal auftritt. Also zum Beispiel, daß beim Wurf von $n=8$ Münzen $k=2$ mal der „Kopf“ oben liegt. Diese Wahrscheinlichkeit konnten wir mittels der Binomialverteilung darstellen, die Graphik zeigte eine ähnliche Gestalt wie die Normalverteilung, nur daß die Variable k nicht stetig war, sondern die diskreten Werte von $k=0$ bis $k = n$ durchlief.

Bilden wir nun für diese Variable die Summenfunktion, d.h. kumulieren wir von links nach rechts die Binomialverteilung auf, dann erhalten wir folgende Funktion:

$$P_{0-u} = \sum_{k=0}^u \binom{n}{k} p^k q^{n-k}$$

Hierbei bedeuten

n = Größe der Stichprobe

u = Häufigkeit eines Merkmals, das nur zwei Ausprägungen haben darf, in der Stichprobe

k = Laufvariable, die von 0 bis u läuft.

p = Wahrscheinlichkeit des Eintretens dieses Merkmals in einer Grundgesamtheit

q = 1-p

Diese Funktion bestimmt die Summe der Wahrscheinlichkeiten P, daß ein Ereignis, für dessen Eintretenswahrscheinlichkeit p steht, in einer Stichprobe vom Umfang n nullmal oder einmal oder zweimal oder ...oder u-mal auftritt. Also zum Beispiel, daß beim Wurf von n=8 Münzen der Kopf nullmal oder 1mal oder 2mal oder...oder 4mal oben liegt.

Diese Summe von Wahrscheinlichkeiten P_{0-u} kann minimal 0 und maximal 1 (entsprechend 100 %) sein, da u maximal den Wert von n annehmen kann.

Für die Prüfung unserer Nullhypothese benutzen wir nun diese Summenfunktion als Lieferanten unseres Kriteriums c, das den Annahme- und Ablehnungsbereich der Nullhypothese in unserer Testfunktion voneinander trennt. Wir legen c beispielsweise bei 0,05 bzw. 5 % fest und berechnen P_{0-u} . Ist P_{0-u} kleiner als c, dann lehnen wir die Nullhypothese ab, ist P_{0-u} größer als c, dann behalten wir die Nullhypothese bei.

Rechenbeispiel:

In einer Untersuchung über Drogengebrauch geben 15 Personen aus einer befragten Gruppe (n = 15) an, keine Drogen zu benutzen. Man vermutet aber, daß die Stichprobe einer Grundgesamtheit entnommen wurde, in der 50 % Drogen gebrauchen. Nun ist die Frage, ob es Zufall ist, wenn alle, alle bis auf einen, alle bis auf zwei, alle bis auf drei usw. den Drogengebrauch verneinen. Kodieren wir das Zugeben des Drogengebrauchs mit 1, das Verneinen mit 2, dann können wir mittels der Berechnung der kumulierten Wahrscheinlichkeiten ablesen, wann die Nullhypothese beibehalten werden kann, bei 5 % Irrtumswahrscheinlichkeit und einseitigem Test, daß eine Stichprobe aus einer Grundgesamtheit mit 50 % Drogenbenutzern gezogen wurde.

In der folgenden Tabelle steht die Gruppe 1 für die den Drogengebrauch zugebenden Personen, die Gruppe 2 sind diejenigen, die den Drogengebrauch verneinen. Untereinander stehen in mehreren Zeilen zum Vergleich Stichproben, die kein Mitglied der Gruppe 1 enthalten, 1 Mitglied enthalten, 2 Mitglieder enthalten etc...:

Test auf Binomialverteilung

	N	Beobachteter Anteil	Testanteil	Kumulierte Punktwahrscheinlichkeit	Punkt-Wahrscheinlichkeit
Gruppe 1	0	,00	,50	,000	,000
Gruppe 2	15	1,00			
Gesamt	15	1,00			
Gruppe 1	1	,07	,50	,000	,000
Gruppe 2	14	,93			
Gesamt	15	1,00			
Gruppe 1	2	,13	,50	,004	,003
Gruppe 2	13	,87			
Gesamt	15	1,00			
Gruppe 1	3	,20	,50	,018	,014
Gruppe 2	12	,80			
Gesamt	15	1,00			
Gruppe 1	4	,27	,50	,059	,042
Gruppe 2	11	,73			
Gesamt	15	1,00			

In der letzten Spalte ist die Wahrscheinlichkeit (Punktwahrscheinlichkeit) dafür berechnet worden, daß exakt kein Mitglied der Gruppe 1 enthalten, 1 Mitglied enthalten, 2 Mitglieder enthalten etc...:

In der vorletzten Spalte ist diese Wahrscheinlichkeit aufkumuliert. Nimmt man diese aufkumulierte Wahrscheinlichkeit als Kriterium für die Entscheidung über die Nullhypothese, dann kann aus der Tabelle folgendes abgelesen werden: Sind 4 Mitglieder der Gruppe 1 in der Stichprobe vorhanden, dann kann die Nullhypothese bei einseitigem Test und bei 5 % Irrtumswahrscheinlichkeit schon beibehalten werden, da die kumulierte Punktwahrscheinlichkeit P_{0-U} auf 0,059 angewachsen ist, also größer als 0,05 geworden ist (unser α wäre bei 5 % Irrtumswahrscheinlichkeit gleich 0,05). Sind nur 3 Mitglieder der Gruppe 1 vorhanden, müßte sie abgelehnt werden, da $P_{0-U} = 0,018 < 0,05$ wäre.

Das praktische Berechnen der kumulierten Punktwahrscheinlichkeit kann sehr mühsam sein, wenn die Stichproben groß werden. Daher hilft man sich bei großen Stichproben mit einer anderen Rechenstrategie. Wenn (nach einem Kriterium von Sidney Siegel) $npq > 9$, rechnet man den z-Test anstelle des Binomialtests. z wird hierbei wie folgt bestimmt:

$$z = \frac{(x - np) + 0,5}{\sqrt{npq}}$$

Hierbei ist z eine Entfernung vom Nullpunkt der standardisierten Normalverteilung, der ein bestimmtes Integral der standardisierten Normalverteilung entspricht. In der Formel steht das darin enthaltene x für die Anzahl, mit der das seltenere Ereignis von zwei möglichen Ereignissen in der Stichprobe aufgetreten ist. Die Subtraktion oder Addition des Wertes 0,5 von x erfolgt nach der Regel:

wenn $x < np$ wird 0,5 addiert
wenn $x > np$ wird 0,5 subtrahiert.

Diese Operation dient der sogenannten "Kontinuitätskorrektur".

Hierzu ein weiterer praktischer Fall:

In einer Stichprobe vom Umfang 80 (N) aus Studierenden der Betriebswirtschaft wurden 26 Frauen gezählt. Man weiß aus anderen Quellen, daß in der Grundgesamtheit der Anteil der Frauen 27% beträgt. Die Frage lautet: Wie gut bildet die Stichprobe die Grundgesamtheit ab, ist sie in bezug auf das Geschlecht für die Grundgesamtheit „repräsentativ“? Durch Überprüfung der Repräsentativität will man ausschließen, daß aufgrund der Stichprobe unzulässige Aussagen über die Grundgesamtheit gemacht werden. Rechnen wir das o.a. Beispiel durch, dann ergibt sich:

$$n \cdot q \cdot p = 80 \cdot 0,27 \cdot 0,73 = 15,77 > 9 \quad (\text{Siegel-Kriterium}).$$

Wir dürfen hiernach den z-Test rechnen; da: $x=26 > 80 \cdot 0,27 = 21,60$, ist in der Formel 0,5 von x zu subtrahieren:

$$z = \frac{(26 \pm 0,5) - 80 \cdot 0,27}{\sqrt{80 \cdot 0,27 \cdot 0,73}} = \frac{25,5 - 21,60}{3,971} = 0,9821$$

Für dieses z können wir feststellen, daß zwischen -0,9821 und + 0,9821 67,4% der Fläche unter der Dichtefunktion der Standardnormalverteilung liegen. Außerhalb dieses Bereichs liegen also 32,6% der Fläche, was der Größe des Ablehnungsbereichs der Nullhypothese bei zweiseitigem Test entspräche. Soll dieser aber nur - beispielsweise - 5% betragen, dann müßten wir in diesem Fall die Nullhypothese beibehalten, daß unsere Stichprobe aus der fraglichen Grundgesamtheit entnommen worden sein kann.

7.3.7. Chiquadrat-Test

7.3.7.1. Allgemeines

Ausgehend von der sogenannten "Pearsonschen Testfunktion",

$$Chi^2 = \sum_{i=1}^r \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^r \frac{(f_b - f_e)^2}{f_e}$$

die für große n näherungsweise chiquadrat-verteilt ist, können verschiedene Anpassungstests formuliert werden, also Tests, die die Gestalt der Verteilungsfunktion einer Zufallsvariable betreffen und nicht den einen oder anderen Parameter der Zufallsvariable. In der o.a. Funktion bedeuten

y_i = Zufallsvariable, Anzahl von Versuchen (z.B. Würfeln),

bei denen das Ereignis A eintritt (z.B. die Zahl 4). y_i entspricht der beobachteten Häufigkeit eines Ereignisses f_b .

p_i = ist die Eintrittswahrscheinlichkeit für das Ereignis A_i .

n = Menge der Versuche in einem Experiment.

r = Menge verschiedener möglicher Ereignisse A_i

$np = f_e$ = erwartete Häufigkeit eines Ereignisses A

Für konkrete Versuche berechnen wir zunächst für jedes Ereignis den Erwartungswert aus $n \cdot p_i$. Dies sind die "erwarteten Häufigkeiten".

Diese subtrahieren wir von den beobachteten Häufigkeiten, die für konkrete Untersuchungen an die Stelle der y_i in der o.a. Formel treten. Wir quadrieren die Differenz $(y_i - np_i)^2$ und dividieren diese durch $(n \cdot p_i)$, so daß wir schließlich aus den Chiquadratwerten der Einzelereignisse die Summe bilden können. Diese Summe ist der Wert der Testfunktion, den wir wiederum mit einem Kriteriumswert zu vergleichen haben, wenn wir über die Annahme oder Ablehnung einer Nullhypothese entscheiden.

7.3.7.2. Eindimensionales und zweidimensionales Chiquadrat-Design

Wie können wir bei Chiquadratverfahren die Nullhypothese formulieren ?

Hierzu gibt es zwei Möglichkeiten:

Für eine monovariante (eindimensionale) Häufigkeitsverteilung:

$H_0 \Rightarrow$ die Variable y ist (gleichverteilt, normalverteilt etc.) in der Grundgesamtheit

$H_1 \Rightarrow$ die Variable y ist nicht (gleichverteilt, normalverteilt etc.) in der Grundgesamtheit

Für eine kombinierte (zweidimensionale) Häufigkeitsverteilung der Variablen v und w :

$H_0 \Rightarrow$ die Variablen vw sind gleichverteilt in der Grundgesamtheit

$H_1 \Rightarrow$ die Variablen vw sind nicht gleichverteilt in der Grundgesamtheit

oder

$H_0 \Rightarrow$ die Variablen vw sind „randverteilt“ in der Grundgesamtheit

$H_1 \Rightarrow$ die Variablen vw sind nicht „randverteilt“ in der Grundgesamtheit

Bei Gleichverteilung sind die Erwartungswerte für jede Häufigkeit der Verteilung gleich groß.

Bei „Randverteilung“ unterscheiden sich die Erwartungswerte der verschiedenen kombinierten Häufigkeiten. Sie werden aus den Randsummen einer $n \times m$ -Tabelle ermittelt, und zwar so, daß pro Häufigkeit der folgende Ausdruck berechnet wird:

Erwartungswert pro Zelle = (Zeilensumme * Spaltensumme) / Tabellensumme

Wie bestimmen wir jetzt die Freiheitsgrade jener Chiquadratfunktion, aus der wir unser Kriterium für die Entscheidung über die Nullhypothese gewinnen ? Hier sind die folgenden zwei Fälle zu unterscheiden:

- Bei monovariaten Häufigkeitsverteilungen wird der Kriteriumswert für die Entscheidung über die Nullhypothese aus einer Chiquadratverteilung mit $k-1$ Freiheitsgraden gewonnen, wobei k die Menge der Kategorien ist, in die die Variable eingeteilt ist.
- Bei bivariaten Häufigkeitsverteilungen, die auch in Tabellenform dargestellt werden können, wird der Kriteriumswert für die Entscheidung über die Nullhypothese aus einer Chiquadratverteilung mit $(r-1) \cdot (c-1)$ Freiheitsgraden gewonnen, wobei r die Menge der Zeilen einer solchen Tabelle ist, c die Menge der Spalten.

Rechenbeispiele:

„Eindimensionales Chiquadrat“

Es wird 1026 mal gewürfelt. Die Nullhypothese ist: Die Augenzahl ist gleichverteilt. Erwartet werden daher für die verschiedenen Augenzahlen jeweils $n/k = 1026/6 = 171$ -maliges Auftreten. Die untenstehende Tabelle enthält neben den tatsächlich aufgetretenen Häufigkeiten der verschiedenen Augen noch die erwarteten Häufigkeiten und deren Differenz zu den erwarteten Häufigkeiten. Setzt man jetzt die Werte in die Testfunktionsformel ein, dann erhält man die unten abgedruckte Chiquadratsumme. Diese beträgt 10,76 und muß mit einem Kriteriumswert verglichen werden, der aus einer Chiquadratverteilung mit $k-1 = 6-1 = 5$ Freiheitsgraden gewonnen wird. Führen wir eine einseitigen Chiquadratstest mit einer Irrtumswahrscheinlichkeit von 5 %, dann ist c , der Kriteriumswert, gleich 11,7.

AUGEN

	Beobachtetes N	Erwartete Anzahl	Residuum
1,00	192	171,0	21,0
2,00	178	171,0	7,0
3,00	142	171,0	-29,0
4,00	158	171,0	-13,0
5,00	189	171,0	18,0
6,00	167	171,0	-4,0
Gesamt	1026		

Statistik für Test

	AUGEN
Chi-Quadrat ^a	10,760
df	5
Exakte Signifikanz	,056

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 171,0.

Da das berechnete Chiquadrat kleiner ist als der Kriteriumswert für 5% Irrtumswahrscheinlichkeit bei fünf Freiheitsgraden, behalten wir die Nullhypothese bei. Die Angabe der „exakten Signifikanz“ in der o.a. Tabelle entspricht der Größe des Integrals in einer Chiquadratverteilung mit fünf Freiheitsgraden von 10,76 bis unendlich.

„Zweidimensionales Chiquadrat“

In einer Untersuchung über die Einstellung von Menschen zu Hunden werden zwei Vorgaben gemacht:

- (a) Mein Hund soll schön und auffallend sein. Item-Stufen: Keinesfalls, eher nicht, vielleicht doch, eher ja.
 (b) Wenn ein Hund seine Aufgaben nicht erfüllt, sollte man sich einen anderen Hund anschaffen. Item-Stufen: Keinesfalls, eher nicht, vielleicht doch, eher ja.

Die wissenschaftliche Hypothese, die zu dieser kombinierten Auszählung (Bildung einer sogenannten Kreuztabelle) Anlaß gibt, lautet: Menschen, denen es vor allem auf das schöne Aussehen eines Hundes ankommt, sind eher bereit, den Hund abzuschaffen, wenn er seine Aufgaben nicht erfüllt, als Menschen, denen es nicht so sehr auf die Schönheit ankommt.

Aus dieser wissenschaftlichen Hypothese wird die statistische Hypothese entwickelt:

$H_0 \Rightarrow$ die Variablen *vw* sind „randverteilt“ in der Grundgesamtheit

$H_1 \Rightarrow$ die Variablen *vw* sind nicht „randverteilt“ in der Grundgesamtheit

Was heißt nun „randverteilt“? Randverteilung liegt vor, wenn alle Zellen-Häufigkeiten (in allen Zeilen und allen Spalten) im gleichen Verhältnis stehen wie die Randsummen der Tabelle. Ist dies der Fall, dann unterscheiden sich beobachtete und erwartete Häufigkeiten nicht. Es gibt dann keine Häufigkeiten, die auf Unter- oder Überrepräsentation bestimmter Wertekombinationen hindeuten.

Sind die Zellenhäufigkeiten aber anders verteilt (in bestimmten Zeilen oder Spalten) als die entsprechenden Randsummen, dann liegen in diesen Zellen Über- oder Unterrepräsentationen von Wertekombinationen vor, die auf einen statistischen Zusammenhang zweier Variablen hindeuten. Für das oben aufgeführte Beispiel müßten also Menschen, die auf einen schönen Hund Wert legen, signifikant häufiger als andere Menschen bereit sein, sich einen anderen Hund anzuschaffen, wenn ihr derzeitiger Hund seine Aufgaben nicht erfüllt.

Hier nun die empirischen Daten aus einer Untersuchung des Autors:

Mein Hund soll schön und auffallend sein / Wenn der Hund seine Aufgaben nicht erfüllt: Anderen anschaffen

			Anderen anschaffen				Gesamt
			keines- falls	eher nicht	vielleicht doch	eher ja	
Schöner Hund	keines- falls	Anzahl	86	20	6	10	122
		Erwartete Anzahl	71,6	27,3	11,1	12,0	122,0
		Residuen	14,4	-7,3	-5,1	-2,0	
	eher nicht	Anzahl	120	51	10	15	196
		Erwartete Anzahl	115,1	43,8	17,8	19,3	196,0
		Residuen	4,9	7,2	-7,8	-4,3	
	viel- leicht doch	Anzahl	74	33	18	13	138
		Erwartete Anzahl	81,0	30,8	12,5	13,6	138,0
		Residuen	-7,0	2,2	5,5	-,6	
	eher ja	Anzahl	30	14	14	14	72
		Erwartete Anzahl	42,3	16,1	6,5	7,1	72,0
		Residuen	-12,3	-2,1	7,5	6,9	
Gesamt	Anzahl	310	118	48	52	528	
	Erwartete Anzahl	310,0	118,0	48,0	52,0	528,0	

Chi-Quadrat-Tests

	Wert	df
Chi-Quadrat nach Pearson	35,491 ^a	9
Anzahl der gültigen Fälle	528	

a. 0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 6,55.

Die obige Kreuztabelle enthält in ihren Zellen die beobachteten Häufigkeiten, die erwarteten Häufigkeiten und deren Differenz, die sog. Residuen. Würden alle beobachteten Häufigkeiten randverteilt sein, würden sie den erwarteten entsprechen und die Residuen wären gleich null. Weichen die beobachteten Häufigkeiten positiv von den erwarteten ab, dann liegt eine Überrepräsentation der entsprechenden Wertekombination vor, weichen die beobachteten Häufigkeiten negativ von den erwarteten ab, dann liegt eine Unterrepräsentation der entsprechenden Wertekombination vor.

Wie eine Inspektion der Tabelle zeigt, sind vor allem die Zellen der sogenannten Hauptdiagonale (von oben links nach unten rechts) mit Häufigkeiten versehen, die auf Überrepräsentation hindeuten, Zellen die oben rechts und unten links von der Hauptdiagonale liegen, weisen eher Unterrepräsentationen auf. Dieses Phänomen kann zufällig auftreten oder auf einen statistischen Zusammenhang hindeuten. Ob ein solcher Zusammenhang angenommen werden kann, hängt davon ab, ob unsere Nullhypothese („Randverteilung“) beibehalten werden kann oder ob sie abgelehnt werden muß. Das Kriterium hierzu gewinnen wir aus einer Chiquadratverteilung mit $(r-1)(c-1)$ Freiheitsgraden. Bei 4 Zeilen (r) und 4 Spalten (c) ist df (Menge der Freiheitsgrade) gleich $3 \cdot 3 = 9$. Testen wir unsere Hypothese einseitig bei 1 % Irrtumswahrscheinlichkeit wäre unser Kriteriums-Chiquadrat-Wert gleich 21,67.

Aus der Tabelle berechnen wir nach der Formel

$$Chi^2 = \sum_{i=1}^r \frac{(f_b - f_e)^2}{f_e}$$

eine Chiquadratsumme für die ganze Tabelle von 35,491.

Da unser Kriteriums-Chiquadrat (21,67) kleiner als das berechnete ist, lehnen wir bei einseitigem Test mit 1 % Irrtumswahrscheinlichkeit die Nullhypothese ab.

Dies heißt zunächst nur, daß unsere Stichprobe nicht zufällig einer Grundgesamtheit entnommen worden sein kann, in der die Zellenverteilungen sich nach den Randsummenverteilungen richten. Es müssen also Über- und Unterrepräsentationen in der Tabelle vorkommen. Welche sind das ? Dadurch, daß die positiven Residuen

(Überrepräsentationen) überwiegend auf der Hauptdiagonale liegen, kommen wir zu dem Schluß, daß die wachsende Priorität für die Schönheit des Hundes mit einer wachsenden Bereitschaft einhergeht, ihn gegebenenfalls zu ersetzen.

Generell kann nun gesagt werden:

Es läßt sich für alle χ^2 -Verteilungen (entweder mit $k-1$ Freiheitsgraden oder mit $(r-1)*(c-1)$ Freiheitsgraden) jener kritische Wert c bestimmen, der ein Integral der entsprechenden χ^2 -Verteilung rechts begrenzt, welches einer bestimmten Irrtumswahrscheinlichkeit α entspricht. Liegt unser berechnetes χ^2 oberhalb dieses kritischen Wertes, dann müssen wir die Null-Hypothese ablehnen.

Sonderfall:

Für den eindimensionalen Test von Übereinstimmung einer Verteilung mit einer von unbekannten Parametern (z.B. von μ und σ) abhängigen Funktion (z.B. der Normalverteilung) verringert sich die Menge der Freiheitsgrade, die sonst $df = r-1$ ist, um die Menge der Parameter m , so daß für diesen Fall

$$df = r - 1 - m \quad \text{beträgt.}$$

Testen wir also die Übereinstimmung mit einer Normalverteilung, dann müssen wir für die Berechnung der p_i und der f_i Meßwertklassen bilden; die Freiheitsgrade für den Test wären dann hiernach gleich "Menge der Meßwertklassen minus 3". (vgl. Bosch, a.a.O., S. 103 ff., hierzu siehe S. 108).

7.3.7.3. Standardisierte, korrigierte Chiquadratresiduen

Will man bei einem zweidimensionalen Chiquadrat-Test den Beitrag einzelner Zellen zum Tabellen-Chiquadrat ermitteln, dann bieten sich hierzu mehrere Möglichkeiten an:

Beispiel:

Die Differenz von beobachteter Häufigkeit und erwarteter Häufigkeit ergibt die Residuen. Sind diese groß, dann ist der Beitrag der Zelle zum Tabellen-Chiquadrat groß.

Diese Chiquadrat-Residuen kann man „standardisieren“, indem man sie durch die Wurzel aus den erwarteten Häufigkeiten teilt. Auch hier ist der Beitrag zum Tabellen-Chiquadrat erkennbar.

Die standardisierten Residuen kann man durch eine weitere Operation in die sogenannten „korrigierten“ standardisierten Residuen verwandeln:

$$\frac{f_b - f_e}{\sqrt{f_e - \left(1 - \frac{Z}{n}\right)\left(1 - \frac{S}{n}\right)}}$$

Es bedeuten: Z = Zeilensumme für die jeweilige Zelle
 S = Spaltensumme für die jeweilige Zelle
 n = Stichprobengröße (Tabellensumme).

Diese korrigierten, standardisierten Residuen sind z-verteilt, d.h. man kann aus ihnen ablesen, ob der Beitrag einer Zelle zum Tabellen-Chiquadrat „signifikant“ ist. Das Auftreten von z-Werten oberhalb bestimmter Schwellenwerte ist umso weniger wahrscheinlich, je größer diese Schwellenwerte sind. Sinkt diese Wahrscheinlichkeit unter bestimmte kritische Werte ab, z. B. auf 1 % oder 0,1%, dann kann man dem entsprechenden z bzw. dem standardisierten, korrigierten Chiquadrat-Residuum schon die Eigenschaft zusprechen, eine „signifikante“ Über- oder „Unterrepräsentation“ der jeweiligen Kombination von Werten für eine bestimmte Zelle in einer Kreuztabelle zu signalisieren. So kann man sowohl über die Signifikanz der

Beziehung zwischen den Variablen in der Tabelle als auch über die Bedeutung der einzelnen Zellen hierbei eine Aussage machen.

Versuchen Sie, dies anhand der folgenden Daten nachzuvollziehen: Die Daten entstammen einer Befragung von Hundehaltern, die entscheiden sollten, ob ihr Hund eher Wach- und Schutzhund sein sollte oder nicht. Auffällig sind beispielsweise die standardisierten, korrigierten Chiquadratresiduen bei der Ausprägung „Auf jeden Fall“. Sie liegen bei

- 4,3 (für die Frauen) und + 4,3 (für die Männer).

Die Wahrscheinlichkeit, daß oberhalb von z-Werten von 4,3 noch weitere z-Werte auftreten, ist mit $p = 0,0000008540$ annähernd Null. Analog wäre das Auftreten von standardisierten, korrigierten Chiquadratresiduen oberhalb von 4,2 ähnlich wenig wahrscheinlich. Wir deuten daher solche Werte als Zeichen für signifikante Über- oder Unterrepräsentation der entsprechenden Kombination von Ausprägungen in der Tabelle:

Kreuztabellen

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Geschl. d. Befragten * Mein H. soll Wach- und Schutzhund sein	961	96,9%	31	3,1%	992	100,0%

Geschlecht der Befragten * Mein H. soll Wach- und Schutzhund sein, Kreuztabelle

		Mein H. soll Wach- und Schutzhund sein					Gesamt
		Auf keinen Fall	Eher nicht	Teils-teils	eher ja	Auf jeden Fall	
Geschl. d. weiblich Befragten	Anzahl	72	212	143	114	40	581
	Erwartete Anzahl	70,1	188,6	149,9	112,5	59,9	581,0
	Standardisierte Residu	,2	1,7	-,6	,1	-2,6	
	Korrigierte Residuen	,4	3,3	-1,0	,3	-4,3	
männlich	Anzahl	44	100	105	72	59	380
	Erwartete Anzahl	45,9	123,4	98,1	73,5	39,1	380,0
	Standardisierte Residu	-,3	-2,1	,7	-,2	3,2	
	Korrigierte Residuen	-,4	-3,3	1,0	-,3	4,3	
Gesamt	Anzahl	116	312	248	186	99	961
	Erwartete Anzahl	116,0	312,0	248,0	186,0	99,0	961,0

Chi-Quadrat-Tests

	Wert	df	Asymptotisch e Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	24,968 ^a	4	,000
Anzahl der gültigen Fälle	961	1	,000

a. 0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 39,15.

Symmetrische Maße

	Wert	Näherung sweise Signifikanz
Nominal-Maß Cramer-V	,161	,000
Anzahl der gültigen Fälle	961	

- a. Die Null-Hyphothese wird nicht angenommen.
 b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

7.3.8. Mann-Whitney U-Test

Beispiel: Gegeben seien zwei unabhängige Stichproben A und B von ordinal skalierten Meßwerten. Meßwerte aus der Stichprobe A heißen a_i , Meßwerte aus der Stichprobe B sind die b_i . Die Nullhypothese ist, daß die beiden Stichproben aus identisch verteilten Grundgesamtheiten stammen (wobei die Form der Verteilung gleichgültig ist). Die Alternativhypothese ist, daß die Elemente der Stichprobe A größere Werte annehmen als die aus der Stichprobe B.

$$H_0 \Rightarrow p(a > b) = 0,5$$

$$H_1 \Rightarrow p(a > b) > 0,5$$

Gesucht ist nun ein empirischer Wert einer Funktion, die einen kritischen Wert für die Entscheidung der Nullhypothese enthält. Dieses ist die sog. U-Funktion. Der empirische U-Wert wird wie folgt berechnet:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

Hierbei ist n_1 der Umfang der 1. Stichprobe, n_2 der Umfang der zweiten Stichprobe und R_1 die Summe der Ränge in der 1. Stichprobe.

Rechenbeispiel: Zwei Gymnasien (1 und 2) werden aufgrund kleiner Stichproben verglichen in bezug auf die Lösungsgeschwindigkeit bei einer Mathematikaufgabe in Minuten. Aus dem ersten Gymnasium kommen 5 Schüler, aus dem zweiten vier. Deren Lösungsdauer in Minuten werden in der folgenden Tabelle zusammengefügt und in eine Rangordnung gebracht. Die Ränge derjenigen Schüler, die im 1. Gymnasium sind, werden addiert. Das Ergebnis ist $R_1 = 26$.

Punktesumme im Abiturzeugnis	Gymnasium	Rang	Menge der Meßwerte aus Gymnasium 1, die besser sind als bestimmte Meßwerte aus Gymnasium 2
45,00	1,00	1,000	
51,00	2,00	2,000	1
53,00	2,00	3,000	1
64,00	1,00	4,000	
70,00	2,00	5,000	2
75,00	1,00	6,000	
78,00	1,00	7,000	
82,00	1,00	8,000	
110,00	2,00	9,000	5
			Summe = 9 = U.

Nun setzen wir die Werte in die Formel ein, dann ergibt sich

$$U = 5 * 4 + \frac{5(5+1)}{2} - 26 = 9$$

Es gilt nun:

$$U = n_1 n_2 - U'$$

Eingesetzt ergibt sich:

$$9 = 5 * 4 - U'$$

$U' = 11$ ist daher größer als $U=9$.

Den kleineren von beiden Werten U und U' nimmt man als empirischen Testwert.

Dieser ist hier die Menge der Meßwerte aus Gymnasium 1, die jeweils besser sind als bestimmte Meßwerte aus Gymnasium 2.

Diese empirisch ermittelte Zahl 9 wird nun mit der U-Tabelle verglichen. Wäre U' kleiner als U , würden wir U' wählen.

Für die Prüfung der Nullhypothese suchen wir in der U-Tabelle die Teiltabelle für die größere Stichprobe ($n=5$). Wir suchen jene Wahrscheinlichkeit der U-Funktion, die unserem $U = 9$ und der Stichprobengröße der kleineren Stichprobe ($n=4$) entspricht. Dort finden wir (Siegel, S. 271) die Größe $p = 0.452$. Da diese Zahl größer ist als ein mögliches Niveau der Irrtumswahrscheinlichkeit von $\alpha = 0,05$, entscheiden wir uns, die Nullhypothese beizubehalten. Wie Siegel ausführt, sind unterschiedliche Verfahren der Berechnung des kritischen Wertes von U maßgeblich, je nachdem, wie groß die größere der beiden Stichproben ist. Wächst deren Umfang über 20 an, kann man mittels eines z-Tests entscheiden, ob die Nullhypothese beibehalten werden kann. Hier gilt:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Dieses z ist – wie oben ausgeführt – standardnormalverteilt mit dem arithmetischen Mittel 0 und der Standardabweichung 1. Diese Verteilung wird dann benutzt wie beim z-Test.

Tauchen ties auf (siehe hierzu Kendalls τ_b), dann wird den entsprechenden Rängen ihr durchschnittlicher Rang zugewiesen und z wird nach einer korrigierten Formel berechnet, die Siegel auf S. 125 referiert.

7.4. Tabellarische Übersicht über kombinatorische Probleme und die Prüfverteilungen

Übersicht über Fragestellungen und Lösungen kombinatorischer Probleme:

Permutationen: (Menge der möglichen Reihenfolgen aller Elementen einer Menge)		Kombinationen: Zusammensetzung einer Teilmenge, die aus einer Gesamtmenge von verschiedenen Elementen ausgewählt wurde			
Alle Elemente sind verschieden	Einige Elemente sind identisch	Mit Berücksichtigung der Reihenfolge der Elemente		Ohne Berücksichtigung der Reihenfolge der Elemente	
		Mit Zurücklegen von Elementen bei der Auswahl	Ohne Zurücklegen der Elemente bei der Auswahl	Mit Zurücklegen von Elementen bei der Auswahl	Ohne Zurücklegen der Elemente bei der Auswahl
$n!$	$\left(\frac{n!}{n_1!n_2!\dots n_u!} \right)$	n^k	$\frac{n!}{(n-k)!}$	$\binom{n+k-1}{k}$	$\frac{n!}{k!(n-k)!} = \binom{n}{k}$
Beispiel: Als Elemente existieren die Buchstaben: S A B I N E. Wieviele verschiedene Buchstabenfolgen kann man aus diesen Buchstaben herstellen ?	Beispiel: Als Elemente existieren die Buchstaben: O T T O K A R. Wieviele verschiedene Buchstabenfolgen kann man aus diesen Buchstaben herstellen ? (Merke: Vertauschungen identischer Buchstaben sind nicht unterscheidbar)	Beispiel: Aus den Buchstaben: T U B C H A M P I O N werden 3 Buchstaben ausgewählt, aber nach der Auswahl zurückgelegt, bei der Menge der Kombinationen sei die Reihenfolge bedeutsam: Wieviele unterscheidbare Kombinationen gibt es ?	Beispiel: Aus den Buchstaben: T U B C H A M P I O N werden 3 Buchstaben ausgewählt, und nach der Auswahl nicht zurückgelegt, bei der Menge der Kombinationen ist die Reihenfolge bedeutsam: Wieviele unterscheidbare Kombinationen gibt es ?	Beispiel: Aus den Buchstaben: T U B C H A M P I O N werden 3 Buchstaben ausgewählt, aber nach der Auswahl zurückgelegt, bei der Menge der Kombinationen ist die Reihenfolge nicht bedeutsam: Wieviele unterscheidbare Kombinationen gibt es ?	Beispiel: Aus den Buchstaben: T U B C H A M P I O N werden 3 Buchstaben ausgewählt und nach der Auswahl nicht zurückgelegt, bei der Menge der Kombinationen ist die Reihenfolge nicht bedeutsam: Wieviele unterscheidbare Kombinationen gibt es ?
Lösung:	Lösung:	Lösung:	Lösung:	Lösung:	Lösung:
720	1260	1331	990	286	165

Verteilungsfunktionen von Wahrscheinlichkeiten für das Eintreten bestimmter Ereignisse: Entweder für diskrete Variablen (z. B. k in der Binomialverteilung) oder für stetige Variablen (z.B. Wahrscheinlichkeitsdichte der Normalverteilung etc.)			
Name/ Erklärung der Symbole	Formel für die Wahrscheinlichkeit/bzw. die Wahrscheinlichkeitsdichte/Anwendungsfälle:	arithmetisches Mittel	Standardabweichung bzw. Varianz/ Erläuterung
Gleichverteilung x: stetige Variable, Meßwert. a: Untergrenze b:Obergrenze eines Intervalls	$f(x) = 1/(b-a)$ für das Intervall a bis b	$\mu = (a + b)/2$	$\sigma = \sqrt{\frac{(b-a)^2}{12}}$
Binomialverteilung n: Umfang einer Menge k: Umfang einer ausgewählten Teilmenge u= maximales k p: Wahrscheinlichkeit eines Ereignisses von zwei möglichen Ereignissen (z.B. weiblich/männlich) q = 1-p: Wahrscheinlichkeit des komplementären Ereignisses	$P_{n,p,k} = \binom{n}{k} p^k q^{n-k}$; $P_{0-u} = \sum_{k=0}^u \binom{n}{k} p^k q^{n-k}$ Anwendungsfall: Binomialtest: Stammt eine Stichprobe (n=30) mit bis zu 2 Kranken aus einer Grundgesamtheit mit 20% Kranken, bei einer Irrtumswahrscheinlichkeit von 2 %?	$\mu = n \cdot p$	$\sigma = \sqrt{npq}$; Die Verteilung ist die Basis des Binomialtests
Poisson-Verteilung: Symbole wie bei Binomialverteilung, e: Eulersche Zahl 2,718281828...	$f(k) = \frac{(np)^k}{k!} e^{-np}$ Anwendungsfall: Es geht um die Wahrscheinlichkeit seltener Ereignisse: Wenn z.B. „die Wahrscheinlichkeit, in einer Klausur zu versagen gleich $p = 0,01$ ist, ist die Wahrscheinlichkeit berechenbar, unter $n=40$ Klausurteilnehmern $k = 0, 1, 2, 3, 4$ „Ausfälle“ zu haben. Sie nimmt folgende Werte an: 0,67; 0,26; 0,053; 0,007.	$\mu = np$	$\sigma = \sqrt{np}$; Die Verteilung geht aus der Binomialverteilung hervor, wenn p gegen 0 geht und n gegen ∞ , und wenn np eine endliche und konstante Größe bleibt.
Normalverteilung – allgemeine Form und Standardform μ : arithmetisches Mittel der Grundgesamtheit \bar{x} : arithmetisches Mittel der Stichprobe σ : Standardabweichung der Grundgesamtheit s: Standardabweichung der Stichprobe π : Kreiskonstante: 3,141592654... z: Standardisierter Meßwert	$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$ Wenn $\mu = 0$ und $\sigma = 1$ sind, ergibt sich: $f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$ Anwendungsfall: Stammt eine Stichprobe mit $n = 49$, $s = 9,3$ und $\bar{x} = 75$ mit 1 % Irrtumswahrscheinlichkeit aus einer normalverteilten Grundgesamtheit mit $\mu = 72$, $\sigma = 9$? Lösung: $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = 2,333$ $p = \alpha = 0,0098 < 0,01$. Die Nullhypothese $\bar{x} = \mu$ wird abgelehnt, die Alternativhypothese $\bar{x} > \mu$ angenommen.	μ ist definiert zwischen $-\infty$ und $+\infty$. Für die Standardnormalverteilung ist $\mu = 0$.	σ definiert zwischen 0 und $+\infty$. Die Verteilung geht aus der Binomialverteilung hervor, wenn n gegen ∞ geht, und $npq > 9$ ist. Sie ist die Basis des z-Tests . Jede beliebige Normalverteilung läßt sich durch Standardisierung der Meßwerte in z-Werte $z = \frac{x_i - \bar{x}}{s}$ umrechnen. Es entsteht dann die Standardnormalverteilung $f(z)$. Bei ihr ist $\sigma = 1$.

<p>χ^2-Verteilung</p> $\chi^2 = \sum_{i=1}^{df} z_i^2$ $\chi^2 = \sum_{i=1}^k \frac{(f_b - f_e)^2}{f_e}$ <p>χ^2: stetige Variable auf der Abszisse der χ^2-Verteilung df: Freiheitsgrad. f_b = beobachtete Häufigkeit f_e = erwartete Häufigkeit Γ Zeichen für die Gammafunktion</p>	$f(\chi^2) = \frac{e^{-\frac{\chi^2}{2}} * \left(\chi^2\right)^{\left(\frac{df-1}{2}\right)}}{2^{\frac{df}{2}} * \Gamma\left(\frac{df}{2}\right)}$ <p>Anwendungsfälle: Eindimensionales χ^2: Die Echtheit eines Würfels wird überprüft. df = k-1. Zweidimensionales χ^2: Der Zusammenhang zweier nominalskalierter Variablen, der in einer Kreuztabelle mit r Zeilen und c Spalten dargestellt ist, wird überprüft (df = (r-1)*(c-1)).</p>	<p>$\mu = df$; df = (r-1)* (c-1)</p>	<p>$\sigma = \sqrt{2 * df}$; Die Verteilung geht aus der Addition von quadrierten z-Verteilungen hervor, wobei die Menge der Freiheitsgrade df der Menge der Summanden entspricht. Sie ist die Basis des χ^2-Tests. Die sog. Gamma-Funktion liefert $\Gamma(n) = (n-1)!$ für alle natürlichen Zahlen. Die Gammafunktion ist jedoch darüber hinaus die stetige Erweiterung der Fakultät für nichtganzzahlige Werte.</p>
<p>t-Verteilung</p> $t_{df} = \frac{z}{\sqrt{\frac{\chi_{df}^2}{df}}}$ <p>t: stetige Variable auf der Abszisse der t-Verteilung</p>	$f(t_{df}) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\pi} * \Gamma\left(\frac{df}{2}\right) * \left(1 + \frac{t^2}{df}\right)^{\frac{df+1}{2}}}$ <p>Anwendungsfälle: a) Ist die Differenz der Mittelwertes einer kleinen Stichprobe (n<30) und einer normalverteilten Grundgesamtheit mit bekannter Varianz signifikant? (df = n-1) b) Ist die Differenz der Mittelwerte zweier (b1) unabhängiger (df = n₁ + n₂ - 2) bzw. (b2) zweier abhängiger (df = n₁ - 1) Stichproben mit homogenen Varianzen „signifikant“. c) Ist ein Korrelationskoeffizient r aus einer Stichprobe mit dem Umfang n signifikant? (df = n - 2).</p>	<p>$\mu = 0$</p>	<p>für df > 2 ist $\sigma = \sqrt{\frac{df}{df-2}}$ Die Verteilung geht aus der Division einer z-Verteilung durch die Wurzel aus einer χ^2-Verteilung, deren Werte ihrerseits durch die Menge der Freiheitsgrade geteilt werden, hervor. Sie ist die Basis verschiedener t-Tests.</p>
<p>F-Verteilung</p> <p>F: stetige Variable auf der Abszisse der F-Verteilung</p> $F_{m,n} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$ <p>m = df₁ n = df₂</p>	$f(F_{m,n}) = \frac{\Gamma\left(\frac{m+n}{2}\right) \left(\frac{m}{n}\right)^{\frac{m}{2}} * F^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right) \left(1 + \frac{m}{n} (F)\right)^{\frac{m+n}{2}}}$ <p>Anwendungsfälle: a) Homogenitätsprüfung für die Varianzen zweier unabhängiger Stichproben beim t-Test. b) Verschiedene Formen der Varianzanalyse. Hier geht es z.B. um die Frage, ob die Mittelwerte mehrerer unabhängiger Stichproben, die verschiedenen Ausprägungen einer nominalskalierten Variable zugehören, in mindestens einem Fall signifikant verschieden sind (einfaktorielle Varianzanalyse).</p>	<p>für n > 2 $\mu = \frac{m}{n-2}$</p>	<p>für m > 5 $\sigma = \sqrt{\frac{2m^2(m+n-2)}{n(m-4)(m-2)^2}}$ Die Verteilung geht aus der Division zweier χ^2-Verteilungen hervor, die ihrerseits durch die Menge ihrer Freiheitsgrade geteilt wurden.</p>

8. Bivariate Verteilungen - Zusammenhangsdarstellungen und -maße

Erst wenn mindestens zwei Merkmale gleichzeitig an einer Erhebungseinheit festgestellt werden, können Zusammenhänge von Merkmalen beschrieben werden.

Der Begriff "**Zusammenhang**" hat hierbei eine spezielle Bedeutung. Er besagt, daß die Kenntnis der Ausprägung eines Merkmals eine **Prognose** der Ausprägung eines anderen Merkmals zuläßt.

Dabei ist diese Prognose mehr oder minder genau. **Statistische Maßzahlen (Koeffizienten), die die Genauigkeit dieser Prognose beschreiben, sind in der Regel zwischen 0 und 1 definiert.** Ist der Genauigkeitsgrad gering, geht der Koeffizient gegen 0, ist er groß, geht der Koeffizient gegen Eins.

Eine spezielle Problematik des Ansatzes für eine solche Prognose ist durch das Skalierungsniveau der Daten gegeben. Bei Daten auf dem Nominalniveau kann ich nur Häufigkeiten für verschiedene Kategorien vorhersagen, bei Daten auf dem Ordinal-Niveau nur Häufigkeiten für Ränge; erst bei Daten auf dem Intervall- und Ratio-Niveau kann man Funktionswerte und Abweichungskorridore für solche Funktionen beschreiben. Dabei kann es sich um lineare und nicht-lineare Funktionen mit einer abhängigen und einer unabhängigen Variable handeln oder um Funktionen mit mehreren unabhängigen Variablen.

Die folgenden Ausführungen geben eine Auswahl aus den Möglichkeiten der Darstellung statistischer Zusammenhänge wieder.

9. Korrelation und Regression

9.1 Zwei intervallskalierte Merkmale

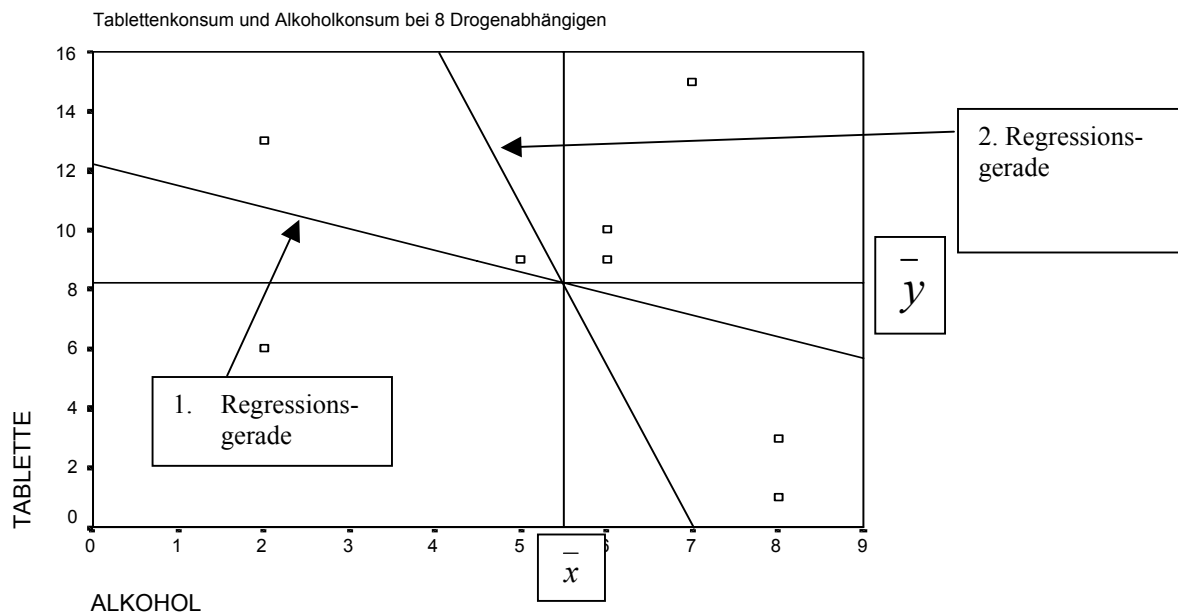
9.1.1. Die Regressionsgeraden und die Regressions-Koeffizienten

Wir gehen von folgendem praktischen Beispiel aus:

In einer empirischen Untersuchung von Drogenabhängigen stellte sich folgender Konsum von Alkohol und Tabletten (Beruhigungsmittel) heraus:

	x_i	y_i
	Alkohol in cl pro Tag	Beruhigungstabletten pro Tag
Name		
Paul	5	9
Peter	2	13
Jutta	8	1
Carla	6	9
Annette	6	10
Eduard	8	3
Siegfried	7	15
Urs	2	6

Stellen wir diese Verteilung in einem y/x-Achsenkreuz graphisch dar, dann ergibt sich das folgende Bild:



Die Wertepaare sind in dieser Graphik, die auch „Scatterplot“ genannt wird, als kleine Quadrate eingezeichnet. Die senkrechte Linie liegt dort, wo bei dem Alkoholgebrauch pro Tag das arithmetische Mittel (5,5) liegt, die waagerechte Linie liegt dort, wo beim Tablettengebrauch das arithmetische Mittel (8,25) liegt. Durch den Kreuzungspunkt beider Linien verlaufen die sogenannten „Regressionsgeraden“, auf die wir unten noch eingehen.

Wie wir später sehen werden, können wir aufgrund obiger Rechenergebnisse zwar sagen, daß ein Verdacht besteht, in der Stichprobe hingen Tablettenkonsum und Alkoholkonsum in dem Sinne zusammen, daß das eine das andere ersetzt. Wie wir aber weiterhin sehen werden, kann dieser Verdacht aufgrund der vorhandenen Daten statistisch nicht gesichert werden. Wie gehen wir zur Erlangung dieser Aussagen nun im einzelnen vor ?

Nach den uns bekannten Methoden können wir nun für die Meßwerte x_i und y_i die arithmetischen Mittel und die Standardabweichungen berechnen. Hierdurch erfahren wir aber noch nichts über den Zusammenhang der Variablen.

Was bedeutet nun "Zusammenhang" im Sinne der Statistik? Der Begriff "Zusammenhang" meint die Möglichkeit, in Kenntnis des Wertes der einen Variablen den Wert der anderen vorhersagen zu können. Eine solche Vorhersage ist bei intervallskalierten Daten gut formulierbar, wenn eine Funktion $x=f(y)$ benannt werden kann, die dieser Vorhersage zugrunde liegt. Weiterhin kann die Vorhersage mehr oder minder genau sein, so daß der vorhergesagte Wert eintrifft oder knapp bzw. weit verfehlt wird.

In der Statistik löst man dieses Problem so, daß eine **Funktion** gesucht wird, bei der **die Summe der quadrierten Differenzen zwischen prognostizierten und eingetretenen Werten der Ordinate in unserer Graphik ein Minimum wird**.

Zunächst gilt es, den Funktionstyp festzulegen. Hierbei können wir uns fragen, ob die Funktion ein Polynom vom 1. Grade (eine lineare Funktion) oder ein Polynom höheren Grades sein soll (Parabel, kubische Funktion etc.). Die Entscheidung hierüber liegt in der Regel bei der Theorie über den Gegenstand. Haben wir den Funktionstyp festgelegt (wir wählen hier eine Gerade), dann wählen wir die Methode der Berechnung des Verlaufs der gesuchten Funktion. Diese Methode soll, wie oben gesagt, gewährleisten, daß die gesuchte Funktion die **Bedingung minimaler Abweichungsquadrate** erfüllt.

Der Fall, den wir hier behandeln, geht von dem Funktionstyp "Gerade" aus. Da eine Gerade durch zwei Konstanten, den Abschnitt auf der y-Achse a und die Steigung b , festgelegt wird, müssen wir a und b berechnen. Die Koeffizienten a und b nennen wir die Regressionskoeffizienten.

Wenn wir von x ausgehend y vorhersagen wollen, ergibt sich allerdings eine andere Gerade als wenn wir von y ausgehend x vorhersagen wollen.

Deshalb suchen wir zwei Geraden, die sogenannten **Regressionsgeraden**:

$$(1) y = a_1 + b_1 x$$

$$(2) x = a_2 + b_2 y$$

Zu beachten ist, daß wir in der ersten Gleichung y, in der anderen x vor dem Gleichheitszeichen haben.

Dies ist wichtig, wenn wir die Geraden zeichnen wollen, weil a_1 ein Abschnitt auf der y-Achse, a_2 ein **Abschnitt auf der x-Achse ist**.

Die Berechnung von b_1 , b_2 , a_1 und a_2 kann nun nach folgendem Verfahren erfolgen, das allerdings umständlich ist, und das wir in der Praxis nicht anwenden:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_j - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \begin{array}{l} \text{(Kovarianz)} \\ \text{(Summe der quadrierten Abweichungen)} \end{array}$$

$$a_1 = \bar{y} - b_1 \bar{x}$$

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_j - \bar{y})}{\sum_{i=1}^n (y_j - \bar{y})^2} \quad \begin{array}{l} \text{(Kovarianz)} \\ \text{(Summe der quadrierten Abweichungen)} \end{array}$$

$$a_2 = \bar{x} - b_2 \bar{y}$$

Vereinfachte Berechnungsweisen sind die folgenden (weniger Rechenaufwand, obwohl die Formeln komplizierter aussehen):

Wir bestimmen zunächst folgende Ausdrücke:

n Menge der Meßwertpaare

$\sum_{i=1}^n x_i$ Summe der Meßwerte der Variablen x

$\sum_{i=1}^n x_i^2$ Summe der quadrierten Meßwerte der Variablen x

$\sum_{i=1}^n y_j$ Summe der Meßwerte der Variablen y

$\sum_{i=1}^n y_j^2$ Summe der quadrierten Meßwerte der Variablen y

$\sum_{i=1}^n x_i y_j$ Summe der Produkte der Meßwertpaare

$\left(\sum_{i=1}^n x_i \right)^2$ Quadrierte Summen der Meßwerte der Variablen x bzw. y:

$\left(\sum_{i=1}^n y_j \right)^2$

Diese setzen wir in die unten aufgeführten Formeln ein:

$$b_1 = \frac{n \sum_{k=1}^n x_i y_j - \left(\sum_{k=1}^n x_i \right) \left(\sum_{k=1}^n y_j \right)}{n \sum_{k=1}^n x_i^2 - \left(\sum_{k=1}^n x_i \right)^2}$$

$$a_1 = \frac{\left[\sum_{k=1}^n y_j \right] \left[\sum_{k=1}^n x_i^2 \right] - \left[\sum_{k=1}^n x_i \right] \left[\sum_{k=1}^n x_i y_j \right]}{n \sum_{k=1}^n x_i^2 - \left[\sum_{k=1}^n x_i \right]^2}$$

bzw. vereinfacht: (s.o.)

$$a_1 = \bar{y} - b_1 \bar{x}$$

9.1.2. Der Produkt-Moment-Korrelationskoeffizient r

Während die Regressionskoeffizienten für gegebene Meßwertpaare nur die Geraden angeben, an die die Meßwerte in Richtung x- oder y-Achse nach dem Kriterium minimaler Abstandsquadrate optimal angepaßt sind, soll der Produktmoment-Korrelationskoeffizient den Anpassungsgrad selber signalisieren. Dabei sind folgende Fälle zu unterscheiden:

- Totale Anpassung : $r = \pm 1$
- keine Anpassung : $r = 0$
- "mittlere" Anpassung: $0 < |r| < 1$

Das Vorzeichen von r ist das gleiche wie das der Steigung von b_1 .

r wird aus folgender "Urform" gewonnen und dann zu einer rechentechnisch optimierten Form umgeformt:

$$r = \frac{\sum_{i=j=1}^n (x_i - \bar{x}) * (y_j - \bar{y})}{n * \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} * \sqrt{\frac{\sum_{j=1}^n (y_j - \bar{y})^2}{n}}} \quad \text{(Kovarianz)}$$

$$= \frac{\sum_{i=j=1}^n (x_i - \bar{x}) * (y_j - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 * \sum_{j=1}^n (y_j - \bar{y})^2}} \quad \text{Produkt der Standardabweichungen, multipliziert mit n).}$$

Die rechentechnisch optimierte Formel lautet:

$$r = \frac{n * \sum_{i=1}^n x_i y_j - \sum_{i=1}^n x_i * \sum_{j=1}^n y_j}{\sqrt{\left[n * \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] * \left[n * \sum_{j=1}^n y_j^2 - \left(\sum_{j=1}^n y_j \right)^2 \right]}}$$

In der "Urform" enthält r im Zähler deutlich sichtbar die "Kovarianz", im Nenner das geometrische Mittel der Summen der quadrierten Abweichungen von x bzw. y. Die Kovarianz ist ein Ausdruck für die gemeinsame Veränderung der Variablen x und y. Ist sie gleich 0, besteht kein Zusammenhang. Sie wird gleich 0, wenn die Regressionsgeraden entweder zur X-Achse oder zur Y-Achse parallel verlaufen. Sie ist positiv, wenn die Gerade nach rechts ansteigt, negativ, wenn sie nach rechts fällt. Das Produkt der Abweichungs-Quadratsummen im Nenner ist gleich 0, wenn auf einer Variablen keine Variation vorliegt. Dann ist der Koeffizient nicht berechenbar (wegen des Verbots, durch 0 zu teilen).

Bei Linearität des Zusammenhangs erreicht die Kovarianz genau das geometrische Mittel der summierten Abweichungsquadrate, so daß $r = 1$; bei Abweichungen von der Linearität der Wertepaare ist die Kovarianz kleiner als das geometrische Mittel der summierten Abweichungsquadrate.

Daher ist in solchen Fällen $0 < |r| < 1$.

9.1.3. Ein Rechenbeispiel

Wir gehen von den oben aufgeführten Daten zum Zusammenhang von Alkoholkonsum und Tablettenkonsum aus, für den oben schon eine Graphik erstellt wurde.

Wertepaare und Arbeitstabelle für Rechengang Nr. 2:

f _k	Alkohol in cl		Tabletten pro Tag		x _i * y _j
	x _i	x _i ²	y _j	y _j ²	
1	5	25	9	81	45
1	2	4	13	169	26
1	8	64	1	1	8
1	6	36	9	81	54
1	6	36	10	100	60
1	8	64	3	9	24
1	7	49	15	225	105
1	2	4	6	36	12
n= 8	$\sum_{i=1}^n x_i = 44$	$\sum_{i=1}^n x_i^2 = 282$	$\sum_{j=1}^n y_j = 66$	$\sum_{j=1}^n y_j^2 = 702$	$\sum_{i=j=1}^n x_i y_j = 334$
	$(\sum x_i)^2 = 1936$		$(\sum y_j)^2 = 4356$		

Wir berechnen zunächst b₁:

$$b_1 = \frac{8 * 334 - 44 * 66}{8 * 282 - 1936} = \frac{-232}{320} = -0,725$$

Sodann ermitteln wir a₁:

$$a_1 = 8,25 - (-0,725 * 5,5) = 12,238$$

Danach ermitteln wir b_2 :

$$b_2 = \frac{8 * 334 - 44 * 66}{8 * 702 - 4356} = \frac{-232}{1260} = -0,184$$

sodann a_2 :

$$a_2 = 5,5 - (-0,184 * 8,25) = 7,018$$

Schließlich setzen wir die errechneten Werte für r ein:

$$r = \frac{8 * 334 - 44 * 66}{\sqrt{(8 * 282 - 1936)(8 * 702 - 4356)}} = -0,365$$

Die Regressionsgeraden sind somit bestimmt als

1. Regressionsgerade:

$$y = a_1 + b_1x = 12,238 - 0,725 x$$

2. Regressionsgerade:

$$x = a_2 + b_2y = 7,018 - 0,184y$$

Merke :hier ist $x = f(y)$

Beide Regressionsgeraden sind in die o.a. Graphik eingezeichnet. Ihr Winkel korrespondiert mit dem Korrelationskoeffizienten r insofern, als ein Korrelationskoeffizient von 1 dem Winkel von 0° entspräche, ein Korrelationskoeffizient von 0 einem Winkel von 90° .

Die Voraussetzung für die Berechnung des Korrelationskoeffizienten ist, daß beide Variablen x und y normalverteilt sind und daß es sinnvoll ist, eine Gerade als Repräsentantin entsprechender „Punktwolken“ aus Wertepaaren anzusehen.

Hier muß nur noch darauf hingewiesen werden, daß das Quadrat des Korrelationskoeffizienten

r^2

auch als Determinationskoeffizient bezeichnet wird und angibt, in welchem Maße die Varianz der Variable y durch die Varianz der Variable x erklärt wird (siehe hierzu auch η^2 in der Varianzanalyse unten).

9.2. Zwei ordinal skalierte Merkmale

9.2.1. Kruskals Gamma

Auch für den Fall ordinal skalierten Merkmale ist man an der statistischen Beschreibung von Zusammenhängen interessiert. Das heißt, man will in Kenntnis des Wertes - der Ausprägung - einer Variablen den Wert der anderen vorhersagen. Während nun bei intervallskalierten Merkmalen oder Variablen mit Abständen und Abstandsquadraten argumentiert werden kann, ist dies bei Rangskalen verboten; daher ergibt sich die Notwendigkeit, neue Konzepte für Rangkorrelationen zu entwickeln. Hierbei geht eine Gruppe von Statistikern vom Konzept der konkordanten und diskordanten Paare aus.

Beispiel:

13 Schüler geben ein Urteil ab über 3 Lehrer, die sie in einem System von je drei Rängen plazieren, einmal auf der Dimension "Sachbezogenheit", zum anderen auf der Dimension "positiver Beitrag zum Gruppenklima".

Das Ergebnis sei die folgende Tabelle mit 39 Urteilen, die für beide Dimensionen vorliegen; für jeden Lehrer gibt es die Angabe der Position auf beiden Rangskalen, die jeweils 3 Ränge umfassen:

39 Urteile über 3 Lehrer durch 13 Schüler, betreffend deren Sachbezogenheit und ihren Beitrag zum Gruppenklima.

		Rang: "Gruppenklima"			
Rang		1	2	3	
Sach- bezogen- heit	1	5	7	2	Die Zahlen in den einzelnen Zellen bedeuten: obere Zahl: Häufigkeit untere Zahl: Kombination der zellenspezifischen Ränge.
		11	12	13	
	2	1	9	0	
		21	22	23	
	3	1	3	11	
		31	32	33	

Die Zelle mit der Zugehörigkeit zu den Rängen 1 auf beiden Skalen hat insofern die Kennzeichnung 11.
Die Ränge werden hier so interpretiert, *daß kleine Rangzahlen höhere Ränge anzeigen als große Rangzahlen.*

Ein Lehrer mit den Rängen

$$x = 1 / y = 1$$

und ein Lehrer mit den Rängen

$$x = 2 / y = 2$$

bilden nun ein sogenanntes **konkordantes Paar**, da beide Rangunterschiede auf beiden Variablen "gleichsinnig" sind.

$$(x: 1 > 2; y: 1 > 2)$$

Ein Lehrer, der die Ränge

$$x = 1 / y = 3$$

zugesprochen bekommt, und einer, der die Ränge

$$x = 3 / y = 1$$

zugesprochen bekommt, bilden ein sogenanntes **"diskordantes Paar"**, da die Rangunterschiede "ungleichsinnig sind" ($x: 1 > 3; y: 3 < 1$).

Die Menge aller konkordanten Paare wird nun mit N_c bezeichnet, die Menge der diskordanten Paare mit N_d .
Kruskals γ („gamma“), ein einfaches Maß für die Rangkorrelation, wird aus

$$\gamma = \frac{N_c - N_d}{N_c + N_d}$$

bestimmt.

Die Berechnung der Menge konkordanten Paare (N_c) erfolgt durch Aufsummieren der aller Produkte von Häufigkeiten in Zellen, die konkordante Paare enthalten, wie

Zelle	Häufigkeit	Zelle	Häufigkeit	Produkt
11	5	22	9	$5 * 9 = 45$
11	5	23	0	$5 * 0 = 0$
11	5	32	3	$5 * 3 = 15$
11	5	33	11	$5 * 11 = 55$
12	7	23	0	$7 * 0 = 0$
12	7	33	11	$7 * 11 = 77$
21	1	32	3	$1 * 3 = 3$
21	1	33	11	$1 * 11 = 11$
22	9	33	11	$9 * 11 = 99$
				$N_c = 305$

Die Berechnung der Menge diskordanter Paare erfolgt analog:

Zelle	Häufigkeit	Zelle	Häufigkeit	Produkt
13	2	22	9	$2 * 9 = 18$
13	2	21	1	$2 * 1 = 2$
13	2	32	3	$2 * 3 = 6$
13	2	31	1	$2 * 1 = 2$
12	7	21	1	$7 * 1 = 7$
12	7	31	1	$7 * 1 = 7$
23	0	32	3	$0 * 3 = 0$
23	0	31	1	$0 * 1 = 0$
22	9	31	1	$9 * 1 = 9$
				$N_d = 51$

$$\gamma = \frac{305 - 51}{305 + 51} = 0,713$$

In diese Rechnung sind allerdings noch keine Paare einbezogen, bei denen zeilen- oder spaltengleiche Ränge existieren, wie z.B. bei den Zellen

11 und 12, bei denen der Rang in der Variable Gruppenklima ungleich ist, der bei Sachbezogenheit aber gleich,
oder 11 und 21, bei denen der Rang in der Variable Gruppenklima gleich ist, der bei Sachbezogenheit aber ungleich.

Paare, die in sich solchen Zellen befinden, bilden die sogenannten ties, also (teilweise) ranggleiche Elemente in der Menge der Paare.

Die Berechnung von Gamma berücksichtigt nur die anderen (konkordanten, diskordanten) Paare. Gamma ist wie r zwischen -1 und +1 definiert. Da es "ties" (Übereinstimmungen von Rängen in einer der Variablen) nicht berücksichtigt, zeigt es jedoch oft zu "günstige" Werte an und täuscht Zusammenhänge vor.

9.2.2. Kendalls τ_b ("tau_b").

Ein Maß, das die ties berücksichtigt, ist das Maß τ_b (tau_b) von Kendall, das eine erweiterte Form von Kendalls τ_a („tau_a“) ist. τ_a ist Kruskals γ vergleichbar:

$$\tau_a = -\frac{N_c - N_d}{\frac{n(n-1)}{2}}$$

Hier steht im Nenner eine Konstante, die Menge der Zellen einer quadratischen Matrix oberhalb der Hauptdiagonale. Sie betrüge für eine 3*3-Matrix genau 3.

τ_b hat einen anders definierten Nenner als Kruskals γ :

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}}$$

In diesem Nenner bedeutet T_x die Menge der "ties" auf der Variablen x (hier „Sachbezogenheit“), T_y die Menge der „ties“ auf der Variablen y (hier „Beitrag zum Gruppenklima“).

Die T_x errechnen sich wie folgt:

Zelle	Häufigkeit	Zelle	Häufigkeit	Produkt
11	5	21	1	5 * 1 = 5
11	5	31	1	5 * 1 = 5
21	1	31	1	1 * 1 = 1
12	7	22	9	7 * 9 = 63
12	7	32	3	7 * 3 = 21
22	9	32	3	9 * 3 = 27
13	2	23	0	2 * 0 = 0
13	2	33	11	2 * 11 = 22
23	0	33	11	0 * 11 = 0
				$T_x = 144$

Die T_y errechnen sich wie folgt:

Zelle	Häufigkeit	Zelle	Häufigkeit	Produkt
11	5	12	7	5 * 7 = 35
11	5	13	2	5 * 2 = 10
12	7	13	2	7 * 2 = 14
21	1	22	9	1 * 9 = 9
21	1	23	0	1 * 0 = 0
31	1	32	3	1 * 3 = 3
31	1	33	11	1 * 11 = 11
32	3	33	11	3 * 11 = 33
				$T_y = 115$

$$T_x = 144$$

$$T_y = 115$$

Setzt man diese Werte in die Formel ein, dann ergibt sich

$$\tau_b = \frac{305 - 51}{\sqrt{(305 + 51 + 144)(305 + 51 + 115)}} = 0,523$$

Da γ , in dem die "ties" nicht berücksichtigt wurden, den (größeren) Wert 0,713 hatte, ist der "dämpfende" Einfluß der Einbeziehung der ties in die Formel von τ_b hier gut sichtbar. Welche ties diesen Einfluß haben, kann man am Nenner ablesen: Der Einfluß der $T_x=144$ ist stärker als der Einfluß der $T_y=115$.

9.3 Zwei nominal skalierte Merkmale

9.3.1 Cramér's V

Für den Fall nominal skalierten Merkmale ist die Möglichkeit nicht mehr gegeben, zwischen einer gleichsinnigen oder ungleichsinnigen Veränderung der Variablen y bei Veränderung der Variablen x zu unterscheiden, sondern nur noch die Möglichkeit, zu messen, wie stark die beobachteten Werte von erwarteten Werten abweichen. Genau dies haben wir beim zweidimensionalen Chi-Quadrat-Test getan. Ist diese Abweichung groß, dann wird das Bestehen eines Zusammenhangs vermutet. Was dies bedeutet, kann an folgendem Beispiel demonstriert werden:

Wir benutzen wiederum das oben zum Chi-Quadrat-Test angeführte Beispiel:

Mein Hund soll schön und auffallend sein / Wenn der Hund seine Aufgaben nicht erfüllt: Anderen anschaffen

			Anderen anschaffen				Gesamt
			keines-falls	eher nicht	vielleicht doch	eher ja	
Schöner Hund	keines-falls	Anzahl	86	20	6	10	122
		Erwartete Anzahl	71,6	27,3	11,1	12,0	122,0
		Residuen	14,4	-7,3	-5,1	-2,0	
	eher nicht	Anzahl	120	51	10	15	196
		Erwartete Anzahl	115,1	43,8	17,8	19,3	196,0
		Residuen	4,9	7,2	-7,8	-4,3	
	viel-leicht doch	Anzahl	74	33	18	13	138
		Erwartete Anzahl	81,0	30,8	12,5	13,6	138,0
		Residuen	-7,0	2,2	5,5	-,6	
	eher ja	Anzahl	30	14	14	14	72
		Erwartete Anzahl	42,3	16,1	6,5	7,1	72,0
		Residuen	-12,3	-2,1	7,5	6,9	
Gesamt	Anzahl	310	118	48	52	528	
	Erwartete Anzahl	310,0	118,0	48,0	52,0	528,0	

Chi-Quadrat-Tests

	Wert	df
Chi-Quadrat nach Pearson	35,491 ^a	9
Anzahl der gültigen Fälle	528	

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 6,55.

Hier gilt es nur noch, ein Maß der Stärke des Zusammenhangs nachzureichen. Die Stärke des Zusammenhangs wird durch folgenden Koeffizienten gemessen, der zwischen 0 und 1 definiert ist:

$$V = \sqrt{\frac{\chi^2}{n(\text{Min}(r-1); (c-1))}}$$

Dieser Koeffizient, Cramér's V, hat im Zähler das oben berechnete χ^2 , im Nenner das Produkt aus n (der Tabellensumme) und dem kleineren von zwei Werten (das bedeutet "Min"):

der Menge der Zeilen minus 1 (=r-1); unsere obige Tabelle hat 4 Zeilen;
oder
der Menge der Spalten minus 1 (=c-1); unsere obige Tabelle hat 4 Spalten.

Für unser konkretes Beispiel errechnen wir

$$V = \sqrt{\frac{35,419}{528 * (4-1)}} = 0,149$$

Da Zeilenmenge und Spaltenmenge identisch sind, ist es hier egal, ob wir r-1 oder c-1 in den Nenner einbeziehen. Die Definition des Nenners stellt sicher, daß V den Wert 1 auch dann erreichen kann, wenn keine quadratische Matrix vorliegt. Im vorliegenden Fall haben wir es mit einem schwachen Zusammenhang zu tun, der aber – weil die Nullhypothese im Chiquadrat-Test abgelehnt wurde, als statistisch auf dem 1%-Niveau der Irrtumswahrscheinlichkeit gesichert gilt.

9.3.2 Phi und korrigiertes χ^2

Die Messung von Zusammenhängen zwischen nominal skalierten Merkmalen durch Cramers V bezog sich auf den Fall von Tabellen beliebig großen Formats. Sinkt das Tabellenformat auf 2 x 2 ab, dann sinkt der Freiheitsgrad von χ^2 auf 1. In diesem Fall wirkt sich die Tatsache, daß bei der oben gezeigten Berechnungsmethode für Erwartungswerte auch Brüche vorkommen können, in der Realität aber nur ganze Zahlen als absolute Häufigkeiten denkbar sind, verfälschend auf den Wert von χ^2 aus. Weiterhin gibt es für Vierfeldertabellen ein vereinfachtes Verfahren der Berechnung des statistischen Zusammenhangs, d.h. es brauchen nicht mehr einzelne Erwartungswerte etc. ermittelt zu werden, sondern nur noch der folgende Ausdruck (Phi):

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Das zur Kontinuitätskorrektur nach Yates modifizierte Chiquadrat errechnet man für Vierfeldertafeln nach

$$\chi_y^2 = \frac{\left[|ad - bc| - \frac{n}{2}\right]^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Die Buchstaben a, b, c und d in diesen Formeln bedeuten absolute Häufigkeiten in den Zellen einer quadratischen Tabelle mit 4 Feldern:

a	b
c	d

"a" ist also immer die Häufigkeit "links oben", "d" die Häufigkeit "rechts unten" etc.

Rechenbeispiel:

In einer Untersuchung über die geschlechtsspezifischen Einstellungen zur Nutzung der Atomkraft verteilten sich die Aussagen für bzw. gegen ein sofortiges Abschalten aller Atomkraftwerke wie folgt:

für sofortiges Abschalten der AKW	Frauen	Männer	Summe
	7	4	11
gegen sofortiges Abschalten der AKW	5	9	14
Summe	12	13	25

Setzt man die Werte in die Formel für Φ ein, ergibt sich:

$$\Phi = \frac{7 * 9 - 4 * 5}{\sqrt{(7+4)(5+9)(7+5)(4+9)}} = 0,28$$

$$\chi^2_y = \frac{\left[7 * 9 - 4 * 5 - \frac{25}{2}\right]^2 * 25}{(11)(14)(12)(13)} = 0,968$$

Prüft man die Nullhypothese, daß die Häufigkeiten „randverteilt“ sind bei 1 % Irrtumswahrscheinlichkeit einseitig, dann müßte man dies an einer Chiquadratverteilung mit einem Freiheitsgrad tun. Der Kriteriumswert betrüge dann 6,63.

Da das berechnete Chiquadrat nur 0,968 beträgt, behalten wir die Nullhypothese bei, d.h. wir rechnen nicht mit einem Zusammenhang von Geschlecht und Einstellung zur Nutzung der Atomkraft.

Der Statistiker Cole schlägt noch eine Korrektur von Phi vor, weil Phi den Maximalwert 1 nur annehmen kann, wenn die Felder einer Diagonale leer sind. Um eine Unterschätzung des Zusammenhangs zu vermeiden, schlägt er vor, Phi nach verschiedenen Formeln zu berechnen, je nachdem, ob der Zähler ein positives oder negatives Vorzeichen hat. Einzelheiten referieren Clauß und Ebner (Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, 2. Auflage 1975, S. 285).

10. Signifikanztests für Korrelations- und Assoziationsmaße

10.1. t-Test für den Produktmomentkorrelationskoeffizienten

Die Nullhypothese lautet in diesem Fall: "In der Grundgesamtheit ist die Korrelation gleich Null".

Der Test dieser Hypothese ist möglich mit Hilfe eines kritischen Wertes "t", der wie folgt berechnet wird.

$$t = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

df= n-2

Wenden wir diese Formel auf unser Rechenbeispiel vom Zusammenhang des Alkohol- und Tablettenkonsums an, dann können wir einsetzen:

$$t = 0,365 \frac{\sqrt{8-2}}{\sqrt{1-0,133}} = 0,96$$

für $8-2 = 6$ Freiheitsgrade wäre der Kriteriumswert der t-Verteilung bei einseitigem Test und 1 % Irrtumswahrscheinlichkeit $t = 3,14$. Da unser empirischer t-Wert mit $t = 0,96$ kleiner ist, behalten wir die Nullhypothese bei und Verneinen die Vermutung eines Zusammenhangs von Alkohol- und Tablettenkonsum.

Einfacher ist es noch, eine Tabelle von Zufallshöchstwerten von r zu Rate zu ziehen (vgl. Erna Weber, Jena 1972, S. 652).

10.2. Z-Test für Kendalls "tau"

Für Kendalls tau existiert ein Signifikanztest auf der Basis der standardisierten Normalverteilung. Der kritische z-Wert

$$|Z| = \frac{(N_c - N_d) \pm 1}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}$$

wird unter folgender Nebenbedingung berechnet; für $N_c < N_d$ ist die Zahl 1 dem Zähler hinzuzuaddieren; ist $N_c > N_d$, ist 1 zu subtrahieren. Bedenken gegen diesen Koeffizienten sind insofern legitim, als die z-Verteilung und die t-Verteilung Abstände voraussetzen, die bei Rangkorrelationen nicht gegeben sind. Weiterhin wird die Menge der ties nicht in die Kriteriumsrechnung einbezogen.

10.3. Einschränkungen der Chiquadrat-Tests für Cramers V und Phi

Es sei hier noch nachgetragen, daß bei Tabellen mit im Durchschnitt weniger als 5 Elementen pro Zelle der Chiquadrattest problematisch wird, weil die Chiquadratverteilung große n voraussetzt. In der Praxis wird man daher durch Modifikation von Tabellen (mittels Zusammenfassung von Kategorien oder Weglassen schlecht besetzter Kategorien) dafür Sorge tragen, daß nicht mehr als 20 % der Zellen Erwartungswerte von weniger als 5 haben. Allerdings ist in der Berichterstattung auf eine solche Prozedur hinzuweisen.

Für kleine n wird statt der Chiquadrat-Methoden "Fishers exakter Test" empfohlen, der allerdings nur den Vierfelder-Fall betrifft.

10.4 Fisher's exakter Test

Eine Methode zur Analyse von Vierfeldertafeln in bezug auf die Frage, ob bestehende Zusammenhänge auch in der Grundgesamtheit vorhanden sind, bietet Fisher's "exakter" Test. Er gestattet es, die Wahrscheinlichkeiten für das Auftreten bestimmter Zellenbesetzungen zu berechnen, wenn die Randverteilungen gegeben sind. Für folgenden Fall

		10
		10
10	10	

sind folgende Zellenbesetzungen denkbar:

0	10
10	0

1	9
9	1

2	8
8	2

3	7
7	3

4	6
6	4

5	5
5	5

6	4
4	6

7	3
83	7

8	2
2	8

9	1
1	9

10	0
0	10

Der Test berechnet für eine konkrete Verteilung, z.B. für die Verteilung

3	7
7	3

die Wahrscheinlichkeit ihres Eintretens nach Maßgabe der hypergeometrischen Verteilung mit der Formel

$$\text{Formel: } p(a)_x = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Hierbei sind a, b, c und d die Häufigkeiten einer Vierfeldertabelle

a	b
c	d

$p(a)_x$ ist die Wahrscheinlichkeit, daß - bei gegebenen Randsummen - a den Wert x annimmt. Für unser o.a. Beispiel können wir ermitteln:

Punktwahrscheinlichkeiten

aufsummierte Werte:

$p(a)_0 = 0,000005413$	0,000005413
$p(a)_1 = 0,000541254$	0,001087921
$p(a)_2 = 0,010960402$	0,011507069
$p(a)_3 = 0,077940635$	0,089447703
$p(a)_4 = 0,238693193$	0,328140896
$p(a)_5 = 0,343718198$	0,6718591
$p(a)_6 = 0,238693193$	0,410552287
$p(a)_7 = 0,077940635$	0,988492922
$p(a)_8 = 0,010960402$	0,99453324
$p(a)_9 = 0,000541254$	0,999994578
$p(a)_{10} = 0,000005413$	1,0000

Der Fisher Test summiert nun die Wahrscheinlichkeiten auf, die für ein konkretes $a = x$ **und alle möglichen kleineren $a < x$ gelten**. Diese Summe

$$\sum_{i=0}^x p(a = i)$$

wird nun wie folgt interpretiert: Ist ihr Wert größer als ein Kriteriumswert (die zugelassene Wahrscheinlichkeit für die irrtümliche Ablehnung der Nullhypothese, daß in der Grundgesamtheit kein Zusammenhang besteht), wird die Nullhypothese angenommen.

Der Fisher-Test wird als "einseitiger" Test interpretiert: er ist nur anwendbar auf die schwächstbesetzte Zelle einer Tabelle! Diese muß beim Test immer in der Position von a sein:

a	
---	--

Modifikationen referiert Sidney Sigel, a.a.O., S. 101 ff.

Die o.a. Berechnungen zeigen, daß der Fall

daß aber der Fall

2	8
8	2

noch zur Ablehnung der Nullhypothese führt,

3	7
7	3

schon zur Annahme der Nullhypothese führt.

Das gleiche Ergebnis ist aber auch näherungsweise mit Hilfe der Chiquadrattechnik zu erzielen, wenn man die Yates-Korrektur benutzt.

11. Einfaktorielle Varianzanalyse

Während beim t-Test aufgrund der Mittelwertsdifferenzen geprüft werden kann, ob zwei Stichproben aus einer gemeinsamen Grundgesamtheit gezogen worden sein können, bietet die einfaktorielle Varianzanalyse die Möglichkeit, mehrere gleichgroße Stichproben in bezug auf die Frage zu untersuchen, ob sie alle aus einer gemeinsamen Grundgesamtheit stammen können oder nicht. Die Nullhypothese würde sein:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Die einfaktorielle Varianzanalyse erlaubt es, zu überprüfen, ob sich verschiedene Stichproben, deren Mitglieder aus unterschiedlichen "Bereichen" kommen, so unterscheiden, daß man die Unterschiede durch die Herkunft aus unterschiedlichen Bereichen "erklären" kann. Bereiche können dabei z.B. regional, temporal, institutionell oder auch experimentell definiert sein.

An einer Volkshochschule werden Spanier in Deutsch unterrichtet. Alle Dozenten benutzen das gleiche Lehrmaterial, aber die Personen, die unterrichten, praktizieren informell sehr verschiedene persönliche Lehrstile. Diese wirken sich auf die Motivation der Beteiligten fördernd oder hinderlich aus. Die Menge der Fehltage in einem Jahr schwankt deshalb beträchtlich. Die Frage ist nun, liegt das am Lehrstil?

Jeweils

5 Teilnehmer bei den 4 verschiedenen Dozenten werden nun befragt, wie oft sie bei dem Dozenten im letzten Semester gefehlt haben. Die Antworten lauten:

Dozent Nr.	1	2	3	4
1. Person Fehltage	1	2	3	4
2. Person Fehltage	1	2	5	3
3. Person Fehltage	3	5	7	2

4. Person Fehltag	1	6	5	1
5. Person Fehltag	0	3	6	0

Zunächst können wir die durchschnittliche Fehldauer pro Dozent feststellen: Sie beträgt bei

Dozent Nr.	1	2	3	4
\bar{x}_k	1,2	3,6	5,2	2,0

Hiernach scheint es so zu sein, als ob es aufgrund der Mittelwertdifferenzen deutliche Unterschiede zwischen den einzelnen Teilstichproben gibt. Um dies genauer zu klären, stellen wir den Mittelwert aller Meßwerte fest. Er beträgt für unsere Daten

$$\bar{x}_{tot} = 3 \quad \text{„tot“ ist ein Kürzel für "total - alle".}$$

Die weitere Analyse der Streuungen geht nun folgenden Gang:

Von der Varianz, die - wenn die Grundgesamtheitsvarianz geschätzt werden soll - als

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

definiert ist, wird zunächst nur der Zähler, also die Summe der Abweichungsquadrate, auch "Quadratsumme" genannt, abgekürzt QS, betrachtet.

Beziehen wir in die Berechnung von QS alle Meßwerte aus allen Stichproben ein, und wählen wir

$$\bar{x}_{tot} \quad \text{als Bezugsgröße,}$$

dann erhalten wir

$$QS_{tot} = \sum (x_i - \bar{x}_{tot})^2$$

Dies ist die Quadratsumme, über alle Meßwerte gerechnet.

Dieser Ausdruck wird als Resultat einer Addition aufgefaßt:

$$QS_{tot} = QS_{treat} + QS_{fehler}$$

"Treat" ist hier eine Abkürzung für "Treatment" (Behandlung).

QS_{treat} ist die Quadratsumme, deren Zustandekommen durch

unterschiedliche "Treatments" entstanden ist.

QS_{Fehler} ist eine durch zufällige Variation entstandene Quadratsumme.

Diese Gleichung spiegelt wieder, daß die Unterschiede zwischen den Meßwerten einen systematisch verursachten Anteil, -

$$QS_{\text{treat}}$$

in unserem Falle durch die spezifische Lehrmethode des Dozenten verursachten Anteil - ,

und einen zufälligen Anteil

$$QS_{\text{Fehler}}$$

haben.

Diese Quadratsummen sind ihrerseits chiquadratverteilt, mit Freiheitsgraden, für die folgendes gilt:

$$df_{\text{tot}} = df_{\text{treat}} + df_{\text{fehler}}$$

Teilt man QS_{treat} durch QS_{tot} , dann erhält man, wenn der

Quotient mit 100 multipliziert wird, den „Prozentsatz der Varianzaufklärung“:

$$\text{Varianzaufklärung} = \frac{QS_{\text{treat}}}{QS_{\text{tot}}} * 100 = \eta^2 + 100$$

Die QS_{treat} wird wie folgt berechnet:

$$QS_{\text{treat}} = n \sum (\bar{x}_k - \bar{x}_{\text{tot}})^2$$

Hier werden die Differenzen der "dozenten-spezifischen" Mittelwerte und des Gesamtmittelwerts quadriert, aufsummiert und mit n multipliziert.

Die QS_{fehler} wird wie folgt berechnet:

Für jeden Bereich k - also für jeden Dozenten - wird die dozenten-spezifische Quadratsumme ermittelt. Alle diese Summen werden aufaddiert. Daher die Doppelsumme.

$$QS_{\text{treat}} = \sum \sum (x_i - \bar{x}_k)^2$$

Jeder Quadratsumme entspricht nun eine Schätzung der entsprechenden Varianz der Grundgesamtheit, also der totalen Varianz, der durch das "Treatment verursachten" Varianz und der durch "Fehler" verursachten Varianz.

Diese Schätzungen erhält man, wenn man jeweils eine dieser Quadratsummen durch die ihr entsprechende Menge an Freiheitsgraden teilt:

$$\bar{\sigma}_{tot}^2 = \frac{QS_{tot}}{df_{tot}}$$

$$\bar{\sigma}_{treat}^2 = \frac{QS_{treat}}{df_{treat}}$$

$$\bar{\sigma}_{fehler}^2 = \frac{QS_{fehler}}{df_{fehler}}$$

Die Freiheitsgrade, df, werden hier wie folgt bestimmt:

$df_{tot} = n \cdot p - 1$, wobei in unserem Beispiel n die Menge der Schüler pro Dozent ist, p die Menge der Dozenten.

$$df_{treat} = p - 1$$

$$df_{fehler} = p \cdot (n - 1)$$

Die Frage, ob der Grad der Varianzaufklärung

$$Varianzaufklärung = \frac{QS_{treat}}{QS_{tot}} * 100 = \eta^2 * 100$$

zufällig zustandekommt oder statistisch gesichert ist, wird nun mit Hilfe des F-Tests geprüft. Hierbei gilt die Beziehung:

$$F = \frac{\bar{\sigma}_{treat}^2}{\bar{\sigma}_{fehler}^2}$$

wobei die Menge der Freiheitsgrade jene ist, die im Nenner der in diesen Bruch einbezogenen Ausdrücke steht.

Wie oben schon erläutert, entsteht die F-Verteilung aus zwei Chi-Quadrat-Verteilungen, mit jeweils einer spezifischen Menge an Freiheitsgraden. Die Entscheidung, ob ein konkreter Bruch F zufällig zustande kommt oder nicht, wird wieder - wie schon bei den anderen oben gezeigten Verteilungen - an einem kritischen Wert c orientiert. c entspricht einem vorher festgelegten Niveau der Irrtumswahrscheinlichkeit für eine bestimmte F-Verteilung. Dieses c wird mit dem konkreten berechneten F, für die jeweils relevante Kombination von Freiheitsgraden, verglichen. Ist das empirisch ermittelte $F > c$, wird die o.a. Nullhypothese abgelehnt.

Für unser Beispiel errechnen wir:

$QS_{treat} = 47,2 \quad df_{treat} = 3 \quad \bar{\sigma}_{treat}^2 = 15,733$	
--	--

$QS_{fehler} = 36,8$ $df_{fehler} = 16$ $\overline{\sigma}_{fehler}^2 = 2,300$	$F = \frac{15,733}{2,3} = 6,841$
$QS_{tot} = 84,0$ $df_{tot} = 19$ $\overline{\sigma}_{tot}^2 = 4,442$	$c_{3;16;1\%} = 5,29$

Da $F > c$ ist, können wir die Nullhypothese auf dem 1% Niveau ablehnen. Es zeigt sich eine

$$Varianzaufklärung = \frac{QS_{treat} * 100}{QS_{tot}} = 56,19\%$$

Damit wird in diesem Beispiel das Problem der Fehlzeiten zur über 50 % durch die dozentenspezifische Art des Lehrens erklärt.

Der entsprechende SPSS-Ausdruck sah für frühere SPSS-Versionen (SPSS6 oder 7) folgendermaßen aus:

```
ANOVA /VARIABLES haeuf BY dozent (1,4) /STATISTICS ALL.
```

```

      * * *   C E L L   M E A N S   * * *
      HAEUF
      BY DOZENT

TOTAL POPULATION
      3.00
      (   20)
DOZENT
      1           2           3           4
      1.20        3.60        5.20        2.00
      (    5)    (    5)    (    5)    (    5)

      * * *   A N A L Y S I S   O F   V A R I A N C E   * * *
      HAEUF
      BY DOZENT

Source of Variation   Sum of Squares   DF   Mean Square   F   Signif
Main Effects         47.200           3    15.733       6.841   .004
DOZENT               47.200           3    15.733       6.841   .004

Explained             47.200           3    15.733       6.841   .004
Residual              36.800          16     2.300
Total                84.000          19     4.421

20 Cases were processed.
0 Cases ( .0 PCT) were missing.
```

Rechenschritte einer einfachen Varianzanalyse:

Gegeben sind m Teilstichproben vom Umfang n_i

- Man berechnet die arithmetischen Mittel und die Varianzen für alle Teilstichproben und für den gesamten Datensatz.
- Die Varianzen der Teilstichproben werden mit $(n-1)$ multipliziert. Es resultieren die Quadratsummen (QS) der Teilstichproben.
- Man addiert diese und erhält die Fehlerquadratsumme (QS_{Fehler})
- Man ermittelt die Freiheitsgrade für QS_{Fehler} , indem man die Menge der Teilstichproben m mit $(n-1)$ multipliziert. Es resultiert df_{Fehler} . Dann teilt man QS_{Fehler} durch df_{Fehler} . Es resultiert die Fehlervarianz, die auch Binnenvarianz pro Freiheitsgrad genannt wird und durch σ^2_{Fehler} bezeichnet wird.
- Man ermittelt die Treatmentquadratsumme $QS_{\text{Treatment}}$, in dem man die Differenzen zwischen den Mittelwerten der Teilstichproben und dem Mittelwert der Gesamtstichprobe bildet, diese Differenzen quadriert und aufsummiert. Die Summe multipliziert man mit n . Man ermittelt die Freiheitsgrade $df_{\text{Treatment}}$ von $QS_{\text{Treatment}}$, indem man von der Menge der Teilstichproben m die Zahl 1 subtrahiert ($m-1$). Man teilt $QS_{\text{Treatment}}$ durch $df_{\text{Treatment}}$ und erhält die Zwischenvarianz pro Freiheitsgrad $\sigma^2_{\text{Treatment}}$.
- Man addiert QS_{Fehler} und $QS_{\text{Treatment}}$ und erhält QS_{tot} .
- Man ermittelt den F-Wert, indem man $\sigma^2_{\text{Treatment}}$ durch σ^2_{Fehler} teilt. Dem Zähler werden die Freiheitsgrade des Treatment $df_{\text{Treatment}}$ zugeschrieben, dem Nenner die Freiheitsgrade für die Fehler df_{Fehler} .
Es wird – beispielsweise bei einer Irrtumswahrscheinlichkeit von 1 % der kritische Wert in der F-Verteilung gesucht, der $F_{df_{\text{Treatment}}, df_{\text{Fehler}}, 1\%}$ entspricht. Ist der empirische F-Wert größer als der kritische F-Wert, wird die Nullhypothese verworfen, daß alle Teilstichproben aus der gleichen Grundgesamtheit gezogen worden sein können. Ist der empirische F-Wert größer als der kritische F-Wert, behält man die Nullhypothese bei.
- Man ermittelt den Grad der Varianzaufklärung aus dem Quotienten $QS_{\text{Treatment}} * 100 / QS_{\text{tot}} = \eta^2 * 100$
- Man interpretiert das Ergebnis etwa folgendermaßen: Die Nullhypothese.....wird beibehalten/abgelehnt bei% Irrtumswahrscheinlichkeit. Der Grad der Varianzaufklärung ist....%.

Es gibt in der Statistik viele Symbole, deren Bedeutung bekannt sein müssen, wenn entsprechende Formeln verstanden und sinngemäß angewendet werden sollen. Leider ist die Schreibweise und Definition einiger Symbole nicht bei allen Autoren einheitlich, so daß nicht selten aus dem Kontext erschlossen werden muß, was gemeint ist. Eine grundlegende Einführung auch in die mathematische Symbolsprache ist erfahrungsgemäß deshalb vonnöten, weil viele Studierende sehr schlechte Erfahrungen mit dem Mathematikunterricht gemacht und das meiste aus der Schule wieder vergessen haben.

Da eine kurze Wiederholung eine gründlichere Beschäftigung mit diesen Symbolen nicht ersetzen kann, seien hier noch einschlägige Titel angegeben:

GELLERT, W. u.a. /Hrsg): Kleine Enzyklopädie Mathematik, Frankfurt/M. 1972

WALKER, Helen M.: Mathematik für den statistischen Grundkurs, Weinheim 1971

BORTZ, Jürgen: Lehrbuch der Statistik, 2. Auflage, Anhang A, B und C, Heidelberg 1985.

CLAUß, G., EBNER, H.: Grundlagen der Statistik, Kap. 6, Frankfurt Main 1975

ZÖFEL, Peter: Statistik in der Praxis. Kap (1.2), Stuttgart 1985

Statistik hat es mit Zahlen zu tun.

Deshalb zunächst einiges über Zahlen:

(1) Natürliche Zahlen:

Kardinalzahlen: 1, 2, 3, ...

Kardinalzahlen entstehen durch Abzählen der Elemente einer Menge

Ordinalzahlen: 1., 2., 3., 4., ...

(der erste, zweite, dritte, vierte) entstehen durch Zuweisung von Plätzen oder Rängen.

Die Null gilt auch hier als natürliche Zahl.

(2) Die Grundrechenarten und ihre Schreibweisen:

Addition ($a + b$)

Subtraktion ($a - b$)

(Multiplikation) ($a \cdot b$); (ab)

Diese drei Grundrechenarten sind im Bereich der natürlichen Zahlen unbeschränkt ausführbar.

Die Division ($(a : b)$, $(a : b)$, (a/b) und der waagerechte Bruchstrich) ist nur beschränkt ausführbar, da einige Zahlen sich nicht ohne Rest durch andere natürliche Zahlen teilen lassen.

(3) Ganze Zahlen

Sind mit Vorzeichen versehene natürliche Zahlen

also $-5 -4 -3 -2 -1 0 +1 +2 +3 +4$

Sie sind erforderlich, um alle Subtraktionsaufgaben lösen zu können.

(4) Absolutzahlen

sind Zahlen ohne Vorzeichen: für die ganzen Zahlen $-3 +5 -7$

bzw. sind die AbsolutBETRÄGE $|3| |5| |7|$

(5) Für das Rechnen mit ganzen Zahlen gilt

$a + b = b + a$ (Kommutativgesetz)

$(a+b) + c = a + (b+c)$ (Assoziativgesetz)

Aus $a < b$ folgt $a + c < b + c$ (Monotoniegesetz)

(6) Subtraktionen können auch durch Addition negativer Zahlen ausgeführt werden

$6 - 2 = 6 + (-2)$

(7) Addition und Subtraktion sind Rechenoperationen erster Stufe

Multiplikationen und Divisionen sind Rechenoperationen zweiter Stufe

Letztere haben bei Rechnungen Vorrang (Punkt- geht vor Strichrechnung).

(8) Multiplizieren

kann als fortlaufende Addition des gleichen Betrages aufgefaßt werden.

Für Multiplikationen gilt:

$$a * b = b * a \text{ (Assoziativgesetz)}$$

bei Zahlen > 0 (nicht für negatives c gültig):

aus $a < b$ folgt $a * c < b * c$ (Monotoniegesetz)

Vorzeichenregel:

$$(+u) * (+v) = (-u) * (-v) = u * v$$

$$(+u) * (-v) = (-u) * (+v) = -u * v$$

(9) Division

ist die Umkehrung der Multiplikation. Es gelten die gleichen Vorzeichenregeln

$$(+u) : (+v) = (-u) : (-v) = \frac{u}{v}; \quad (-u) : (+v) = (+u) : (-v) = -\frac{u}{v};$$

(10) Brüche

Bei allen Divisionen entstehen Brüche

$$\text{z.B. der Form } \frac{p}{q}$$

p heißt Zähler, q Nenner.

Brüche mit dem Nenner 0 sind ausgeschlossen.

(11) Erweitern

$$\text{z.B. } \frac{1}{3} = \frac{2}{6} = \frac{4}{12} = \frac{20}{60} \quad \text{Zähler und Nenner werden mit gleicher Zahl multipliziert}$$

Kürzen

$$\text{z.B. } \frac{20}{80} = \frac{10}{40} = \frac{5}{20} = \frac{1}{4} \quad \text{Zähler und Nenner werden durch gleiche Zahl dividiert}$$

(12) Bruchrechnen

Brüche werden addiert, indem man sie gleichnamig macht.

Der Hauptnenner ist das kleinste gemeinsame Vielfache der Nenner:

$$\text{z.B. } \frac{1}{6} + \frac{3}{10} - \frac{11}{15} = \frac{5 + 9 - 22}{30} = \frac{-8}{30} = -\frac{4}{15}$$

(13) Produkt von Brüchen

$$\frac{a}{b} * \frac{c}{d} = \frac{a * c}{b * d}$$

(14) Quotient von Brüchen

$$\frac{a}{b} : \frac{c}{d} = \frac{a}{b} * \frac{d}{c}$$

(15) Distributivgesetz: (Vgl. oben (8)).

$$a * (b+c) = a*b + a*c$$

(16) Dezimalbrüche

Jeder gemeine Bruch läßt sich als endlicher oder periodischer Dezimalbruch schreiben:

$$\text{z.B. } \frac{1}{2} = \frac{5}{10} = 0,5; \quad \frac{34}{99} = 0,\overline{34} = 0,3434343434\dots$$

(17) Rechnen mit Dezimalbrüchen

Addition: **Komma unter Komma:**

$$\begin{array}{r} 0,45 \\ +23,43 \\ \hline =23,88 \end{array}$$

Multiplikation: Wie bei natürlichen Zahlen, Menge der Kommastellen richtet sich nach der Summe der Kommastellen der Faktoren:

$$50 * 0,00004 = 0,002$$

50 hat zwei Vorkommastellen
0,00004 hat fünf Nachkommastellen.
2+5 = -3, also muß das Ergebnis drei Nachkommastellen haben.

Division:

Der Divisor wird zu einer ganzen Zahl gemacht. Entsprechend wird beim Dividieren das Komma verschoben:

$$0,88 : 0,0022 = 8800 : 22 = 400$$

(18) Potenzen und Wurzeln

Potenzen sind Vorschriften für mehrfache Multiplikation:

Wurzeln sind Potenzen mit gebrochenem Exponenten.

$$a * a * a = a^3$$

a ist hier die **Basis**, **3** der **Exponent**. Der Exponent kann im Prinzip jeden Zahlenwert annehmen, also positiv und negativ sein, ganzzahlig oder gebrochen. Auch die Basis darf positiv und negativ sein.

Ausnahmen siehe unten.

Beispiele:

0^0 = verbotene Operation

$$\begin{array}{ll}
10^0 = 1 & (\pm 10)^{-4} = +0,0001 \\
10^1 = 10 & (-10)^{-5} = -0,00001 \\
10^2 = 100 & \\
10^8 = 100\,000\,000 & 100^{\frac{1}{2}} = \sqrt[2]{100} = 10 \\
10^{-1} = 1/10 = 0,1 & 100^{-\frac{1}{2}} = 0,1 = \frac{1}{\sqrt[2]{100}} \\
10^{-8} = 1/10^8 = 0.00000001 & (-100)^{-\frac{1}{2}} = \text{verbotene Operation} \\
& (\text{wenn Basis } a < 0, \text{ darf Exponent nur ganze Zahl sein}) \\
(-10)^2 = +100 & \\
(-10)^3 = -1000 &
\end{array}$$

(19) Regeln zur Potenzrechnung

$$(a \cdot b)^n = a^n \cdot b^n \quad \text{Multiplikation}$$

$$a^n \cdot a^m = a^{n+m}$$

$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n} \quad \text{Division}$$

$$\frac{a^m}{a^n} = a^{m-n}$$

$$a^0 = 1 \quad ; \quad (\text{für } a \neq 0) \quad \text{Erweiterung des Potenzbegriffs auf Potenzen von 0 und den Bereich kleiner 0}$$

$$a^{-n} = \frac{1}{a^n}$$

$$(a^n)^m = a^{nm}$$

$$a^{\frac{n}{m}} = \sqrt[m]{a^n} \quad \text{Wurzeln als Bruch im Exponenten}$$

(20) Logarithmen

Beispiele:

für die Beziehung $10^2 = 100$

kann gesagt werden, daß der Exponent 2 der Logarithmus von 100 auf der Basis 10 ist.

für die Beziehung $2^3 = 8$ kann gesagt werden, daß der Exponent 3 der Logarithmus von 8 auf der Basis 2 ist.

Logarithmengesetze (bezogen auf die Basis b):

$$\log_b (n_1 \cdot n_2) = \log_b n_1 + \log_b n_2$$

Der Logarithmus eines Produkts ist gleich der Summe der Logarithmen der Faktoren

$$\log_b \frac{n_1}{n_2} = \log_b n_1 - \log_b n_2$$

Der Logarithmus eines Quotienten ist gleich der Differenz der Logarithmen des Dividenden und des Divisors

\log_b heißt „Logarithmus auf der Basis b “. **Die Basis muß größer als eins sein.**

$$\log_b (p^r) = r \cdot \log_b p$$

Logarithmus einer Potenz

$$\log_b \sqrt[r]{w} = \frac{1}{r} \log_b w$$

Logarithmus einer Wurzel

Bei den Logarithmen auf der Basis 10 verweist eine positive Zahl vor dem Komma auf die Menge der Stellen der Zahl, ist sie negativ, verweist sie auf die Menge der Nullen nach dem Komma. Die Zahlen der Logarithmen nach dem Komma nennt man auch **Mantisse**.

Beispiele:

$$\begin{aligned}\log_{10} 2000 &= 3,301029996 && (2000 \text{ hat } 4 \text{ Stellen; } 3+1 = 4) \\ \log_{10} 200 &= 2,301029996 \\ \log_{10} 20 &= 1,301029996 \\ \log_{10} 0,02 &= -1,698970004 && (0,02 \text{ hat zwei Stellen nach dem Komma.} \\ &&& - (1+1) = -2\end{aligned}$$

(21) Logarithmen auf der Basis 2, e und 10

Da Logarithmen auf der Basis aller Zahlen, die größer als 1 sind, definiert sein können, ist zu fragen, was „üblich“ ist. Üblich sind die Basis 2, die Basis e und die Basis 10. Die Logarithmen auf der Basis 10 heißen „dekadische“, die auf der Basis e „natürliche“ und die auf der Basis 2 „logarithmus dualis“. Die dekadischen Logarithmen schreibt man lg, die natürlichen ln.

Die Basis „e“ ist eine Zahl, die sich ergibt, wenn man die Reihe

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n, \text{ bzw. deren „Grenzwert“ berechnet.}$$

Ihr Betrag ist 2,718281828..... (e ist eine „transzendente Zahl“, gehört also zu einer Teilmenge der „reellen Zahlen“, Definition siehe unten (23)).

Der Vorteil der Zahl e ist, daß die Funktion e^x einen Differentialquotienten hat, der ihrem Betrag proportional ist, das heißt, er läßt sich leicht berechnen. Der Differentialquotient ist ein Ausdruck für die Steigung.

Die Umrechnung von Logarithmen der einen in die der anderen Basis ist leicht:
 $\lg n = 1/(\ln 10) * (\ln n) = 0,4342945 * (\ln n)$

- (22) Vorteil des Rechnens mit Logarithmen: Jede Rechnung wird auf der nächst-niedrigeren Stufe durchgeführt, Multiplikationen werden zu Additionen, die Potenz wird zur Multiplikation. Dadurch lassen sich in Brüchen Probleme lösen, die sonst nicht lösbar wären. Außerdem lassen sich nichtlineare Funktionen teilweise in lineare Funktionen überführen (Linearisierung).
- (23) Viele Logarithmen, Wurzeln und die Zahlen e (s.o.) und π gehören zu den „reellen“ Zahlen“. Diese sind unendliche - positive und negative - Dezimalbrüche. Die oben erwähnten „transzendenten“ Zahlen genügen keiner algebraischen Gleichung mit ganzzahligen Koeffizienten.
- (24) **Algebraische** Gleichungen sind solche, in denen mit den Unbekannten nur **algebraische** Rechenoperationen vorgenommen werden, d.h., + - * ÷, Wurzel, Potenz. Transzendente Zahlen können in algebraischen Gleichungen vorkommen. Grundformen algebraischer Gleichungen sind die lineare Gleichung, die quadratische Gleichung und schließlich die allgemeine algebraische Gleichung; sie werden hier in der

Die Folge muß keine der Größe nach geordneten Zahlen enthalten, dies gilt nur für bestimmte (z.B.: arithmetische, geometrische) Folgen.

Reihen sind aufsummierte Folgen, z.B.

$$a_1 + a_2 + a_3 + a_4 + a_5 + \dots + a_n = \sum_{i=1}^n a_i \quad \Sigma \text{ heißt „Summe“}$$

(lies: Summe über a_i für $i=1$ bis n)

Wenn keine Mißverständnisse befürchtet werden müssen, können die Summationsgrenzen (hier 1 und n) weggelassen werden.

Beispiele:

$$A_7 + A_8 + A_9 + A_{10} + A_{11} + A_{12} = \sum_{i=7}^{12} A_i$$

$$a_5 \tau_5 + a_6 \tau_6 + a_7 \tau_7 + a_8 \tau_8 + a_9 \tau_9 = \sum_{i=5}^9 a_i \tau_i$$

$$(x_1 - a)^2 + (x_2 - a)^2 + (x_3 - a)^2 + (x_4 - a)^2 = \sum_{i=1}^4 (x_i - a)^2$$

$$\chi_{12}^2 + \chi_{13}^2 + \chi_{14}^2 + \chi_{15}^2 + \chi_{16}^2 = \sum_{i=12}^{16} \chi_i^2$$

$$\begin{aligned} \sum_{i=1}^5 (\chi_i - 2) &= (\chi_1 - 2) + (\chi_2 - 2) + (\chi_3 - 2) + (\chi_4 - 2) + (\chi_5 - 2) \\ &= (\chi_1 + \chi_2 + \chi_3 + \chi_4 + \chi_5) - 10 \\ &= \sum_{i=1}^5 \chi_i - 10 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^4 a \chi_i &= a \chi_1 + a \chi_2 + a \chi_3 + a \chi_4 \\ &= a (\chi_1 + \chi_2 + \chi_3 + \chi_4) = a \sum_{i=1}^4 \chi_i \end{aligned}$$

$\chi_3 + \chi_4 + \chi_7 + \chi_{12}$ kann nicht mit Hilfe des Summationszeichens geschrieben werden.

Doppelte Indizes

Wenn Daten nach mehr als einem Gesichtspunkt eingeteilt werden, ist der Gebrauch von doppelten oder sogar dreifachen Indizes sehr nützlich. Als einfaches Beispiel sei angenommen, daß von N Schülern c verschiedene Tests durchgeführt worden sind, also $N \cdot c$ Messungen. Es sei X_{12} (lies „X eins zwei“) der Testwert von Schüler 1 beim Test 2, X_{71} der Wert von Schüler 7 beim Test 1 und allgemein X_{ij} der Wert von Schüler i beim Test j .

Die Nc Testwerte können in einer rechteckigen Formation, genannt Matrix, angeordnet werden:

In dieser Matrix liegen die Werte jedes einzelnen Schülers in einer Reihe; die Werte aller Schüler in einem einzigen Test sind in einer Spalte. Die Reihe der Werte $X_{i1}, X_{i2}, \dots, X_{ic}$ repräsentiert die Werte von irgendeinem nicht spezifizierten Schüler. Die Spalte $X_{1j}, X_{2j}, \dots, X_{Nj}$ repräsentiert die Werte aller Schüler bei irgendeinem nicht spezifizierten Test.

Algebraische Symbole

$$\begin{array}{cccccc}
 X_{11} & X_{12} & \cdot & X_{1j} & \cdot & X_{1c} \\
 X_{21} & X_{22} & \cdot & X_{2j} & \cdot & X_{2c} \\
 X_{31} & X_{32} & \cdot & X_{3j} & \cdot & X_{3c} \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 X_{i1} & X_{i2} & \cdot & X_{ij} & \cdot & X_{ic} \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 X_{N1} & X_{N2} & \cdot & X_{Nj} & \cdot & X_{Nc}
 \end{array}$$

Man möchte eventuell die Summe der Werte eines einzelnen Schülers bei allen Tests oder die Summe der Werte aller Schüler bei einem einzigen Test angeben oder die Summe der Werte aller Schüler bei allen Tests. Studiere dazu die folgenden Gleichungen und drücke sie dabei in Worten aus.

Für die Werte von Schüler 6 gilt:

$$X_{61} + X_{62} + \dots + X_{6j} + \dots + X_{6c} = \sum_{j=1}^c X_{6j} \quad (\text{Die Summe der } X_{6j}, \text{ wobei } j \text{ von } 1 \text{ bis } c \text{ läuft})$$

Für die Werte in Test 4 gilt:

$$X_{14} + X_{24} + \dots + X_{i4} + \dots + X_{N4} = \sum_{i=1}^N X_{i4} \quad (\text{Die Summe der } X_{i4}, \text{ wobei } i \text{ von } 1 \text{ bis } N \text{ läuft})$$

Für die Werte irgendeines nicht spezifizierten Schülers gilt:

$$X_{1j} + X_{2j} + \dots + X_{ij} + \dots + X_{Nj} = \sum_{i=1}^N X_{ij}$$

Für die Werte irgendeines nicht spezifizierten Tests gilt:

$$X_{i1} + X_{i2} + \dots + X_{ij} + \dots + X_{ic} = \sum_{j=1}^c X_{ij}$$

Übung

Für alle Werte gilt, wenn man sie erst spalten- und dann reihenweise addiert:

$$(X_{12} + X_{21} + \dots + X_{N1}) + (X_{12} + X_{22} + \dots + X_{N2}) + \dots + (X_{1c} + X_{2c} + \dots + X_{Nc})$$

$$\begin{aligned}
&= \sum_{i=1}^N X_{i1} + \sum_{i=1}^N X_{i2} + \dots + \sum_{i=1}^N X_{iC} \\
&= \sum_{j=1}^C \sum_{i=1}^N X_{ij}
\end{aligned}$$

Für alle Werte gilt, **wenn man sie erst reihen- und dann spaltenweise addiert:**

$$\begin{aligned}
&(X_{11} + X_{12} + \dots + X_{1C}) + (X_{21} + X_{22} + \dots + X_{2C}) + \dots + (X_{N1} + X_{N2} + \dots + X_{NC}) \\
&= \sum_{j=1}^C X_{1j} + \sum_{j=1}^C X_{2j} + \dots + \sum_{j=1}^C X_{Nj} \\
&= \sum_{i=1}^N \sum_{j=1}^C X_{ij}
\end{aligned}$$

Offenbar ist das Endergebnis stets gleich, egal ob man zuerst die Reihen oder die Spalten addiert.

(29) Fortlaufendes Produkt.

In Analogie zur Reihe kann eine Folge auch fortlaufend multipliziert werden.
Die Schreibweise ist dann

$$a_1 * a_2 * a_3 \dots * a_n = \prod_{i=1}^n a_i$$

\prod heißt „Produkt“

(lies: Produkt über a_i für $i=1$ bis n)

(30) Fakultät, Permutationen

Das Produkt $1*2*3*4$ wird auch als $4!$ geschrieben (allgemein: $n!$)

(lies: vier Fakultät).

n -Fakultät bedeutet: $n! = 1*2*3*\dots*(n-1)*n$

Es gilt: $0! = 1$; $1! = 1$.

(n -Fakultät entspricht der Zahl der möglichen Vertauschungen der Reihenfolge der Elemente einer Menge mit n verschiedenen Elementen:

Beispiel: Die Buchstaben abc können in $3!=6$ verschiedenen Reihenfolgen geschrieben werden: abc, acb, bac, bca, cab, cba. Solche Vertauschungen heißen auch Permutationen).

(31) Kombinationen und „n über k“

der Ausdruck $\binom{n}{k}$ (lies: n über k) ist eine abgekürzte Schreibweise

von $\frac{n!}{k!(n-k)!}$. Der gleiche Ausdruck wird auch als $\binom{n}{r}$ oder

als $C_{n,r}$ oder als C_r^n oder als ${}_r C_n$ geschrieben. Er ist ein Symbol für die Menge verschiedener

Kombinationen von k Elementen einer Menge mit dem Umfang n , aus der k (bzw. r) Dinge ausgewählt werden, ohne daß es auf die Reihenfolge ankäme.

Beispiel: Beim Spiel 6 aus 49 errechnen sich $\binom{49}{6} = \frac{49!}{6!(49-6)!} = 13983816$

verschiedene Kombinationen von 6 Lottozahlen.

(32) Multiplikation algebraischer Summen

$$(a+b)(c+d) = a(c+d) + b(c+d) = ac + ad + bc + bd$$

(33) Binomische Formeln als Sonderfall der Multiplikation algebraischer Summen

$$(a \pm b)^2 = a^2 \pm 2ab + b^2$$

$$(a+b)(a-b) = a^2 - b^2$$

Bei höheren Potenzen ergibt sich durch die Einführung des Summenzeichens und des Ausdrucks $\binom{n}{k}$ eine vereinfachte Schreibweise:

$$(a+b)^n = \binom{n}{0} a^n b^0 + \binom{n}{1} a^{n-1} b^1 + \binom{n}{2} a^{n-2} b^2 + \dots + \binom{n}{n} a^0 b^n$$

$$= \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

Die Ausdrücke $\binom{n}{k}$ heißen auch **Binomialkoeffizienten**, weil sie in der o.a. Gleichung die Berechnung des Koeffizienten vor den Produkten von a und b gestatten.
Beispiel (die Binomialkoeffizienten werden fett gedruckt)

$$(a+b)^7 = 1a^7 b^0 + 7a^6 b^1 + 21a^5 b^2 + 35a^4 b^3 + 35a^3 b^4 + 21a^2 b^5 + 7a^1 b^6 + 1a^0 b^7$$

Die Entwicklung der Folge der Binomialkoeffizienten ist symmetrisch.

(34) **Matrizenrechnung:** Definition einer Matrix

Matrix = rechteckige Anordnung von Zahlen in Zeilen und Spalten (n = Menge der Zeilen, m = Menge der Spalten in einer n*m Matrix)

Die Matrix wird mit **fettgedruckten Großbuchstaben** bezeichnet, die Elemente haben Indizes, deren erster die zugehörige Zeile, deren zweiter die zugehörige Spalte bezeichnet:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \quad \mathbf{A} = a_{ij} \quad (i=1,2,3; j=1,2,3,4)$$

(35) **Transponierte**

Zeilen und Spalten der ursprünglichen Matrix werden vertauscht: Wenn A z.B.

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & 6 \\ 5 & 0 & 4 \\ 1 & 2 & 1 \end{pmatrix} \quad \text{wäre die Transponierte } \mathbf{A}' = \begin{pmatrix} 2 & 5 & 1 \\ 3 & 0 & 2 \\ 6 & 4 & 1 \end{pmatrix}$$

(36) Matrizen, die nur eine Zeile oder Spalte haben, werden **Vektoren** genannt,

$$\mathbf{v} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \\ \mathbf{V}_4 \end{pmatrix}$$

Spaltenvektoren werden mit fettgedruckten Kleinbuchstaben bezeichnet,
 Zeilenvektoren mit fettgedruckten Kleinbuchstaben mit Strich
 $\mathbf{u}' = (u_1, u_2, u_3)$,
 wenn sie als transponierte Spaltenvektoren angesehen werden.

(37) Matrizen, die nur einen Wert haben, heißen **Skalar**. In einer **Diagonalmatrix** sind nur die Werte in der Diagonale von links oben nach rechts unten besetzt. In der **Identitätsmatrix** ist diese Diagonale mit Einsen besetzt. Die Diagonalmatrix wird mit **D**, die Identitätsmatrix mit **I** bezeichnet. **I wird oft auch als Einheitsmatrix bezeichnet.**

(38) **Addition von Matrizen**

Jedes Element der Summenmatrix ist gleich der Summe der korrespondierenden Matrizenelemente

$$\begin{pmatrix} 3 & 1 \\ 5 & 2 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 8 & 5 \\ 6 & 4 \\ 3 & 7 \end{pmatrix}$$

A + **B** = **C**

(39) **Multiplikation mit einem Skalar oder mit einer Matrix:**

Multiplikation mit einem **Skalar**:

Jedes Matricelement wird mit dem Skalar multipliziert.

Multiplikation mit einer **Matrix** folgt der Regel

$$\mathbf{C} = \mathbf{A} * \mathbf{B} \quad c_{ij} = \sum_{k=1}^s a_{ik} * b_{kj} \quad \text{für } i=1,2,\dots,n; j=1,2,\dots,m \text{ und } k=1,2,\dots,s.$$

A ist hierbei eine n*s Matrix, **B** ist eine s*m Matrix und **C** eine n*m Matrix.

Beispiel:

$$\begin{pmatrix} 5 & 3 & 0 \\ 6 & 9 & 7 \end{pmatrix} * \begin{pmatrix} 0,5 & 0,2 \\ 4 & 8 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} 14,5 & 25 \\ 46 & 115,2 \end{pmatrix}$$

Die einzelnen Elemente der Ergebnismatrix C werden wie folgt errechnet:

$$c_{11} = 5*0,5 + 3*4 + 0*1 = 14,5$$

$$c_{12} = 5 \cdot 0,2 + 3 \cdot 8 + 0 \cdot 6 = 25$$

$$c_{13} = 6 \cdot 0,5 + 9 \cdot 4 + 7 \cdot 1 = 46$$

$$c_{14} = 6 \cdot 0,2 + 9 \cdot 8 + 7 \cdot 6 = 115,2$$

Würde man die Reihenfolge von **A** und **B** in dieser Multiplikation vertauschen, dann hätte **C** nicht wie oben die Dimension 2*2 sondern 3*3:

$$\begin{matrix} \mathbf{A} & * & \mathbf{B} & = & \mathbf{C} \\ \begin{pmatrix} 0,5 & 0,2 \\ 4 & 8 \\ 1 & 6 \end{pmatrix} & & \begin{pmatrix} 5 & 3 & 0 \\ 6 & 9 & 7 \end{pmatrix} & & \begin{pmatrix} 3,7 & 3,3 & 1,4 \\ 68 & 84 & 56 \\ 41 & 57 & 42 \end{pmatrix} \end{matrix}$$

$$c_{11} = 0,5 \cdot 5 + 0,2 \cdot 6 = 3,7$$

$$c_{12} = 0,5 \cdot 3 + 0,2 \cdot 9 = 3,3$$

$$c_{13} = 0,5 \cdot 0 + 0,2 \cdot 7 = 1,4$$

$$c_{21} = 4 \cdot 5 + 8 \cdot 6 = 68$$

$$c_{22} = 4 \cdot 3 + 8 \cdot 9 = 84$$

$$c_{23} = 4 \cdot 0 + 8 \cdot 7 = 56$$

$$c_{31} = 1 \cdot 5 + 6 \cdot 6 = 41$$

$$c_{32} = 1 \cdot 3 + 6 \cdot 9 = 57$$

$$c_{33} = 1 \cdot 0 + 6 \cdot 7 = 42$$

Es kommt bei der Matrizenmultiplikation daher auf die Reihenfolge an. Bedingung für die Matrizenmultiplikation ist, daß die Menge der Spalten der links stehenden Matrix gleich der Menge der Zeilen der rechts stehenden Matrix ist.

(40) Die Matrizenmultiplikation ist distributiv und assoziativ, aber nicht kommutativ:

$$(\mathbf{A} + \mathbf{B}) * \mathbf{C} = \mathbf{A} * \mathbf{C} + \mathbf{B} * \mathbf{C}$$

$$\mathbf{A} * (\mathbf{B} + \mathbf{C}) = \mathbf{A} * \mathbf{B} + \mathbf{A} * \mathbf{C}$$

$$(\mathbf{A} * \mathbf{B}) * \mathbf{C} = \mathbf{A} * (\mathbf{B} * \mathbf{C}) = \mathbf{A} * \mathbf{B} * \mathbf{C}$$

(41) Die Multiplikation mit einer Diagonalmatrix verändert im Fall der Einheitsmatrix die Ursprungsmatrix nicht, gleichgültig, ob von links oder von rechts multipliziert wird.

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

(42) In eine Determinante gehen sämtliche Elemente der Matrix ein. Sie ist eine Zahl, die als **gewichtete Summe der Elemente einer beliebigen Zeile oder Spalte berechnet wird**. (Zum Berechnungsverfahren siehe Bortz (1989, S. 816f).

Die Determinante wird für die Matrix **A**

$|A|$ geschrieben. Die Determinante ist für Ursprungsmatrix und Transponierte identisch: $|A| = |A'|$.

Die Determinante eines Matrixprodukts ist gleich dem Produkt der entsprechenden Determinanten:

$$|AB| = |A| * |B|$$

(43) Die **Inverse** einer Matrix ist diejenige Matrix, die mit der ursprünglichen Matrix

multipliziert, die Einheitsmatrix ergibt.

Die Inverse von \mathbf{A} wird \mathbf{A}^{-1} geschrieben.

Es gilt die Beziehung

$$\mathbf{A} * \mathbf{A}^{-1} = \mathbf{I}$$

$$\mathbf{A}^{-1} * \mathbf{A} = \mathbf{I}$$

Zum Berechnungsverfahren siehe BORTZ (1989, S. 820f)

Sowohl die Berechnung von Inverse als auch Determinante sind bei größeren Matrizen so rechenaufwendig, daß ausschließlich von Rechenprogrammen Gebrauch gemacht wird, um sie zu bestimmen.

- (44) Die Inverse existiert nur für quadratische Matrizen und nur für solche, deren Determinante von 0 verschieden ist. Für symmetrische Matrizen ist die Inverse ebenfalls symmetrisch. Die Inverse einer Transponierten ist gleich der Transponierten der Inversen:
 $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$

- (45) Lineare Gleichungssysteme vgl. oben 24) lassen sich mit Hilfe der Matrizenrechnung wie folgt lösen:

Gegeben sei ein Gleichungssystem mit folgenden Gleichungen:

$$9x_1 + 5x_2 + 3x_3 = 0$$

$$5x_1 + 4x_2 + 5x_3 = 0$$

$$0x_1 + 4x_2 + 4x_3 = 3$$

Wir definieren aus diesen Gleichungen die Matrix

$$\mathbf{A} = \begin{pmatrix} 9 & 5 & 3 \\ 5 & 4 & 5 \\ 0 & 4 & 4 \end{pmatrix}$$

den Spaltenvektor

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

den Spaltenvektor

$$\mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix}$$

Gesucht sind die Elemente des Spaltenvektors \mathbf{x} .

Sie können aus folgender Gleichung errechnet werden:

$$\mathbf{X} = \mathbf{A}^{-1} * \mathbf{c}; \text{ also: } \mathbf{A}^{-1} * \mathbf{c} = \mathbf{x}$$

$$\begin{pmatrix} 0,05 & 0,11 & -0,17 \\ 0,26 & -0,47 & 0,39 \\ -0,26 & 0,47 & -0,14 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix} \quad \begin{pmatrix} -0,51 \\ 1,18 \\ -0,43 \end{pmatrix}$$

Es ergeben sich: $x_1 = -0,51$; $x_2 = 1,18$; $x_3 = -0,43$

(es zeigen sich bei der Probe kleine Rundungsfehler, weil nur zwei Nachkommastellen angezeigt wurden).

(46) Bestimmtes Integral

Für viele Probleme der Statistik ist es erforderlich, Integrale zu berechnen. In der Regel handelt es sich hierbei um Integrale von Verteilungsfunktionen. Deshalb soll hier kurz angedeutet werden, wie die mathematische Schreibweise ist, ohne tiefer auf Probleme der Integralrechnung einzugehen:

Man kann das Integral als Fläche zwischen einer stetigen Funktion und der x-Achse beschreiben. Integrale werden für bestimmte Intervalle berechnet. Die Intervallgrenzen werden unterhalb und oberhalb des Integralzeichens angegeben.

Für die stetige Funktion $y = f(x)$ (lies: y ist eine Funktion von x) schreibt man das bestimmte Integral zwischen den Grenzen $x = a$ und $x = b$ wie folgt:

$$\int_a^b f(x) dx \quad (\text{lies: Integral von a bis b über } f(x) dx).$$

Diese Schreibweise ist entstanden aus folgender Vorstellung: Man schneide die Fläche unter der Funktion in lauter kleine Streifen, deren Breite mit Δx bezeichnet wird und deren Höhe entweder mit m_i oder M_i bezeichnet wird, je nachdem, ob die oberen Kanten der (rechteckigen) Streifen die Funktion von unten (Untersumme) oder von oben (Obersumme) berühren. Werden die Streifen unendlich schmal, dann gleichen sich Ober- und Untersumme an. Das Integral ergibt sich dann als Grenzwert folgender Summen:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n m_i \Delta x = \lim_{n \rightarrow \infty} \sum_{i=1}^n M_i \Delta x = F_b^a = \int_a^b f(x) dx$$

(47) Differentialrechnung

Der Ausgangspunkt der Fragestellung ist: Welche Steigerung (Gefälle) hat eine stetige Funktion. Stellt man die Steigerung als Verhältnis von zwei Strecken dar, dann kann sie näherungsweise durch den sog. **Differenzenquotienten** bestimmt werden.

Dieser lautet:

$$\frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0} = \tan \alpha \quad \alpha \text{ ist näherungsweise der Steigungswinkel}$$

Geht Δx gegen 0, dann geht $\frac{\Delta y}{\Delta x}$ in $\frac{dy}{dx}$ über.

Dieses ist der genannte Differentialquotient.

In diesem Quotienten heißt dy **Differential** der Funktion an der Stelle x_0 .

$$dy = f'(x_0) dx; \quad \frac{dy}{dx} = f'(x_0)$$

dx ist das Differential der abhängigen Variablen..

$f'(x)$ wird als 1. Ableitung der Funktion $f(x)$ bezeichnet.

Viele Funktionen haben mehr als eine Ableitung, d.h. man kann von der Ableitung wieder eine Ableitung bilden (1., 2., n-te Ableitung).

In Formeln drückt man das wie folgt aus:

$$d^n y = d(d^{n-1} y) = f^{(n)}(x) (dx)^n \quad \textbf{(Differential n-ter Ordnung)}$$

Ein Differential 3. Ordnung schreibe man:

$$d^3 y = d(d^2 y) = f'''(x) (dx)^3$$

Von den zahlreichen Differenzierungsregeln (Faktor-, Summen-, Produkt- und Quotientenregel) soll hier nur die Potenzregel mitgeteilt werden:

$$\frac{dx^n}{dx} = n x^{n-1} \quad (n \text{ reell})$$

Spezialfall: $\frac{de^x}{dx} = e^x$

Von Funktionen mit mehreren unabhängigen Variablen können partielle oder totale Ableitungen gebildet werden.

13. Kurze Einführung in SPSS 6.0.1 und in die wichtigsten Routinen.

13.1. Die „Fenster“ von SPSS

Diese Einführung setzt die Vertrautheit mit der Fenstertechnik in Windows 3.1. voraus. Der Einführung liegt die Version SPSS 6.0.1 zugrunde.

Wenn man in Windows 3.1. das Programm SPSS 6.0.1 lädt, dann zeigen sich drei Fenster, zwischen denen man durch den Befehl STRG+F6 hin- und her-wechseln kann.

13.1.1. Das Datenfenster

Die unten stehende Tabelle ist dem Datenfenster nachgebildet:

- In der obersten Zeile steht eine Datensatzbezeichnung mit Angabe des Pfades.
- In der zweiten Zeile steht die Fallnummer und die Bezeichnung der Variable, in der sich der Cursor befindet. Davon getrennt befindet sich rechts das Eingabefenster für weitere Daten.
- In der dritten Zeile stehen die bereits vergebenen Variablen-Namen. Kästchen, in denen noch kein Variablenname steht, sind nur mit „var“ bezeichnet.
- In der 4. Zeile steht der 1. Fall, der mit der laufenden Nr. 1 bezeichnet ist. Ihm sind alle Ausprägungen aller Variablen (Merkmale) in der 1. Zeile zuzuordnen, die auf ihn zutreffen.

Klickt man die Schaltfläche zweimal an, auf der VAR steht, dann öffnet sich ein weiteres Fenster, in dem die Möglichkeit besteht, zahlreiche Eingaben zur Definition der Variablen zu machen.

- **Variablenname:** Hier kann ein bis zu 8 Zeichen langer Variablenname eingegeben werden.
- **Variablenbeschreibung:** (Alle folgenden „Einstellungen“ sind jederzeit änderbar):
 - **Typ:** Varianten sind: Numerisch und andere (Tip: immer numerisch nehmen)
 - **Breite** (Tip: immer 8 nehmen)
 - **Dezimalstellen** (Tip: nur bei ganzzahligen Werten 0 nehmen, sonst die notwendige Zahl)
 - **Label:** Hier kann eine ausführliche Variablenbezeichnung eingegeben werden, ohne Beschränkung. Weiterhin können hier für bestimmte Ausprägungen Namen vergeben werden (**Wertelabels**). Missing Werte: Hier können Zahlen angegeben werden, bei deren Verarbeitung der Rechner intern den jeweiligen Fall als „fehlende Angabe“ bewertet. (Tip: Möglichst den Missing Wert 0 nehmen, wenn die Zahl 0 sonst nicht gebraucht wird).
 - **Spaltenformat:** Hier können Breite des Drucks und Ausrichtung der Zahlenangaben beim Druck (links, rechts, zentriert) gewählt werden. (Tip: Spaltenformat immer größer als das Zahlenformat wählen).

Es empfiehlt sich, vor Eingabe der Daten alle Variablen vollständig zu definieren.

Ist dies geschehen, kann die Dateneingabe beginnen. Dies geht wie folgt: Eingabe der Zahl ins Dateneingabefenster, dann die Enter-Taste drücken. Der Cursor springt dann automatisch zum nächsten Fall bei der gleichen Variable:

In der nebenstehenden Tabelle sind die ersten 10 Zahlen für unser Beispiel vom Tag der offenen Tür 1985 eingegeben. Der Cursor steht im Feld 10/1. In der zweiten Zeile ist angegeben, daß der Cursor auf dem 10 Fall der Variable „Besucher“ steht, und - im Eingabefeld- kann abgelesen werden, daß der Wert dieses Falles „24“ ist.

c:\(Pfad)Besuch85.sav		
10:Besucher	24 (Eingabefeld)	
(lfd.Nr.)	Besucher	var
1	26	
2	14	
3	3	
4	8	
5	21	
6	39	
7	4	
8	15	
9	1	
10	24	
...		

13.1.2. Häufigkeitsanalysen als Rechenbeispiel

Für alle Häufigkeitsanalysen kann man nun folgende Befehle im Menue von SPSS anklicken (die entsprechenden Fenster öffnen sich nacheinander: „Statistik“, „Deskriptive Statistik“, „Häufigkeiten“. Es öffnet sich bei „Häufigkeiten“ ein Fenster, in dem links die definierten Variablen stehen. Die blaue Markierung ist mit Hilfe der Maustaste bewegbar. Man kann hiermit die Namen jener Variablen auswählen, die

man in SPSS untersuchen will. Nach der Auswahl werden sie durch Anklicken des nach rechts gerichteten Pfeils in die Variablenliste für den jeweiligen Vorgang kopiert.

Weiterhin kann man in diesem Fenster bestimmen, ob man die Häufigkeitstabelle (nicht) anzeigen will und welche von den Kennwerten (hier „Statistiken genannt“) man berechnet haben möchte.

13.1.3. Das Syntaxfenster.

Wenn man dann alles ausgewählt hat, könnte man den „Befehl“, der von SPSS inzwischen intern aufgebaut worden ist, im „**Syntax-Fenster**“ anzeigen lassen. Hierzu klickt man auf die Schaltfläche „Befehl“. Es öffnet sich das Syntax-Fenster, und in diesem könnte beispielsweise folgender „Befehl“ stehen:

```
FREQUENCIES
  VARIABLES=besuch85
  /PERCENTILES= 5
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
  SKEWNESS KURTOSIS
  /HISTOGRAM NORMAL .
```

← (Beachten, daß hier ein Punkt steht)!!

Solche „Befehle“ kann man im Syntax-Fenster auch manuell vom Tastenfeld aus eingeben oder ändern, und man sollte sie in Forschungsprojekten zur eigenen Orientierung als **File abspeichern (Speichertechnik wie in Windows generell üblich)**, damit man später noch sehen kann, **was** man mit **welchen Daten wann** gerechnet hat. Die Sprache dieser Befehle („Befehls-Syntax“) ist weitgehend (aber nicht vollständig) noch mit der alten unter DOS gültigen Sprache identisch. Die englische Terminologie, die nicht in allen SPSS-Versionen vollständig ins Deutsche übertragen wurde, bedeutet für dieses Beispiel im Deutschen:

SPSS-Ausdruck	Bedeutung im Deutschen
FREQUENCIES :	Führe eine Häufigkeitsanalyse durch
VARIABLES=besuch85	für die Variable mit dem Label „besuch85“
/PERCENTILES= 5	Berechne den Prozentrang 5.
/STATISTICS=	Berechne folgende Kennwerte
STDDEV	Standardabweichung
VARIANCE	Varianz
RANGE	Eingriffsspielraum
MINIMUM	kleinsten Meßwert
MAXIMUM	größten Meßwert
SEMEAN	Standardfehler des arithmetischen Mittels
MEAN	arithmetisches Mittel
MEDIAN	Median
MODE	den kleinsten Modus
SKEWNESS	die Schiefe
KURTOSIS	den Exzess
/HISTOGRAM NORMAL	Zeichne ein Histogramm und lege hierhin den Verlauf einer Normalverteilungs-Kurve mit den entsprechenden Kennwerten für die

Standardabweichung und das arithmetische Mittel.
--

Jeder SPSS-Befehl wird immer mit einem PUNKT enden (Siehe den Befehltext oben). Ohne Punkt ist er nicht ausführbar. Um ihn auszuführen, markiere ich mit der Maus den gesamten Befehltext. Dann klicke ich auf die Schaltfläche „Ausführen“.

13.1.4. Das Ausgabe-Fenster und das Graphik-Karussell

13.1.4.1. Das Ausgabe Fenster

Es öffnen sich jetzt das „**Ausgabe-Fenster**“ und, wenn eine Graphik angefordert wird, zeigt sich das Symbol für das Graphik - Karussell. Im Ausgabefenster könnte folgendes stehen:

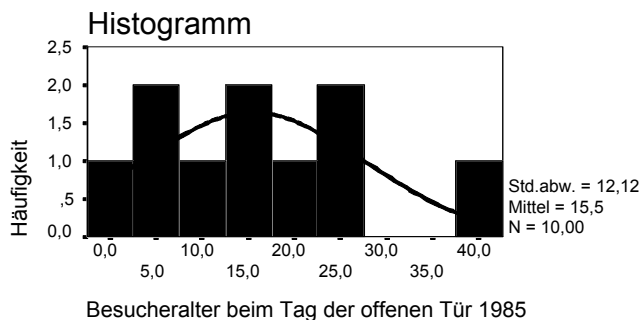
BESUCH85 Besucheralter beim Tag der offenen Tür 1					
Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	1	1	10,0	10,0	10,0
	3	1	10,0	10,0	20,0
	4	1	10,0	10,0	30,0
	8	1	10,0	10,0	40,0
	14	1	10,0	10,0	50,0
	15	1	10,0	10,0	60,0
	21	1	10,0	10,0	70,0
	24	1	10,0	10,0	80,0
	26	1	10,0	10,0	90,0
	39	1	10,0	10,0	100,0
	Total	10	100,0	100,0	
Hi-Res Chart 2:Histogramm vom Besucheralter beim Tag der Offenen Tür 1985					
Mean	15,500	Std err	3,833	Median	14,500
Mode	1,000	Std dev	12,122	Variance	146,944
Kurtosis	-,169	Skewness	,621	Range	38,000
Minimum	1,000	Maximum	39,000		
* Multiple modes exist. The smallest value is shown.					
Percentile	Value				
5,00	1,000				
Valid cases	10	Missing cases	0		

Hierin bedeuten:	Value-Label:	Werte-Etikett (Labels wurden hier nicht vergeben)
	Value:	Wert aus der Datenmatrix
	Frequency	Beobachtete Häufigkeit
	Percent	Rohprozente (wenn fehlende Fälle auftreten, werden sie mit-prozentuiert.
	Valid-Percent	Gültige Prozente (wenn fehlende Fälle auftreten, werden sie nicht mit-prozentuiert.
	Cum Percent	Von unten nach oben aufkumulierte gültige Prozente.

Es empfiehlt sich, auch die Inhalte des Ausgabefensters innerhalb von Projekten abzuspeichern und sich eine systematische Liste der gespeicherten Ausgaben mit den Filenamen und anderen Angaben anzulegen.

13.1.4.2. Das Graphik-Karussell

Klickt man nun das Graphik-Karussell-Symbol an, dann könnte im Graphikfenster folgende Graphik erscheinen.



Die obige Graphik kann im Graphik-Fenster noch in verschiedenster Weise bearbeitet werden. Da man innerhalb einer SPSS-Sitzung mehrere Graphiken (auf einmal oder nacheinander) erzeugen kann, die man nacheinander „durchblättern“ kann, wird der Graphik - Editor auch „Graphik-Karussell“ genannt.

Wie wir sehen, werden hier mit wenigen Handgriffen alle die komplizierten Rechnungen ausgeführt, die oben im Skriptum theoretisch beschrieben wurden.

13.1.5. Übertragung von Ergebnissen aus SPSS nach WinWord

Die Übertragung von Ergebnissen von SPSS nach Word ist eine der wichtigsten Arbeiten, mit denen Studierende vertraut sein sollten. Im Prinzip wird wie folgt verfahren:

SPSS und WINWORD werden nacheinander geöffnet. Über den Befehl STRG+ESC öffnet sich ein Fenster, in dem alle geöffneten Programme angezeigt werden. Man stellt mit dem Cursor die blaue Leiste auf jenes Programm, zu dem man wechseln will und wählt die Schaltfläche „wechseln zu“. Der Rechner wechselt in das gewünschte Programm. Nach dieser Prozedur ist es in der Regel möglich, zwischen den zwei Programmen, SPSS und WinWord durch ALT+TAB zu wechseln.

In den Fenstern von SPSS markiert man für die Übertragung nach Word den gewünschten Text bzw. die gewünschte Graphik durch anklicken und ev. Festhalten der linken Maustaste. Dann gibt man den Befehl: „Bearbeiten“ - „Kopieren“ (durch Anklicken der Optionen im Menue Bearbeiten). Dann wechselt man nach Word. Dort stellt man *vor dem Einfügen von Texten, besonders von Tabellen*, die **Schriftart auf Courier New**. Danach wählt man die Befehle Bearbeiten - Einfügen.

Bei **Graphiken** verfährt man ähnlich:

Man wählt in SPSS „Bearbeiten“ „Graphik kopieren“
geht dann nach Word und wählt „Bearbeiten“ - „Inhalte einfügen“ - „Graphik“ - „OK“

Wenn man hierbei Probleme bekommt (die Graphik wird in Word nicht angezeigt), sollte man andere Wege durch Ausprobieren suchen.

Wichtig ist, daß Graphiken vor der Übertragung nach Word so vollständig bearbeitet sind, daß eine Nachbearbeitung in Word unnötig ist. Hierbei ist beispielsweise sinnvoll, keine Graphik-Titel oder -Nummern in SPSS für Graphiken zu vergeben, da man diese in Word leicht hinzusetzen kann, ohne die Graphik nachbearbeiten zu müssen. Eine Möglichkeit, die Graphik nachzubearbeiten, die für den Umbruch von Texten sehr praktisch ist, besteht allerdings auch noch nach der Übertragung nach Word: Man kann durch Ziehen der Kanten eine Graphik senkrecht, schräg und waagrecht verkleinern oder vergrößern und so dem Seitenformat gut anpassen.

13.2. Praktische Fälle

13.2.1. Untersuchung von Mittelwertsdifferenzen zweier unabhängiger Stichproben.

Gegeben sei das Ergebnis eines sportlichen Tests, anlässlich dessen je eine Stichprobe männlicher und weiblicher Mitglieder einer Sportschule um die Wette laufen. Die Datenmatrix hat folgende Gestalt:

Laufende Nr.	Zeit in Sekunden	Geschlecht (1=m, 2=w)
1	13,1	1
2	12,0	1
3	14,1	2
4	12,6	1
5	15,5	2
6	11,5	2
7	14,4	2
8	13,2	1
9	13,1	1
10	12,0	1
11	12,7	1
12	12,9	2
13	13,5	1
14	12,4	2
15	12,6	1
16	13,0	2
17	15,6	2
18	14,0	1
19	15,6	2
20	14,7	2
21	14,3	2
22	13,4	1
23	14,4	1
24	15,5	2
25	13,9	2
26	15,2	2
27	15,2	2
28	14,7	1
29	12,7	1
30	12,0	1
31	14,0	2
32	15,5	2
33	11,8	2
34	13,6	2
35	12,2	1
36	13,8	2
37	13,0	1
38	13,8	1
39	14,4	1
40	14,3	1

Wir wählen in SPSS aus dem Menue die Befehle:

„Statistik“, „Mittelwertvergleiche“, „t-Test bei unabhängigen Stichproben“.

Wir wählen die Variable „Sekunden“ als „Testvariable“, indem wir sie in das entsprechende Kästchen kopieren (blau hinterlegen, Pfeil anklicken), und wir wählen das Geschlecht als „Gruppierungsvariable“, in dem wir den Variablennamen blau hinterlegen und dann den „Pfeil“ anklicken. Es öffnet sich ein neues Fenster, in dem der Variablenname und ein Klammerausdruck mit 2 Fragezeichen steht. Wir drücken die Schaltfläche „Gruppen definieren“ und schreiben bei „Gruppe 1“ in den vorgesehenen Kasten eine „1“, bei Gruppe 2 eine „2“ hinein. Dann klicken wir die Schaltfläche „Weiter“ an und drücken die Schaltfläche „Befehl“. Im Syntax-Fenster erscheint folgender Befehl:

T-TEST

```
GROUPS=geschlt(1 2)
/MISSING=ANALYSIS
/VARIABLES=sekunden
/CRITERIA=CIN(.95) .
```

Wir markieren diesen Befehltext und klicken auf die Schaltfläche „Ausführen“. Im Ausgabefenster erscheint:

Variable	Number of Cases	Mean	SD	SE of Mean
-----	-----	-----	-----	-----
SEKUNDEN				
männlich	20	13,1788	,860	,192

weiblich	20	14,1288	1,277	,286
----------	----	---------	-------	------

Mean Difference = -,9500

Levene's Test for Equality of Variances: $F = 2,483$ $P = ,123$
t-test for Equality of Means 95%

Variances	t-value	df	2-Tail Sig	SE of Diff	CI for Diff
Equal	-2,76	38	,009	,344	(-1,647; -,253)
Unequal	-2,76	33,30	,009	,344	(-1,651; -,249)

Diesem „Output“ können wir entnehmen, daß zwischen den Jungen und den Mädchen ein auf dem 5 % Niveau im zweiseitigen Test signifikanter Schnelligkeits-Unterschied besteht, der im Durchschnitt bei 0,95 Sekunden liegt (Mean Difference). Daß der Unterschied signifikant ist, erkennen wir daran, daß für „2-Tail Sig“ ein Wert von $0,009 < 0,05$ (Signifikanzniveau) angegeben ist (siehe folgende Tabelle).

Im einzelnen bedeuten:

SPSS-Ausdruck	Bedeutung im Deutschen
SD	Standardabweichung
SE of Mean	Standardfehler
Mean Difference	Durchschnittliche Differenz der Gruppen
Levenes Test for Equality of Variances	Levenes Test der Varianzhomogenität (ist hier $p < 0,05$, wird Varianzhomogenität angenommen)
2-Tail sig	Niveau der Irrtumswahrscheinlichkeit bei zweiseitigem Test, bei der die Grenze zur Ablehnung der Nullhypothese liegt. Oben ist der Wert 0,009. Da $0,009 < 0,05$ ist, wird die Nullhypothese auf dem 5 % Niveau abgelehnt, daß in der Grundgesamtheit keine Differenz besteht
SE of Diff	Standardfehler der Gruppendifferenz
CI for Diff	sog. Konfidenzintervall für die Gruppendifferenz (hier: bei $\alpha = 5\%$)
Equal - Unequal	Ergebnisse für den Fall homogener bzw. inhomogener Varianzen

13.2.2. Untersuchung von Mittelwertsdifferenzen zweier abhängiger Stichproben.

Wir gehen von folgender Datenmatrix aus, in der die Daten von Sportlern vor und nach einem Training im 100m Lauf enthalten sind:

Laufende Nr.	Sekunden vor dem Training	Sekunden nach dem Training
1	13,406	13,424
2	14,399	14,144
3	11,844	11,410
4	13,561	12,658
5	13,226	12,621
6	13,040	12,683
7	14,520	14,626
8	12,861	12,365
9	12,531	12,675
10	11,501	11,169
11	13,335	13,154
12	10,690	10,391
13	11,412	11,449
14	13,408	12,740
15	13,487	12,960
16	13,568	13,795
17	12,235	11,774
18	12,780	12,398
19	11,834	12,027
20	11,727	11,756

Wir wählen im SPSS-Menue den Befehl „Statistik“, „Mittelwertvergleiche“, „T-Test bei gepaarten Stichproben“ und kopieren mit dem Pfeil die beiden Variablen „Vortrai“ und „Nachtrai“ in das Fenster für die gepaarten Variablen.

Wir klicken die Schaltfläche „Befehl“ und erhalten im Syntaxfenster folgendes:

```
T-TEST
  PAIRS= vortrai WITH nachtrai (PAIRED)
  /CRITERIA=CIN(.95)
  /FORMAT=LABELS
  /MISSING=ANALYSIS.
```

Diesen Befehl markieren wir und klicken auf die Schaltfläche

„Ausführen“. Im Ausgabefenster erscheint:

- - - t-tests for paired samples - - -						
Variable	Number of pairs	Corr	2-tail Sig	Mean	SD	SE of Mean
VORTRAI	20	,952	,000	12,7681	1,022	,229
NACHTRAI				12,5109	1,036	,232

Paired Differences						
Mean	SD	SE of Mean		t-value	df	2-tail Sig
-----				-----		
,2573	,318	,071		3,62	19	,002
95% CI (,108; ,406)						

Der Output zeigt, daß das Training im Durchschnitt eine Leistungsverbesserung von ¼ Sekunde gebracht hat, und daß dieses Ergebnis auf dem 1% Niveau von α signifikant ist (untere Zeile: 2-tail Sig = 0,002 < 0,01. Die

Angabe zur „2-tail Sig.“ in der oberen Zeile bezieht sich auf den dort angegebenen Korrelationskoeffizienten, der aber in diesem Zusammenhang immer sehr hoch sein wird, insofern ist diese Angabe banal.

13.2.3. Einfache Varianzanalyse

Nehmen wir an, wir hätten Daten von 5 Gruppen, die miteinander um die Wette laufen, und bei denen sich nur die Trainingsbedingungen unterscheiden, in der unten stehenden Word-Tabelle: Also 5 voneinander unabhängige Stichproben. Die Gruppennummer muß als selbständige Variable aufgefaßt werden: Laufleistung in Sekunden pro 100m in 5 Gruppen

Gruppe 1	Gruppe 2	Gruppe 3	Gruppe 4	Gruppe 5
13,406	13,424	13,325	13,437	13,063
14,399	14,144	12,451	13,189	13,527
11,844	11,410	10,392	12,563	10,140
13,561	12,658	12,919	13,979	13,390
13,226	12,621	12,173	12,369	12,232
13,040	12,683	12,376	12,587	13,633
14,520	14,626	12,679	12,886	13,813
12,861	12,365	12,314	12,777	12,027
12,531	12,675	14,152	12,477	12,305
11,501	11,169	12,535	13,066	14,511
13,335	13,154	12,499	12,768	12,378
10,690	10,391	10,066	10,053	13,511
11,412	11,449	10,286	12,294	11,993
13,408	12,740	13,015	13,810	12,950
13,487	12,960	14,303	15,124	14,871
13,568	13,795	13,344	11,117	11,954
12,235	11,774	10,864	11,827	13,716
12,780	12,398	12,668	13,791	14,372
11,834	12,027	11,222	12,274	12,738
11,727	11,756	12,118	11,583	10,996

Zunächst einmal müssen wir die oben in der Word-Tabelle nebeneinander ausgedruckten Daten im Eingabefenster von SPSS in **einer Spalte** anordnen, denn es handelt sich ja um Daten einer Variable, der Laufleistung. Dazu kopieren wir die Daten aus den verschiedenen Spalten der obigen Word-Tabelle in eine einzige Spalte des Eingabefensters von SPSS, der wir zuvor den Variablennamen Laufleistung („Laufleis“) zuweisen, sowie 6 Stellen für die Zahlenangabe, mit 3 Nachkommastellen.

Der **Kopiervorgang** funktioniert wie folgt: Nehmen wir an, die Fälle oben liegen in einer Word-Tabelle (wie oben) vor. Wir markieren mit dem Cursor alle Fälle in der jeweiligen Spalte (=Variable), die wir kopieren wollen. Dann geben wir den Befehl „STRG+C“. Dann gehen wir mit dem Cursor in das erste freie Feld der Variable „Laufleis“ des SPSS-Eingabefensters. Dort geben wir den Befehl „STRG+V“. Der Rechner kopiert die Daten in die entsprechende Spalte. Sodann müssen wir die jetzt untereinander stehenden Fälle mit der Nummer der Gruppe bezeichnen, aus der sie stammen. Hierzu bilden wir die Variable: „Gruppnr“. Das Eingabefenster in SPSS sieht dann wie folgt aus (gezeigt werden nur die oben fettgedruckten Fälle 15 bis 25, denn bei den Fällen 15-20 ist die Gruppennummer noch 1, ab Fall 21 wächst sie an auf 2:

Lauf. Nr.	Grupp nr	Laufleis
15	1	13,487
16	1	13,568
17	1	12,235
18	1	12,780
19	1	11,834
20	1	11,727
21	2	13,424
22	2	14,144
23	2	11,410
24	2	12,658
25	2	12,621
.	.	.
.	.	.

Haben wir alle Fälle in der neuen Variable „Laufleistung“ angeordnet, dann geben wir folgende Befehle: „Statistik“, „Mittelwertvergleiche“, „einfaktorielle ANOVA“. Dort weisen wir „Laufleis“ der Gruppe der **abhängigen** Variablen zu, dem „Faktor“ weisen wir „Gruppnr“ zu. Wir klicken auf die Schaltfläche „Bereich definieren“ und weisen dem „Minimum“ die Gruppennummer 1, dem „Maximum“ die Gruppennummer 5 zu. Im „Optionsfenster“ wählen wir noch „Deskriptive Statistiken“ und „Varianzhomogenität“. Dann klicken wir auf die Schaltflächen: „Weiter“ und „Befehl“.

Im Syntax-Fenster erscheint das Folgende:

```
ONEWAY
  laufleis BY gruppnr(1 5)
  /HARMONIC NONE
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /FORMAT NOLABELS
  /MISSING ANALYSIS.
```

Diese Befehle führen wir aus. Im Ausgabefenster erscheint jetzt das Folgende:

```

- - - - - O N E W A Y - - - - -
Variable  LAUFLEIS
By Variable  GRUPPNR      Gruppennummer

                                Analysis of Variance
                                Sum of          Mean
                                Squares         Squares
Source      D.F.
Between Groups      4      4,6608      1,1652      ,9495      ,4390
Within Groups      95     116,5749      1,2271
Total             99     121,2357

Group      Count  Mean      Standard      Standard
          20  12,7681  Deviation    Error
Grp 1      20  12,5109  1,0223      ,2286
Grp 2      20  12,2851  1,0360      ,2317
Grp 3      20  12,6986  1,1884      ,2657
Grp 4      20  12,9060  1,0991      ,2458
Grp 5      20  12,9060  1,1818      ,2643
Total     100  12,6337  1,1066      ,1107
          100  12,6337  1,1066      ,1107
          12,4142  TO      12,8533

GROUP      MINIMUM      MAXIMUM
Grp 1      10,6897      14,5199
Grp 2      10,3914      14,6256
Grp 3      10,0659      14,3026
Grp 4      10,0532      15,1242
Grp 5      10,1399      14,8706
TOTAL      10,0532      15,1242

- - - - - O N E W A Y - - - - -
```

Levene Test for Homogeneity of Variances

```

Statistic      df1      df2      2-tail Sig.
,1908          4      95      ,943
```

Der Output zeigt, daß zwischen den 5 Gruppen kein signifikanter Unterschied besteht. Die Varianzen sind homogen (2-tail Sig = 0,943 > 0,05), der F-Wert ist nicht auf dem 5 % Niveau signifikant (,F-Prob. = 0,4390 > 0,05). Die Nullhypothese, daß alle Gruppen aus der gleichen Grundgesamtheit entnommen sein können, wird angenommen. Wäre das Ergebnis signifikant, könnten wir aus den Angaben über die Gruppenmittelwerte entnehmen, welche Gruppen die größten Abstände untereinander aufweisen und dann - durch Ausschluß von extremen Gruppen - prüfen, ob auch für Teilmengen der Gruppen noch signifikante Unterschiede der Mittelwerte bestehen.

13.2.4. Zweidimensionaler CHI-Quadrat -Test mit einer Vierfelder-Tabelle.

Gegeben seien 46 Befragte, Frauen und Männer, die eine Angabe darüber machen, ob sie sehr selten (value=1) oder gelegentlich (value=2) Horoskope lesen. Die Frage ist, ob es eine größere Bereitschaft bei Frauen als bei Männern gibt, Horoskope zu lesen.

Das Eingabefenster (die ersten 10 Fälle) sieht hierbei wie folgt aus:

Geschlecht	Horoskop	Wir wählen über das Menue die Befehle „Statistik“, „Deskriptive Statistik“, „Kreuztabellen“, und weisen mit dem „schwarzen Pfeil“ den Zeilen die Variable „Geschlecht“ und den Spalten die Variable „Horoskop“ zu. Über die Schaltfläche „Statistiken“ öffnen wir ein neues Fenster in dem wir „Chiquadrat“ und „Phi und Cramers V“ durch Anklicken der entsprechenden Kästchen auswählen. Dann klicken wir die Schaltfläche „Weiter“ an und wählen über die Schaltfläche „Zellen“ nach dem gleichen Verfahren „Beobachtete“ und „Erwartete“ Werte, sowie „korrigierte standardisierte“ Residuen. Danach klicken wir auf „Weiter“ und „Befehl“. Im Syntaxfenster zeigt sich folgender Text:
1	1	
2	2	
1	1	
1	2	
2	2	
2	2	
1	1	
1	2	
1	2	
2	2	
.	.	CROSSTABS
.	.	/TABLES=geschl BY horoskop
.	.	/FORMAT= AVALUE NOINDEX BOX LABELS TABLES
.	.	/STATISTIC=CHISQ PHI
.	.	/CELLS= COUNT EXPECTED ASRESID .

Diese Befehle führen wir aus. Im Ausgabefenster erscheint u.a. folgendes (weitere, hier unwesentliche Angaben wurden nicht abgedruckt:

GESCHL Geschlecht by HOROSKOP Lesen von Horoskopen
HOROSKOP

		Count			
GESCHL	Exp Val	sehr	gelegent		
	Adj Res	selten	lich	Row	
		1	2	Total	
	-----+	-----+	-----+		
männlich	1	12	16	28	
		8,5	19,5	60,9%	
		2,3	-2,3		
weiblich	2	2	16	18	
		5,5	12,5	39,1%	
		-2,3	2,3		
		-----+	-----+		
Column		14	32	46	
Total		30,4%	69,6%	100,0%	

Chi-Square		Value	DF	Significance
Pearson		5,21542	1	,02239
Continuity Correction		3,82376	1	,05053
Minimum Expected Frequency		5,478		
Statistic		Value	Significance	
Phi		,33672	,02239 *1	

*1 Pearson chi-square probability
Number of Missing Observations: 0

Der Output zeigt, daß es einen auf dem 5% Niveau signifikanten Zusammenhang zwischen Geschlecht und der Bereitschaft gibt, Horoskope zu lesen. Die Nullhypothese könnte mit 5 % Irrtumswahrscheinlichkeit abgelehnt werden. Wie der Zusammenhang inhaltlich gestaltet ist, kann an der Differenz von beobachteten und erwarteten Häufigkeiten bzw. an den aus diesem Verhältnis pro Zelle errechneten angepaßten, standardisierten Chiquadrat-Residuen abgelesen werden. Die Tabelle zeigt positive Werte hierfür bei den Kombinationen von männlich+sehrselten bzw. weiblich+gelegentlich. Frauen haben hiernach eine größere Affinität zur Lektüre von Horoskopen.

Wir können die Beträge der **angepaßten, standardisierten Chiquadratresiduen wie z-Werte interpretieren**: Sofern z.B. ein z-Wert den Betrag von +2,58 annimmt, können wir feststellen, daß die Wahrscheinlichkeit größerer Z-Werte gleich 0,005 ist. Eine Nullhypothese würde daher bei zweiseitigem Test auf dem 1%Niveau abgelehnt werden müssen, wenn $z > 2,58$. Ist ein angepaßtes, standardisiertes Chiquadratresiduum größer (bzw. kleiner) als $\pm 2,58$, dann kann angenommen werden, daß eine Über (bzw. Unter-) Repräsentation in der entsprechenden Zelle vorliegt, mit einer Irrtumswahrscheinlichkeit von 1% (zweiseitig).

In der obigen Tabelle können wir feststellen, daß die Männer bei „gelegentlich“ unterrepräsentiert sind, die Frauen aber überrepräsentiert sind, mit einer Irrtumswahrscheinlichkeit pro Zelle von 2,2 % (bei $z = 2,3$ ist $\alpha = 0,022$ bei beidseitigem Test). Das Signifikanzniveau für Phi entspricht in der Vierfeldertabelle genau dieser Größe mit $p = 0,022$. Hätte die Tabelle mehr als 4 Zellen, würde statt Phi Cramers V zu interpretieren sein. SPSS druckt immer beides gleichzeitig ab.

Im einzelnen bedeuten:

In der Tabelle:	sind die Zelleninhalte frei wählbar. Welche ausgewählt sind, wird links oben, vor der Kopfzeile, angegeben:
Count	Absolute Häufigkeit (hier: Oberste Zahl in jeder Zelle)
Expected	erwartete Häufigkeit (hier: zweite Zahl in jeder Zelle)
Adj. Res	Angepaßte, standardisierte Chiquadrat-Residuen (hier:dritte Zahl in jeder Zelle)
Row	Zeile
Column	Spalte
Total	Summe
Pearson	Chiquadrat-Wert nach Pearson
Continuity Correction	Chiquadrat-Wert nach Yates korrigiert
Minimum expected Frequency	kleinster Erwartungswert in der Tabelle

Besonderheiten anderer Fälle: Haben wir mehr als 4 Felder in der Tabelle, dann ist - wie gesagt - Cramers V und nicht mehr PHI zu interpretieren. Es ist bei allen solchen Tabellen darauf zu achten, ob in mehr als 20% der Zellen der Erwartungswert kleiner als 5 ist. Sollte dies der Fall sein, ist durch Verzicht auf Kategorien die Tabelle bei gleichem n so zu verkleinern, daß doch noch V berechnet und interpretiert werden kann.

Ist bei Vierfelder-Tabellen die Fallzahl gering, so daß in mehr als 20 % der Zellen der Erwartungswert kleiner als 5 ist, dann druckt SPSS noch die **Ergebnisse für Fisher's exakten Test** ab. Dieser ist auch bei kleinen Fallzahlen zuverlässig.

13.2.5. Zweidimensionaler CHI-Quadrat -Test mit Kontrollvariable.

Der Fall: In einer empirischen Untersuchung von VolkshochschulleitnehmerInnen wird gefragt, wie wichtig den Teilnehmern das Weiterlernen im Beruf ist. Sie können hierauf mit „selten...“, „gelegentlich...“, „öfters...“ und „sehr wichtig“ antworten (vierstufige Ordinalskala). Weiterhin werden sie nach Alter und Geschlecht gefragt. Es soll nun untersucht werden, ob bei den Volkshochschulleitnehmern Unterschiede zwischen den Geschlechtern bestehen, was die Attraktivität beruflichen Weiterlernens in verschiedenen (5) Altersgruppen betrifft. Im Datensatz enthalten sind die Variablen: „Alter“, „V33berw“ (= Bedeutung des beruflichen Weiterlernens) und

Statistic	Value	Significance
Cramer's V	,11689	,00104 *1
Kendall's Tau-b	-,10386	

V33BERW Weiterlernen im Beruf by ALTER
Controlling for V61GEND Geschlecht Value = 2 weiblich

		ALTER					
		Count					
		Row Pct					
		Col Pct					
		Adj Res	1	2	3	4	5
V33BERW							Row Total
	1		151	157	128	155	242
unwichtig			18,1%	18,8%	15,4%	18,6%	29,1%
			44,2%	36,8%	39,5%	44,5%	69,7%
			-1,0	-4,7	-2,8	-,9	9,6
	2		31	35	35	35	15
selten wichtig			20,5%	23,2%	23,2%	23,2%	9,9%
			9,1%	8,2%	10,8%	10,1%	4,3%
			,5	-,2	1,7	1,2	-3,1
	3		31	59	38	33	33
gelegentlich wic			16,0%	30,4%	19,6%	17,0%	17,0%
			9,1%	13,8%	11,7%	9,5%	9,5%
			-1,2	2,3	,6	-,9	-,9
	4		51	59	36	29	17
oefters wichtig			26,6%	30,7%	18,8%	15,1%	8,9%
			14,9%	13,8%	11,1%	8,3%	4,9%
			2,8	2,4	,2	-1,6	-3,9
	5		78	117	87	96	40
sehr wichtig			18,7%	28,0%	20,8%	23,0%	9,6%
			22,8%	27,4%	26,9%	27,6%	11,5%
			-,3	2,3	1,6	2,1	-5,8
Column			342	427	324	348	347
Total			19,1%	23,9%	18,1%	19,5%	19,4%
							1788
							100,0%

Chi-Square	Value	DF	Significance
Pearson	119,65653	16	,00000
Minimum Expected Frequency	27,362		

Statistic	Value	ASE1	Val/ASE0	Approximate Significance
Cramer's V	,12935			,00000 *1
Kendall's Tau-b	-,12761	,01917	-6,66177	

Number of Missing Observations: 380

Der Output zeigt, daß bei den Männern (erste Tabelle) nur zwei Zellen der Tabelle auffällig sind:

- Die letzte Zelle in der ersten Zeile, aus der hervorgeht, daß ältere Männer stark überrepräsentiert sind bei der Aussage, daß berufliche Weiterbildung für sie unwichtig ist.
- Die letzte Zelle in der letzten Zeile, aus der hervorgeht, daß ältere Männer stark unterrepräsentiert sind bei der Aussage, daß berufliche Weiterbildung für sie sehr wichtig ist,
- Die dritte Zelle in der vierten Zeile, aus der hervorgeht, daß Männer im mittleren Alter leicht überrepräsentiert sind bei der Aussage, daß berufliche Bildung für sie öfters wichtig ist.

Insgesamt gibt es nur eine zwar signifikante, aber schwache negative Korrelation zwischen Alter und Bedürfnis nach beruflicher Weiterbildung in der Volkshochschule bei Männern.

Bei Frauen (zweite Tabelle) ist das Ergebnis ähnlich:

Der Output zeigt, daß bei den Frauen (zweite Tabelle) 5 Zellen der Tabelle auffällig sind:

- Die letzte Zelle in der ersten Zeile, aus der hervorgeht, daß ältere Frauen stark überrepräsentiert sind bei der Aussage, daß berufliche Weiterbildung für sie unwichtig ist.
- Die letzte Zelle in der letzten Zeile, aus der hervorgeht, daß ältere Frauen stark unterrepräsentiert sind bei der Aussage, daß berufliche Weiterbildung für sie sehr wichtig ist,
- Aus der ersten und zweite Zelle in der vierten Zeile geht hervor, daß jüngere Frauen überrepräsentiert sind bei der Aussage, daß für sie berufliche Bildung „öfters wichtig ist“.
- Aus der zweiten Zelle der letzten Zeile geht hervor, daß für jüngere Frauen bei der Aussage überrepräsentiert sind, daß für sie berufliche Weiterbildung „sehr wichtig“ ist.

Insgesamt gibt es nur eine zwar signifikante, aber schwache negative Korrelation zwischen Alter und Bedürfnis nach beruflicher Weiterbildung in der Volkshochschule bei den Frauen, die aber etwas stärker ausgeprägt ist als bei den Männern. Tendenziell gilt bei beiden Geschlechtern: Die Nachfrage nach beruflicher Weiterbildung sinkt mit dem Alter.

13.2.6. Korrelations- und Regressionsanalyse

Wir haben den Fall zweier miteinander zusammenhängender Variablen, die beide intervallskaliert sind: Die Menge der km, die ein Autofahrer zum Dienstort benötigt, und die Menge der Punkte in Flensburg wegen irgendwelcher Verkehrsdelikte. Dies alles geschieht in einer Stadt mit extrem schlechter Verkehrsbeschilderung und hoher Kontrollichte der Polizei. Es soll der Nachweis geführt werden, daß die Menge der Punkte in Flensburg vor allem von der Entfernung zum Dienstort abhängt. Das Eingabefenster sähe wie folgt aus (die ersten 4 Fälle):

KmzumD	Punkte
19,0	7
25,0	7
13,0	4
18,0	4
...	...

Wir wählen aus dem Menue folgende Befehle aus: „Statistik“, „Regression“, „linear“; wir weisen in dem dann offenen Fenster „Punkte“ der „abhängigen Variable“ zu, „KmzumD“ der „unabhängigen Variable“. In der Schaltfläche „Statistiken“ fordern wir noch „univariate Statistiken“ an. Das klicken wir die Schaltfläche „Befehl“ an. Im Syntax-Fenster erscheint dann folgendes:

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
```

```
/DEPENDENT punkte
/METHOD=ENTER kmzumd.
```

Wir führen diesen Befehl aus und erhalten folgenden Output im Ausgabefenster (alles hier nicht Relevante ist weggelassen):

```

* * * * M U L T I P L E   R E G R E S S I O N   * * * *
Listwise Deletion of Missing Data
      Mean   Std Dev   Label
PUNKTE    6,174    1,568   Punkte in Flensburg
KMZUMD    19,783    5,257   Entfernung zum Dienstort in km
N of Cases =      46
Correlation, 1-tailed Sig:
      PUNKTE      KMZUMD
PUNKTE    1,000          ,706
          ,          ,000
KMZUMD    ,706    1,000
          ,000          ,

Variable          B          SE B          Beta          T   Sig T
KMZUMD          ,210431    ,031854    ,705658    6,606    ,0000
(Constant)      2,011046    ,651556          3,087    ,0035

```

Wir können dem Output entnehmen, daß es eine deutliche Korrelation zwischen den beiden Variablen gibt, wobei 49,7 % der Varianz der abhängigen Variable durch die unabhängige Variable erklärt werden. Der

Korrelationskoeffizient $r = 0,7056$ kann dem Output an zwei Stellen entnommen werden: bei

„Correlation, 1-tailed Sig“ und bei „Beta“.

Die **Regressionskoeffizienten** entnehmen wir bei $B = 0,210$ (und erhalten die Steigung in der 1.

Regressionsgerade). Der Abschnitt auf der y-Achse ist gleich $a = 2,011$. Wir finden die Angabe hinter

„Constant“. Nach diesen Angaben könnten wir die Regressionsgerade zeichnen (siehe unten die entsprechende Graphik). Weiterhin stellen wir fest, daß der Korrelationskoeffizient signifikant ist ($T = 6,606$), $p = 0,000$. Überdies erhalten wir anfangs im Output noch Angaben über arithmetische Mittel und Standardabweichungen der untersuchten Variablen.

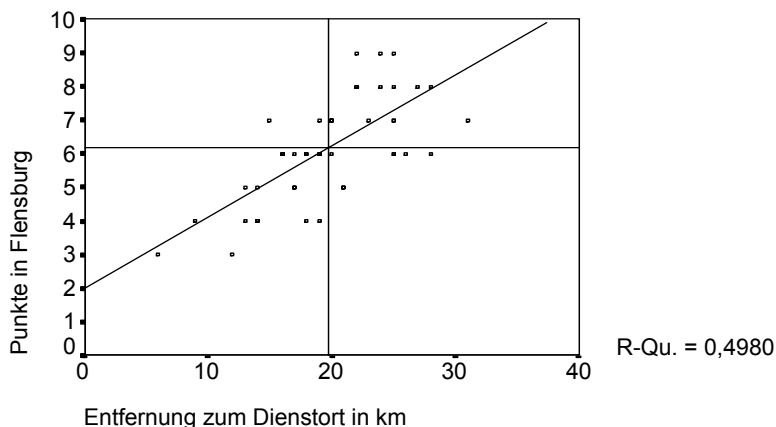
Würden wir eine Graphik der Beziehung haben wollen, wäre ein „Scatterplot“ geeignet. Wir erhalten diese Graphik über die Befehlsfolge: „Graphik“, „Scatterplot“, „Einfach“, „Definieren“ und weisen dort der x-Achse die Variable „KumzumD“ zu und der y-Achse die Variable „Punkte“. Dann klicken wir auf die Schaltfläche „Befehl“ und erhalten im Syntaxfenster:

```
GRAPH
/SCATTERPLOT(BIVAR)=kmzumd WITH punkte
/MISSING=LISTWISE .
```

Nachdem wir diesen Befehl ausgeführt haben, können wir durch Anklicken des Symbols „Graphikkarussell“ folgende Graphik im Graphikeditor herstellen: Zunächst erscheint die Graphik ohne die Regressionsgerade und ohne die senkrechten Linien, die die Position der arithmetischen Mittel von x und y markieren. Aber mit Hilfe des Menues „Bearbeiten“ „Graphik“ etc. lassen sich diese Zusätze zur Erstfassung der Graphik herstellen.

Wichtig für die Praxis ist noch, daß man Farben durch Schraffuren ersetzen kann, daß man den Darstellungsspielraum der Achsen wählen kann (um zum Beispiel den Nullpunkt mit darzustellen), und daß man auch Überschriften hinzufügen kann etc.:

Zusammenhang zwischen der Entfernung zum Arbeitsort und den „Punkten in Flensburg“ bei zufällig ausgewählten Autofahrern in einer Stadt mit problematischer Verkehrsbeschilderung und hoher polizeilicher Kontrollichte:



13.2.7. Binomialtest

Der Fall: „Unter 1000 britischen Ärzten geben 600 an, emotional nicht stabile Patienten im Zweifel lieber zur Ärztin als zum Arzt zu überweisen, da Frauen unter den Ärzten mehr Einfühlungsvermögen zugetraut wird.“ (Ärzte Zeitung vom 17.9.97).

Wenn wir so eine Meldung lesen und mittels SPSS den Binomialtest auf die Daten anwenden wollen, aber nur die Häufigkeiten (und nicht den Datensatz) haben, können wir mittels des Befehls „weight“ dennoch Test durchführen. Wir bauen folgenden Datensatz auf:

ueberw	haeufig
1	600
2	400

Dann wählen wir über das Menue folgende Befehle: Statistik, nichtparametrische Tests, Binomial. Dort wählen wir als Testvariable „ueberw“. Den Test-Anteil lassen wir bei ,50. Wir klicken auf die Schaltfläche: Befehl und erhalten im Syntaxfenster den folgenden Text (ohne die erste Zeile, die wir hinzuschreiben und mit einem Punkt abschließen).

```
weight by haeufigk.
NPAR TEST
  /BINOMIAL (.50)= ueberw
  /MISSING ANALYSIS.
```

Wir führen den Befehl aus und erhalten im Ausgabefenster:

```
- - - - - Binomial Test
```

```
UEBERW      Geschlecht des bevorzugten Arztes

Cases
    400      = 1
    600      = 2
----
Test Prop. = ,5000
Obs. Prop. = ,4000
Z Approximation
```

1000 Total 2-Tailed P = ,0000

Hierbei bedeutet 2-Tailed P = ,000, daß die Wahrscheinlichkeit, aus einer gleichverteilten Grundgesamtheit das vorliegende Ergebnis zufällig zu erhalten, gleich Null ist. Die Nullhypothese wird daher abgelehnt. Die britischen Ärzte haben hiernach ein bestimmtes „Vor“-Urteil über ihre Kolleginnen.

13.2.8. Kolmogorov-Smirnov-Test auf Verteilungsform bei einer Stichprobe

Gegeben sei eine Stichprobe von 200 Zahlen aus Intelligenztests. Gefragt ist, ob diese Stichprobe „normalverteilt“ ist:

95;99;98;98;90;91;98;108;99;98;98;94;107;95;97;91;92;105;103;102;100;95;107;98;99;106;98;98;103;95;99;105;93;98;96;98;94;105;106;93;99;101;100;102;100;98;104;100;100;101;94;86;103;107;105;102;95;97;111;96;97;105;104;94;108;102;96;99;105;105;89;108;97;102;99;92;104;97;95;90;103;97;95;95;89;97;90;91;101;103;97;101;108;100;103;93;94;94;98;101;93;109;106;96;93;102;101;94;105;99;97;107;103;100;104;107;98;91;97;103;97;92;97;101;93;97;102;98;101;102;88;98;98;101;100;104;92;99;101;102;106;94;102;101;100;107;99;97;92;101;88;92;102;102;96;98;94;93;91;96;97;100;103;105;97;94;103;104;92;103;99;98;101;107;92;107;105;110;102;97;104;111;90;103;94;96;99;100;98;93;97;106;100;110;97;94;96;104;107.

Hierzu lesen wir unsere Daten im Datenfenster unter der Variablenbezeichnung: INTELI ein und geben folgende Befehle: „Statistik“, „nichtparametrische Tests“, „1-Stichproben-KS“. Wir wählen die Variable „INTELI“ als Testvariable und wählen bei den Optionen noch „Deskriptive Statistiken“ und „Quartile“. Wir wählen die Prüfung auf „Normalverteilung“. Dann klicken wir die Schaltfläche „Befehl“ an und erhalten im Syntaxfenster:

```
NPART TESTS
  /K-S(NORMAL)= INTELI
  /STATISTICS DESCRIPTIVES QUANTILES
  /MISSING ANALYSIS.
```

Diesen Befehl führen wir aus und erhalten im Ausgabefenster:

	N	Mean	Std Dev	Minimum	Maximum
INTELI	200	99,095	5,134	86	111
		(Median)			
		25th	50th	75th	
	N	Percentile	Percentile	Percentile	
INTELI	200	95,25	99,00	103,00	

- - - - - Kolmogorov - Smirnov Goodness of Fit Test
INTELI

Test distribution	-	Normal	Mean:	99,10
			Standard Deviation:	5,13

Cases: 200

Most extreme differences				
Absolute	Positive	Negative	K-S Z	2-Tailed P
,06445	,06445	-,05161	,9114	,3771

SPSS berechnet eine Irrtums-Wahrscheinlichkeit im zweiseitigen Test von $p = 0,3771$, daß die Daten nicht zufällig einer Normalverteilung entnommen worden sind. Bei dieser hohen Irrtumswahrscheinlichkeit behalten wir die Nullhypothese bei, daß die Daten aus einer Normalverteilung entnommen wurden.

4. Tabellen

4.1. Standardnormalverteilung

Die folgende Tabelle zeigt die Verteilungsfunktion der Standardnormalverteilung. Für ausgewählte z-Werte ist die Wahrscheinlichkeit $W(Z \leq z) = (1 - \alpha)$ angegeben, daß dieser oder ein kleinerer z-Wert auftritt.

Lesebeispiel: Die Tabelle enthält z-Werte, die auf zwei Stellen hinter dem Komma gerundet sind: z.B. -2,03, 1,07 oder 1,96. Damit die Tabelle nicht zu viel Platz einnimmt, ist jeder z-Wert in zwei Teile aufgespalten: Teil 1 mit der ersten Nachkommastelle (Spalte 1) und Teil 2 mit der zweiten Nachkommastelle (alle folgenden Spalten). Gesucht sei die Wahrscheinlichkeit, daß maximal ein z-Wert von -2,03 auftritt (Teil 1: -2,0; Teil 2: 0,03). In der Zeile mit z=-2,0 (erste Spalte) finden Sie in der mit 0,03 überschriebenen Spalte die gesuchte Wahrscheinlichkeit $(1 - \alpha) = 0,0212$.

z-Werte (Teil 1)	Teil 2 des z-Wertes (2. Nachkommastelle)									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224

0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

4.2. χ^2 -Verteilung

Die folgende Tabelle zeigt die inverse Verteilungsfunktion der χ^2 -Verteilung: $\chi^2(1-\alpha|df)$. Für ausgewählte Freiheitsgrade (df) und Wahrscheinlichkeiten (1- α) werden die entsprechenden χ^2 -Werte (χ^2 -Quantile) dargestellt. *Lesebeispiel:* Gesucht sei der χ^2 -Wert, unter dem bei df=17 Freiheitsgraden 95% aller möglichen Werte einer χ^2 -verteilten Zufallsvariablen X^2 liegen. In der Zeile für df=17 finden Sie in der Spalte (1- α)=0,95 den gesuchten Wert $\chi^2=27,59$.

df	(rote/dunkle) Fläche (1- α)								
	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84
4	4,88	5,39	5,99	6,74	7,78	9,49	11,14	13,28	14,86
5	6,06	6,63	7,29	8,12	9,24	11,07	12,83	15,09	16,75
6	7,23	7,84	8,56	9,45	10,64	12,59	14,45	16,81	18,55
7	8,38	9,04	9,80	10,75	12,02	14,07	16,01	18,48	20,28
8	9,52	10,22	11,03	12,03	13,36	15,51	17,53	20,09	21,95
9	10,66	11,39	12,24	13,29	14,68	16,92	19,02	21,67	23,59
10	11,78	12,55	13,44	14,53	15,99	18,31	20,48	23,21	25,19
11	12,90	13,70	14,63	15,77	17,28	19,68	21,92	24,73	26,76
12	14,01	14,85	15,81	16,99	18,55	21,03	23,34	26,22	28,30
13	15,12	15,98	16,98	18,20	19,81	22,36	24,74	27,69	29,82
14	16,22	17,12	18,15	19,41	21,06	23,68	26,12	29,14	31,32
15	17,32	18,25	19,31	20,60	22,31	25,00	27,49	30,58	32,80
16	18,42	19,37	20,47	21,79	23,54	26,30	28,85	32,00	34,27
17	19,51	20,49	21,61	22,98	24,77	27,59	30,19	33,41	35,72
18	20,60	21,60	22,76	24,16	25,99	28,87	31,53	34,81	37,16
19	21,69	22,72	23,90	25,33	27,20	30,14	32,85	36,19	38,58
20	22,77	23,83	25,04	26,50	28,41	31,41	34,17	37,57	40,00
21	23,86	24,93	26,17	27,66	29,62	32,67	35,48	38,93	41,40
22	24,94	26,04	27,30	28,82	30,81	33,92	36,78	40,29	42,80
23	26,02	27,14	28,43	29,98	32,01	35,17	38,08	41,64	44,18
24	27,10	28,24	29,55	31,13	33,20	36,42	39,36	42,98	45,56
25	28,17	29,34	30,68	32,28	34,38	37,65	40,65	44,31	46,93
30	33,53	34,80	36,25	37,99	40,26	43,77	46,98	50,89	53,67
40	44,16	45,62	47,27	49,24	51,81	55,76	59,34	63,69	66,77
50	54,72	56,33	58,16	60,35	63,17	67,50	71,42	76,15	79,49
60	65,23	66,98	68,97	71,34	74,40	79,08	83,30	88,38	91,95
70	75,69	77,58	79,71	82,26	85,53	90,53	95,02	100,43	104,21
80	86,12	88,13	90,41	93,11	96,58	101,88	106,63	112,33	116,32
90	96,52	98,65	101,05	103,90	107,57	113,15	118,14	124,12	128,30
100	106,91	109,14	111,67	114,66	118,50	124,34	129,56	135,81	140,17
150	158,58	161,29	164,35	167,96	172,58	179,58	185,80	193,21	198,36
200	209,99	213,10	216,61	220,74	226,02	233,99	241,06	249,45	255,26
500	516,09	520,95	526,40	532,80	540,93	553,13	563,85	576,49	585,21

4.3. t-Verteilung

Die folgende Tabelle zeigt die inverse Verteilungsfunktion der T-Verteilung: $T(1-\alpha|df)$. Für ausgewählte Freiheitsgrade (df) und Wahrscheinlichkeiten $(1-\alpha)$ werden die entsprechenden t-Werte (t-Quantile) dargestellt, für die gilt: $W(T \leq t|df) = (1-\alpha)$.

Lesebeispiel: Gesucht sei der t-Wert, unter dem bei $df=17$ Freiheitsgraden 95% aller möglichen Werte einer T-verteilten Zufallsvariablen T liegen. In der Zeile für $df=17$ finden Sie in der Spalte $(1-\alpha)=0,95$ den gesuchten Wert $t=1,740$.

df	Fläche (1- α)									
	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656
2	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
30	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
70	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648
80	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
90	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632
100	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626
150	0,386	0,526	0,676	0,844	1,040	1,287	1,655	1,976	2,351	2,609
200	0,386	0,525	0,676	0,843	1,039	1,286	1,653	1,972	2,345	2,601
500	0,386	0,525	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586
1000	0,385	0,525	0,675	0,842	1,037	1,282	1,646	1,962	2,330	2,581

4.4. F-Verteilung für $(1-\alpha)=0,99$

Die folgende Tabelle zeigt die inverse Verteilungsfunktion der F-Verteilung für $(1-\alpha)=0,99$. Für ausgewählte Zähler- und Nenner-Freiheitsgrade (df_1 , df_2) werden die entsprechenden f-Werte (f-Quantile) dargestellt, für die gilt: $W(F \leq f | df_1, df_2) = 0,99$. *Lesebeispiel:* Gesucht sei der f-Wert, unter dem bei $df_1=2$ Zähler-Freiheitsgraden und $df_2=4$ Nenner-Freiheitsgraden 99% aller möglichen Werte einer F-verteilten Zufallsvariablen F liegen. In der Zeile für $df_2=4$ finden Sie in der Spalte für $df_1=2$ den gesuchten Wert $F=18,00$.

df ₂ (Nenner)	df ₁ (Zähler)																		
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	20	22	24	26
1	4052,18	4999,34	5403,53	5624,26	5763,96	5858,95	5928,33	5980,95	6022,40	6055,93	6083,40	6106,68	6143,00	6170,01	6191,43	6208,66	6223,10	6234,27	6244,52
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,43	99,44	99,44	99,45	99,46	99,46	99,46
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,92	26,83	26,75	26,69	26,64	26,60	26,56
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,25	14,15	14,08	14,02	13,97	13,93	13,89
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,77	9,68	9,61	9,55	9,51	9,47	9,43
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,60	7,52	7,45	7,40	7,35	7,31	7,28
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,36	6,28	6,21	6,16	6,11	6,07	6,04
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,56	5,48	5,41	5,36	5,32	5,28	5,25
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,01	4,92	4,86	4,81	4,77	4,73	4,70
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,60	4,52	4,46	4,41	4,36	4,33	4,30
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,29	4,21	4,15	4,10	4,06	4,02	3,99
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,05	3,97	3,91	3,86	3,82	3,78	3,75
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,86	3,78	3,72	3,66	3,62	3,59	3,56
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,70	3,62	3,56	3,51	3,46	3,43	3,40
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,56	3,49	3,42	3,37	3,33	3,29	3,26
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,45	3,37	3,31	3,26	3,22	3,18	3,15
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,35	3,27	3,21	3,16	3,12	3,08	3,05
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,27	3,19	3,13	3,08	3,03	3,00	2,97
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,19	3,12	3,05	3,00	2,96	2,92	2,89
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,13	3,05	2,99	2,94	2,90	2,86	2,83
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,07	2,99	2,93	2,88	2,84	2,80	2,77
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,02	2,94	2,88	2,83	2,78	2,75	2,72
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	2,97	2,89	2,83	2,78	2,74	2,70	2,67
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,93	2,85	2,79	2,74	2,70	2,66	2,63
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,89	2,81	2,75	2,70	2,66	2,62	2,59
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,86	2,78	2,72	2,66	2,62	2,58	2,55
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,82	2,75	2,68	2,63	2,59	2,55	2,52
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,79	2,72	2,65	2,60	2,56	2,52	2,49
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,77	2,69	2,63	2,57	2,53	2,49	2,46
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,74	2,66	2,60	2,55	2,51	2,47	2,44

31	7,53	5,36	4,48	3,99	3,67	3,45	3,28	3,15	3,04	2,96	2,88	2,82	2,72	2,64	2,58	2,52	2,48	2,45	2,41
32	7,50	5,34	4,46	3,97	3,65	3,43	3,26	3,13	3,02	2,93	2,86	2,80	2,70	2,62	2,55	2,50	2,46	2,42	2,39
33	7,47	5,31	4,44	3,95	3,63	3,41	3,24	3,11	3,00	2,91	2,84	2,78	2,68	2,60	2,53	2,48	2,44	2,40	2,37
34	7,44	5,29	4,42	3,93	3,61	3,39	3,22	3,09	2,98	2,89	2,82	2,76	2,66	2,58	2,51	2,46	2,42	2,38	2,35
35	7,42	5,27	4,40	3,91	3,59	3,37	3,20	3,07	2,96	2,88	2,80	2,74	2,64	2,56	2,50	2,44	2,40	2,36	2,33
36	7,40	5,25	4,38	3,89	3,57	3,35	3,18	3,05	2,95	2,86	2,79	2,72	2,62	2,54	2,48	2,43	2,38	2,35	2,32
37	7,37	5,23	4,36	3,87	3,56	3,33	3,17	3,04	2,93	2,84	2,77	2,71	2,61	2,53	2,46	2,41	2,37	2,33	2,30
38	7,35	5,21	4,34	3,86	3,54	3,32	3,15	3,02	2,92	2,83	2,75	2,69	2,59	2,51	2,45	2,40	2,35	2,32	2,28
39	7,33	5,19	4,33	3,84	3,53	3,30	3,14	3,01	2,90	2,81	2,74	2,68	2,58	2,50	2,43	2,38	2,34	2,30	2,27
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,56	2,48	2,42	2,37	2,33	2,29	2,26
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,46	2,38	2,32	2,27	2,22	2,18	2,15
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,39	2,31	2,25	2,20	2,15	2,12	2,08
70	7,01	4,92	4,07	3,60	3,29	3,07	2,91	2,78	2,67	2,59	2,51	2,45	2,35	2,27	2,20	2,15	2,11	2,07	2,03
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,48	2,42	2,31	2,23	2,17	2,12	2,07	2,03	2,00
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52	2,45	2,39	2,29	2,21	2,14	2,09	2,04	2,00	1,97
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,27	2,19	2,12	2,07	2,02	1,98	1,95
125	6,84	4,78	3,94	3,47	3,17	2,95	2,79	2,66	2,55	2,47	2,39	2,33	2,23	2,15	2,08	2,03	1,98	1,94	1,91
150	6,81	4,75	3,91	3,45	3,14	2,92	2,76	2,63	2,53	2,44	2,37	2,31	2,20	2,12	2,06	2,00	1,96	1,92	1,88
175	6,78	4,73	3,90	3,43	3,12	2,91	2,74	2,61	2,51	2,42	2,35	2,29	2,19	2,10	2,04	1,98	1,94	1,90	1,87
200	6,76	4,71	3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,17	2,09	2,03	1,97	1,93	1,89	1,85
250	6,74	4,69	3,86	3,40	3,09	2,87	2,71	2,58	2,48	2,39	2,32	2,26	2,15	2,07	2,01	1,95	1,91	1,87	1,83
300	6,72	4,68	3,85	3,38	3,08	2,86	2,70	2,57	2,47	2,38	2,31	2,24	2,14	2,06	1,99	1,94	1,89	1,85	1,82
400	6,70	4,66	3,83	3,37	3,06	2,85	2,68	2,56	2,45	2,37	2,29	2,23	2,13	2,05	1,98	1,92	1,88	1,84	1,80
500	6,69	4,65	3,82	3,36	3,05	2,84	2,68	2,55	2,44	2,36	2,28	2,22	2,12	2,04	1,97	1,92	1,87	1,83	1,79
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,10	2,02	1,95	1,90	1,85	1,81	1,77
100000 0	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,08	2,00	1,93	1,88	1,83	1,79	1,76

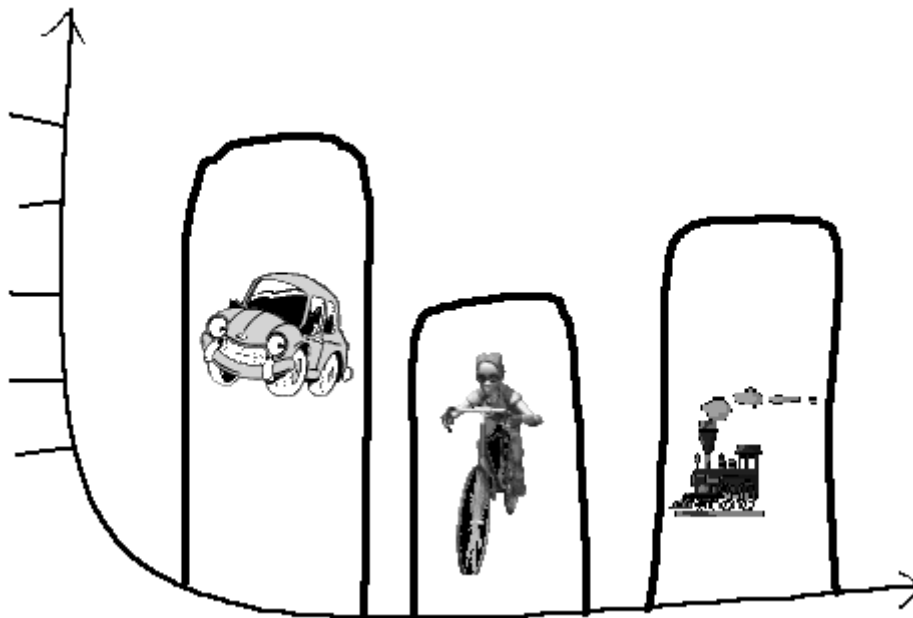
Fortsetzung F-Verteilung für $(1-\alpha) = 0,99$

df ₂ (Nenne r)	df ₁ (Zähler)																		
	27	28	29	30	31	32	33	34	35	36	37	38	40	42	44	46	100	500	10000 00
1	6249,17	6252,90	6257,09	6260,35	651,52	6266,87	6270,13	6272,92	6275,25	6278,05	6280,37	6282,70	6286,43	6290,15	6293,88	6296,67	6333,92	6359,54	6365,59
2	99,46	99,46	99,46	99,47	31,72	99,47	99,47	99,47	99,47	99,47	99,47	99,47	99,48	99,48	99,48	99,48	99,49	99,50	99,50
3	26,55	26,53	26,52	26,50	12,10	26,48	26,47	26,46	26,45	26,44	26,43	26,43	26,41	26,40	26,39	26,37	26,24	26,15	26,13
4	13,88	13,86	13,85	13,84	7,51	13,81	13,80	13,79	13,79	13,78	13,77	13,76	13,75	13,73	13,72	13,71	13,58	13,49	13,46
5	9,42	9,40	9,39	9,38	5,63	9,36	9,35	9,34	9,33	9,32	9,31	9,31	9,29	9,28	9,27	9,26	9,13	9,04	9,02
6	7,27	7,25	7,24	7,23	4,64	7,21	7,20	7,19	7,18	7,17	7,16	7,16	7,14	7,13	7,12	7,11	6,99	6,90	6,88
7	6,03	6,02	6,00	5,99	4,03	5,97	5,96	5,95	5,94	5,94	5,93	5,92	5,91	5,90	5,89	5,88	5,75	5,67	5,65
8	5,23	5,22	5,21	5,20	3,62	5,18	5,17	5,16	5,15	5,14	5,14	5,13	5,12	5,10	5,09	5,08	4,96	4,88	4,86
9	4,68	4,67	4,66	4,65	3,33	4,63	4,62	4,61	4,60	4,59	4,59	4,58	4,57	4,55	4,54	4,53	4,41	4,33	4,31
10	4,28	4,27	4,26	4,25	3,11	4,23	4,22	4,21	4,20	4,19	4,19	4,18	4,17	4,15	4,14	4,13	4,01	3,93	3,91
11	3,98	3,96	3,95	3,94	2,93	3,92	3,91	3,90	3,89	3,89	3,88	3,87	3,86	3,85	3,84	3,83	3,71	3,62	3,60
12	3,74	3,72	3,71	3,70	2,80	3,68	3,67	3,66	3,65	3,65	3,64	3,63	3,62	3,61	3,60	3,59	3,47	3,38	3,36
13	3,54	3,53	3,52	3,51	2,68	3,49	3,48	3,47	3,46	3,45	3,45	3,44	3,43	3,41	3,40	3,39	3,27	3,19	3,17
14	3,38	3,37	3,36	3,35	2,59	3,33	3,32	3,31	3,30	3,29	3,29	3,28	3,27	3,25	3,24	3,23	3,11	3,03	3,00
15	3,25	3,24	3,23	3,21	2,51	3,19	3,18	3,18	3,17	3,16	3,15	3,15	3,13	3,12	3,11	3,10	2,98	2,89	2,87
16	3,14	3,12	3,11	3,10	2,44	3,08	3,07	3,06	3,05	3,05	3,04	3,03	3,02	3,01	3,00	2,99	2,86	2,78	2,75
17	3,04	3,03	3,01	3,00	2,38	2,98	2,97	2,96	2,96	2,95	2,94	2,93	2,92	2,91	2,90	2,89	2,76	2,68	2,65
18	2,95	2,94	2,93	2,92	2,33	2,90	2,89	2,88	2,87	2,86	2,86	2,85	2,84	2,82	2,81	2,80	2,68	2,59	2,57
19	2,88	2,87	2,86	2,84	2,29	2,82	2,81	2,81	2,80	2,79	2,78	2,77	2,76	2,75	2,74	2,73	2,60	2,51	2,49
20	2,81	2,80	2,79	2,78	2,24	2,76	2,75	2,74	2,73	2,72	2,72	2,71	2,69	2,68	2,67	2,66	2,54	2,44	2,42
21	2,76	2,74	2,73	2,72	2,21	2,70	2,69	2,68	2,67	2,66	2,66	2,65	2,64	2,62	2,61	2,60	2,48	2,38	2,36
22	2,70	2,69	2,68	2,67	2,17	2,65	2,64	2,63	2,62	2,61	2,60	2,60	2,58	2,57	2,56	2,55	2,42	2,33	2,31
23	2,66	2,64	2,63	2,62	2,14	2,60	2,59	2,58	2,57	2,56	2,56	2,55	2,54	2,52	2,51	2,50	2,45	2,28	2,26
24	2,61	2,60	2,59	2,58	2,12	2,56	2,55	2,54	2,53	2,52	2,51	2,51	2,49	2,48	2,47	2,46	2,40	2,24	2,21
25	2,58	2,56	2,55	2,54	2,09	2,52	2,51	2,50	2,49	2,48	2,47	2,47	2,45	2,44	2,43	2,42	2,36	2,19	2,17
26	2,54	2,53	2,51	2,50	2,07	2,48	2,47	2,46	2,45	2,45	2,44	2,43	2,42	2,40	2,39	2,38	2,33	2,16	2,13
27	2,51	2,49	2,48	2,47	2,05	2,45	2,44	2,43	2,42	2,41	2,41	2,40	2,38	2,37	2,36	2,35	2,29	2,12	2,10
28	2,48	2,46	2,45	2,44	2,03	2,42	2,41	2,40	2,39	2,38	2,37	2,37	2,35	2,34	2,33	2,32	2,26	2,09	2,06
29	2,45	2,44	2,42	2,41	2,01	2,39	2,38	2,37	2,36	2,35	2,35	2,34	2,33	2,31	2,30	2,29	2,23	2,06	2,03
30	2,42	2,41	2,40	2,39	1,99	2,36	2,35	2,35	2,34	2,33	2,32	2,31	2,30	2,29	2,27	2,26	2,21	2,03	2,01
31	2,40	2,39	2,37	2,36	1,98	2,34	2,33	2,32	2,31	2,30	2,30	2,29	2,27	2,26	2,25	2,24	2,18	2,01	1,98
32	2,38	2,36	2,35	2,34	1,96	2,32	2,31	2,30	2,29	2,28	2,27	2,27	2,25	2,24	2,23	2,22	2,16	1,98	1,96
33	2,36	2,34	2,33	2,32	1,95	2,30	2,29	2,28	2,27	2,26	2,25	2,25	2,23	2,22	2,21	2,20	2,14	1,96	1,93
34	2,34	2,32	2,31	2,30	1,94	2,28	2,27	2,26	2,25	2,24	2,23	2,23	2,21	2,20	2,19	2,18	2,12	1,94	1,91
35	2,32	2,30	2,29	2,28	1,92	2,26	2,25	2,24	2,23	2,22	2,21	2,21	2,19	2,18	2,17	2,16	2,10	1,92	1,89
36	2,30	2,29	2,28	2,26	1,91	2,24	2,23	2,22	2,21	2,21	2,20	2,19	2,18	2,16	2,15	2,14	2,08	1,90	1,87
37	2,28	2,27	2,26	2,25	1,90	2,23	2,22	2,21	2,20	2,19	2,18	2,17	2,16	2,15	2,13	2,12	2,06	1,88	1,85

38	2,27	2,26	2,24	2,23	1,89	2,21	2,20	2,19	2,18	2,17	2,16	2,16	2,14	2,13	2,12	2,11	2,05	1,86	1,84
39	2,26	2,24	2,23	2,22	1,88	2,20	2,19	2,18	2,17	2,16	2,15	2,14	2,13	2,12	2,10	2,09	2,03	1,85	1,82
40	2,24	2,23	2,22	2,20	1,87	2,18	2,17	2,16	2,15	2,14	2,14	2,13	2,11	2,10	2,09	2,08	2,02	1,83	1,80
50	2,14	2,12	2,11	2,10	1,80	2,08	2,07	2,06	2,05	2,04	2,03	2,02	2,01	1,99	1,98	1,97	1,91	1,71	1,68
60	2,07	2,05	2,04	2,03	1,76	2,01	2,00	1,99	1,98	1,97	1,96	1,95	1,94	1,92	1,91	1,90	1,84	1,63	1,60
70	2,02	2,01	1,99	1,98	1,72	1,96	1,95	1,94	1,93	1,92	1,91	1,90	1,89	1,87	1,86	1,85	1,78	1,57	1,54
80	1,98	1,97	1,96	1,94	1,70	1,92	1,91	1,90	1,89	1,88	1,87	1,86	1,85	1,83	1,82	1,81	1,75	1,53	1,49
90	1,96	1,94	1,93	1,92	1,68	1,89	1,88	1,87	1,86	1,85	1,84	1,84	1,82	1,81	1,79	1,78	1,72	1,49	1,46
100	1,93	1,92	1,91	1,89	1,66	1,87	1,86	1,85	1,84	1,83	1,82	1,81	1,80	1,78	1,77	1,76	1,69	1,47	1,43
125	1,89	1,88	1,87	1,85	1,64	1,83	1,82	1,81	1,80	1,79	1,78	1,77	1,76	1,74	1,73	1,72	1,65	1,41	1,37
150	1,78	1,85	1,84	1,83	1,62	1,80	1,79	1,78	1,77	1,76	1,75	1,74	1,73	1,71	1,70	1,69	1,62	1,38	1,33
175	1,85	1,74	1,82	1,81	1,60	1,78	1,77	1,76	1,75	1,74	1,73	1,73	1,71	1,69	1,68	1,67	1,60	1,35	1,30
200	1,84	1,82	1,71	1,79	1,59	1,77	1,76	1,75	1,74	1,73	1,72	1,71	1,69	1,68	1,67	1,65	1,58	1,33	1,28
250	1,82	1,80	1,79	1,77	1,58	1,75	1,74	1,73	1,72	1,71	1,70	1,69	1,67	1,66	1,64	1,63	1,56	1,30	1,24
300	1,80	1,79	1,77	1,76	1,57	1,74	1,73	1,72	1,70	1,70	1,69	1,68	1,66	1,65	1,63	1,62	1,55	1,28	1,22
400	1,79	1,77	1,76	1,75	1,56	1,72	1,71	1,70	1,69	1,68	1,67	1,66	1,64	1,63	1,61	1,60	1,53	1,25	1,19
500	1,78	1,76	1,75	1,74	1,55	1,71	1,70	1,69	1,68	1,67	1,66	1,65	1,63	1,62	1,60	1,59	1,52	1,23	1,16
1000	1,76	1,74	1,73	1,72	1,54	1,69	1,68	1,67	1,66	1,65	1,64	1,63	1,61	1,60	1,58	1,57	1,50	1,19	1,11
100000 0	1,74	1,72	1,71	1,70	1,52	1,67	1,66	1,65	1,64	1,63	1,62	1,61	1,59	1,58	1,56	1,55	1,47	1,15	

Statistische Formelsammlung

Begleitende Materialien zur
Statistik - Vorlesung des Grundstudiums
im Fachbereich IK



Erstellt im Rahmen des studierenden Projektes PROST
Studienjahr 2001/2002
unter Anleitung von Frau Prof. Dr. rer. nat. Kira Klenke
von J.Gelwert, C.Seeger, E.Zotou

Statistik	4
Deskriptive Statistik	4
Analytische Statistik	4
Grundbegriffe der Statistik	5
Urliste	5
Grundgesamtheit	5
Stichprobe	5
Vollerhebung	5
Merkmalsträger	5
Merkmal	5
Merkmalsausprägungen	6
Quantitative und qualitative Merkmale	6
quantitativ	6
stetige / metrische Variablen	6
diskrete Variablen	6
qualitativ	6
nominale Variablen	6
ordinale Variablen	6
Deskriptive Statistik	7
Mittelwerte	7
Arithmetisches Mittel	7
Geometrisches Mittel	7
Schätzung des arithmetischen Mittels für klassierte Daten	8
Median	9
Modus	9
Streuungsmaße	10
Spannweite	10
Varianz und Standardabweichung	10
in der Stichprobe	10
in der Grundgesamtheit	10
Quartile	11
Quartilsabstand	11
Variationskoeffizient	11
Welches Streuungsmaß bei welchem Mittelwert?	11
Häufigkeiten	12
Häufigkeitstabelle	12
Kreuztabelle	13
Klassenbildung	14
Grafiken der deskriptiven Statistik	15
Kreisdiagramm	15
Stabdiagramm	15
Balkendiagramm	15
Stem & Leaf	15
Box-Plot	16
Histogramm	16
Verlaufskurve	16
Analytische Statistik	17
Wahrscheinlichkeitsrechnung	17
Ereignisraum	17
Definitionen der Wahrscheinlichkeit	17
Additionssatz	17

allgemeiner Additionssatz	17
spezieller Additionssatz	17
stochastische (Un-)Abhängigkeit von Ereignissen	18
bedingte Wahrscheinlichkeit	18
Multiplikationssatz	18
allgemeiner Multiplikationssatz	18
spezieller Multiplikationssatz	18
Kombinatorik	19
Fakultät	19
Binomialkoeffizient	19
Wahrscheinlichkeitsverteilungen	19
Verteilungsfunktion	19
Binomialverteilung	19
Gauß- oder Normalverteilung	20
standardisierte Normalverteilung	20
beliebige Zufallsvariable	20
Transformation der Standardnormalverteilung	20
Parameter in der Grundgesamtheit	20
Konfidenzintervalle	21
für einen unbekannten Mittelwert	21
für eine unbekannte Proportion	21
Fallzahlbestimmung	22
Bestimmung des erforderlichen (Mindest-) Stichprobenumfangs	
bei Schätzung eines (prozentualen) Anteils π	22
Bestimmung der erforderlichen (Mindest-)Fallzahl n bei der	
Schätzung eines (unbekannten) Mittelwertes μ	22
Stichwortliste	23

Statistik

Statistik ist eine Wissenschaft, deren Zweck es ist, Informationen in Form von Zahlen oder anderen Daten zusammenzutragen, das gesammelte Material aufzubereiten und die Informationen schließlich so zu verdichten, dass sie aussagekräftig, verständlich und überschaubar werden. Sie ist eine gute Möglichkeit, Fakten exakt und überzeugend darzustellen.

Sie findet Anwendung sowohl in naturwissenschaftlichen als auch wirtschaftlichen, soziologischen und politischen Bereichen.

Sie unterteilt sich in zwei Disziplinen:

Deskriptive Statistik

Deskriptive Statistik bedeutet *beschreibende* Statistik und meint die reine Beschreibung des erhobenen Datenbestandes (in der Regel die Stichprobe).

Entsprechend der Fragestellung werden Daten erhoben, ausgewertet und dargestellt.

Deskriptive Statistik ist also eine Zusammenfassung von unüberschaubar vielen Einzeldaten bzw. ein überschaubares Komprimieren von vielen Einzelinformationen.

Die zusammenfassende Darstellung erfolgt durch die Berechnung der Maßzahlen und die Erstellung von Häufigkeitstabellen und Grafiken.

Analytische Statistik

auch: *Schließende* Statistik, Interferenzstatistik, induktive Statistik.

Die analytische Statistik (insbesondere mit Hilfe der Wahrscheinlichkeitsrechnung) befasst sich damit, wie aufgrund von Daten der Stichproben Rückschlüsse bzw. allgemeingültige Aussagen über die zugrunde liegende Grundgesamtheit hergeleitet werden können.

Grundbegriffe der Statistik

Urliste

Die Urliste ist die Liste der Einzeldaten, so wie erhoben, noch nicht sortiert oder gruppiert.

$x_1, x_2, x_3, x_4, \dots, x_n$

sortierte Urliste

Hier werden die Ausprägungen in eine aufsteigende Reihenfolge gebracht.

$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, \dots, x_{(n)}$

d.h.

$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq \dots, x_{(n)}$

Grundgesamtheit

Die Grundgesamtheit ist die Menge aller interessierenden Untersuchungsobjekte.

Stichprobe

Unter Stichprobe versteht man eine Teilmenge der Grundgesamtheit, die ausgewählt wird. Sie muss repräsentativ sein.

Eine Stichprobe ist dann repräsentativ, wenn jede Untersuchungseinheit aus der Grundgesamtheit dieselbe Chance hat, ausgewählt zu werden.

Vollerhebung

Bei einer Vollerhebung wird die ganze Grundgesamtheit erhoben.

Merkmalsträger / Untersuchungsobjekte / statistische Einheit / Element

Objekte, an denen die interessierenden Merkmale auftreten.

Merkmal / Variable

Interessierende Eigenschaften der Merkmalsträger.

Hierbei ist eine exakte Definition wichtig.

Merkmalsausprägungen

Es sind die möglichen Meßwerte eines Merkmals gemeint.

Quantitative und qualitative Merkmale (Merkmalsausprägungen)

quantitativ:

Ausprägungen sind echte Meßwerte (von Natur aus Zahlen), d.h. addier-, subtrahier- und multiplizierbar. Man unterscheidet folgende Skalierungen:

❖ **stetige / metrische Variablen**

- können fast jeden Wert und Zwischenwert annehmen, d.h. sie sind beliebig fein meßbar

❖ **diskrete Variablen**

- es können nur endlich viele, ganz konkrete Werte angenommen werden (i.d.R. ganzzahlig, keine Brüche)

qualitativ:

Ein Merkmal bezeichnet eine Eigenschaft, „Qualität“, der untersuchten Objekte. Die Ausprägungen eines qualitativen Merkmales sind **keine** echten Zahlen. Man kann mit ihnen **nicht** rechnen! Hier unterscheidet man folgende Skalierungen:

❖ **nominale Variablen**

- sind nur namentlich unterscheidbar (keine Zahlen, keine Ordnung, nur Namen)

❖ **ordinale Variablen**

- sind sortierbar, es gibt eine „innere Ordnung“, es kann damit nicht gerechnet werden

Diese 4 Skalierungstypen sind weisungsgebend für die Art der Datenauswertung!!

Deskriptive Statistik

Mittelwerte

Arithmetisches Mittel (engl. arithmetic mean)

auch: Standard - Mittelwert

Man addiert die Werte aller Elemente und dividiert den Gesamtwert durch die Anzahl der Elemente. Eindeutig zu bestimmen ist das arithmetische Mittel, wenn die Originalwerte in nicht klassifizierter Form vorliegen.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{x}: \text{sprich x-quer}$$

n = Anzahl der Beobachtungen / Messungen

\bar{x} = Ausprägung (kann sehr leicht verzerrt / verfälscht werden durch Extremwerte)
geeignet für stetige Merkmale. Wird in der Praxis auch benutzt bei ordinalen und diskreten Merkmalen.

Gewogenes arithmetisches Mittel

$$\bar{x} = \frac{1}{\sum_{i=1}^k f_i} \cdot \sum_{i=1}^k x_i \cdot f_i$$

k = Anzahl verschiedener Werte des Merkmals X
 f_i = die zugehörigen Häufigkeiten

Geometrisches Mittel (engl. geometric mean)

Ist geeignet zur Berechnung von Wachstumsraten.

durchschnittliches Wachstum w:
Schreibweise in %:

$$w = \sqrt[n]{(100\% + w_1) \cdot (100\% + w_2) \cdot (100\% + w_3) \cdot \dots \cdot (100\% + w_n)} - 100\%$$

wobei $w_1, w_2, w_3, \dots, w_n$ Wachstumsraten in n gleich langen Zeiträumen in %-Angaben.

alternativ: Komma-Schreibweise:

$$w = \sqrt[n]{(1 + w_1) \cdot (1 + w_2) \cdot (1 + w_3) \cdot \dots \cdot (1 + w_n)} - 1$$

Schätzung des arithmetischen Mittels für klassierte Daten (Klassenbildung S.14)

$\bar{x} \approx$ ist eine Näherung!!

Klassen eines stetigen Merkmals	Absolute Häufigkeiten f_i	Klassenmitte a_i $\frac{Min_{Klasse} + Max_{Klasse}}{2}$
... bis unter
... bis unter
... bis unter

$$\bar{x} \approx \frac{\sum_{i=1}^k a_i \cdot f_i}{\sum_{i=1}^k f_i} = \frac{1}{n} \sum_{i=1}^k a_i \cdot f_i$$

k Anzahl verschiedener Klassen
 f_i die zugehörigen Häufigkeiten

$\sum_{i=1}^k f_i = n$ Anzahl von Beobachtungen

Median (engl. median)

auch: Zentralwert

Man sortiert die Urliste vom kleinsten zum größten Wert und der Median ist der Wert, der „in der Mitte“ liegt. Er ist unempfindlich gegen Extremwerte.
Er ist insbesondere für ordinalskalierte Merkmale geeignet und wenn diskrete und stetige Merkmale schief verteilt sind.

50 % der Werte größer oder gleich

50 % der Werte kleiner oder gleich

bei gerader Anzahl der Werte

$$Median = \frac{1}{2} \left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right]$$

Der Median ist bei einer geraden Anzahl von Werten der Mittelwert, der beiden Werte in der Mitte der sortierten Urliste.

bei ungerader Anzahl der Werte:

$$Median = x_{\left(\frac{n+1}{2}\right)}$$

Bei ungerader Anzahl der Werte ist der Median der Wert, der in der sortierten Urliste in der Mitte liegt.

Modus

auch: Modalwert

→ besonders für nominalskalierte Variablentypen geeignet

Der Modus ist derjenige Merkmalswert, der die größte Häufigkeit aufweist, vorausgesetzt es existiert genau 1 Häufigster.

Streuungsmaße (⇒ was befindet sich rechts und links vom Mittelwert?)

Zu jedem Mittelwert sollte immer mind. ein Streuungsmaß angegeben werden.

Spannweite R (engl. range)

Die Spannweite ist die Differenz aus dem maximalen und dem minimalen Wert der sortierten Urliste, der aber sehr leicht durch Extremwerte verzerrt werden kann.

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

$$R = x_{(n)} - x_{(1)}$$

Varianz (engl. variance) und Standardabweichung (engl. standard deviation)

Man unterscheidet zwischen Stichprobe und Grundgesamtheit.

Varianz und Standardabweichung in der Stichprobe werden mit s^2 bzw. s bezeichnet.

Varianz s^2 in der Stichprobe

Die Varianz ist die Summe der Quadrate der Abweichungen aller Werte vom arithmetischen Mittel, geteilt durch die Anzahl der Werte.

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Standardabweichung s in der Stichprobe (mittlere quadratische Abweichung)

Standardabweichung ist die Wurzel aus der Varianz.

$s = \sqrt{s^2}$ Für die Berechnung mit Taschenrechner ist die Taste $\boxed{\sigma n}$ zu benutzen.

Da in der Regel nicht die komplette Grundgesamtheit erfasst wird, werden die Varianz und Standardabweichung der Grundgesamtheit in der analytischen Statistik **geschätzt**. Diese wird, im Gegensatz zu deskriptiver Statistik, mit **$n-1$** statt mit **n** berechnet. Deswegen existieren zwei Tasten auf den Taschenrechner, eine für analytische und eine für deskriptive Statistik.

Schätzung der Varianz in der Grundgesamtheit (analytische Statistik)

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Schätzung der Standardabweichung in der Grundgesamtheit (analytische Statistik)

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ Für die Berechnung mit Taschenrechner ist die Taste $\boxed{\sigma n-1}$ zu benutzen.

Quartile

Sie sind insbesondere für ordinalskalierte Merkmale geeignet und wenn diskrete und stetige Merkmale schief verteilt sind.

Die geordnete Reihe von Merkmalswerten wird in vier gleiche Teile geteilt (also $4 \times 25\%$):

Unteres Quartil (1.Quartil):

25% der Werte sind kleiner oder gleich
75% der Werte sind größer oder gleich

Zweites Quartil → Median

50% der Werte sind kleiner oder gleich
50% der Werte sind größer oder gleich

Oberstes Quartil (3.Quartil):

75% der Werte sind kleiner oder gleich
25% der Werte sind größer oder gleich

Das **vierte Quartil** ist das Maximum:

100% der Werte sind kleiner oder gleich

Quartilabstand QA (engl. Interquartilerange IQR)

Quartilabstand ist die Differenz zwischen dem oberen und dem unteren Quartil (Spannweite der mittleren 50%)

$$QA = Q_3 - Q_1$$

Variationskoeffizient (engl. coefficient of variation)

Variationskoeffizient dient dazu, mehrere Standardabweichungen miteinander zu vergleichen.

$$CV = \frac{s}{\bar{x}} \quad \text{bzw.} \quad CV = \frac{s}{\bar{x}} \cdot 100\% \quad (\text{für Angaben in Prozent})$$

CV= Verhältnis der Standardabweichung s zum Mittelwert \bar{x} in %.
„Wie viel % des arithmetischen Mittels ist s ?“

Welches Streuungsmaß wird bei welchem Mittelwert angegeben?

Mittelwert	Streuungsmaß
\bar{x}	s^2 und s Spannweite beim Vergleichen mehrerer Streuungsmaße CV
Median	IQR / QA Spannweite

Häufigkeiten

Häufigkeitsverteilung (engl. frequency)

Bei Häufigkeitsverteilungen wird ausgezählt, wie oft einzelne Merkmalsausprägungen in der Stichprobe aufgetreten sind.

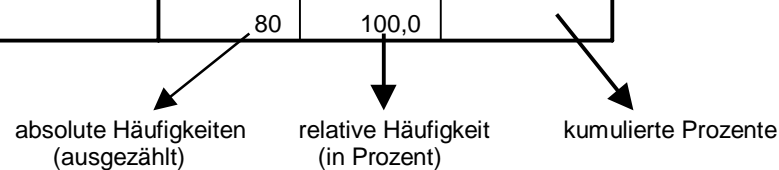
Besonders geeignet für nominale, ordinale und diskrete Ausprägungen.

Für stetige Merkmale müssen Klassen gebildet werden (siehe unten).

Die Ergebnisse der Auszählung werden in tabellarischer Form dargestellt.

Beispiel: eine Häufigkeitstabelle in SPSS

Fachrichtung		Häufigkeit	Prozent	Kumulierte Prozente
Gültig	Allgemeine Dokumentation	11	13,8	13,8
	Wirtschaftsinformatik	10	12,5	26,3
	Bibliothekswesen	7	8,8	35,0
	Psychologie	9	11,3	46,3
	Mathematik	3	3,8	50,0
	E-Technik	6	7,5	57,5
	Jura	8	10,0	67,5
	Sozialwissenschaften	3	3,8	71,3
	Lehramt für berufsbild.Schulen	6	7,5	78,8
	Geschichte & Politikwissens.	4	5,0	83,8
	Germanistik & Politikwissens.	6	7,5	91,3
	Architektur	7	8,8	100,0
	Gesamt	80	100,0	



Relative Häufigkeiten: $\frac{\text{Anzahl_mit_der_spez.Ausprägung}}{\text{Gesamtzahl}(n)} \cdot 100$

Kumulierte Prozente: Aufaddierung der Prozente zeilenweise.

Kreuztabelle

Mehrere Merkmale werden gegeneinander ausgezählt. (Normalfall: 2 Merkmale, ab 4 Merkmalen ist die Tabelle schwer überschaubar).

Softwarepakete geben die Zelleninhalte der Kreuztabelle i.d.R. folgendermaßen aus:

Häufigkeit absolut (frequency)

Reihenprozent (Rowpercent)

Spaltenprozent (Colpercent)

Gesamtprozent (Percent)

Beispiel erstellt mit SPSS

Geschlecht d.Stud. * welche Hochschule? Kreuztabelle

			welche Hochschule?		Gesamt
			Universität	Fachhochschule	
Geschlecht d.Stud.	weiblich	Anzahl	4	8	12
		% von Geschlecht d.Stud.	33,3%	66,7%	100,0%
		% von welche Hochschule?	36,4%	88,9%	60,0%
		% der Gesamtzahl	20,0%	40,0%	60,0%
	männlich	Anzahl	7	1	8
		% von Geschlecht d.Stud.	87,5%	12,5%	100,0%
		% von welche Hochschule?	63,6%	11,1%	40,0%
		% der Gesamtzahl	35,0%	5,0%	40,0%
Gesamt	Anzahl	11	9	20	
	% von Geschlecht d.Stud.	55,0%	45,0%	100,0%	
	% von welche Hochschule?	100,0%	100,0%	100,0%	
	% der Gesamtzahl	55,0%	45,0%	100,0%	

Inhaltliche Interpretation:

4 Personen sind weiblich und studieren an der Universität.

$$4/12 \cdot 100 = 33,3\%$$

33,3% der weiblichen Befragten studieren an der Universität.

$$4/11 \cdot 100 = 36,4$$

36,4% der Studierenden an der Universität sind weiblich.

$$4/20 \cdot 100 = 20\%$$

20% der Befragten sind weiblich und studieren an der Universität.

Klassenbildung bei stetigen Merkmalen

Da stetige Merkmale ganz unterschiedliche Merkmalswerte besitzen können, würde die direkte Auswertung der Daten durch Häufigkeitstabellen wenig Sinn machen. Deswegen werden die Daten nach unterschiedlichen Aspekten gruppiert, d.h. es werden Klassen gebildet, die in Häufigkeits- oder Kreuztabellen ausgewertet werden können.

Festgelegt werden muss:

- Wie viele Klassen lege ich fest?
- Wo lege ich Klassengrenzen fest?

Wie viele Klassen?

Anzahl der Klassen wird bestimmt durch die Größe der Stichprobe.

Faustregel: Anzahl der Klassen = \sqrt{n} (bei Dezimalstellen sinnvoll runden)

Klassengrenzen

- Grenzen dürfen sich nicht überschneiden (klare Definition)
- Klassengrenzen sollen aneinander grenzen
- 1.Klasse fängt bei dem minimalen Wert an
- letzte Klasse hört bei dem maximalen Wert auf
- zur besseren Lesbarkeit sollten die Klassengrenzen „glatte Zahlen“ sein

Gleich breite Klassen:

Jede Klasse besteht aus gleich breiten Abschnitten.

$$\text{Klassenbreite} = \frac{\text{Spannweite}}{\text{Anzahl_der_Klassen}}$$

Prozentual gleich starke Klassen

Jede Klasse besitzt die gleiche Anzahl von n.

$$\text{Klassenhäufigkeit} = \frac{n}{\text{Anzahl_der_Klassen}}$$

Inhaltliche Klassenbildung

Sinnvolle Kombination der Auswertungsmöglichkeiten.

Klassen sind inhaltlich vorgegeben, z.B. bei Blutwerten Grenzwerte für normal und pathologisch.

Grafiken der deskriptiven Statistik

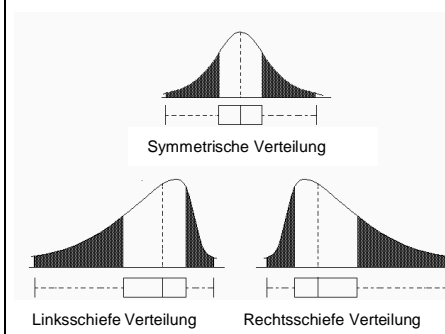
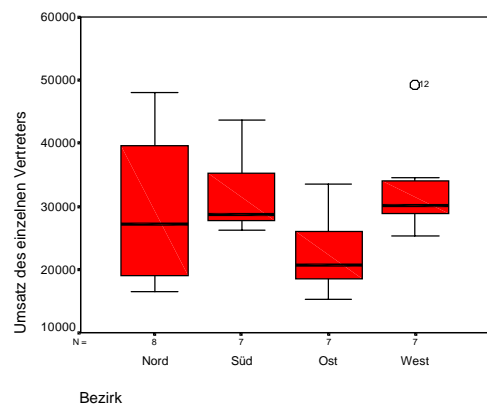
Merkmalstyp	Grafiktyp																								
<p>Skalierung: nominal, ordinal, diskret</p> <p>Anwendungstipp: max. 7 Ausprägungen nicht für feine Unterschiede geeignet</p>	<p>Kreisdiagramm</p> <p>Wie oft nutzen Sie das Internet?</p> <table border="1"><thead><tr><th>Nutzungsfrequenz</th><th>Anteil</th></tr></thead><tbody><tr><td>täglich</td><td>45,0%</td></tr><tr><td>1-4x in der Woche</td><td>40,0%</td></tr><tr><td>5-10x im Monat</td><td>10,0%</td></tr><tr><td>Fehrend</td><td>5,0%</td></tr></tbody></table>	Nutzungsfrequenz	Anteil	täglich	45,0%	1-4x in der Woche	40,0%	5-10x im Monat	10,0%	Fehrend	5,0%														
Nutzungsfrequenz	Anteil																								
täglich	45,0%																								
1-4x in der Woche	40,0%																								
5-10x im Monat	10,0%																								
Fehrend	5,0%																								
<p>Skalierung: ordinal, nominal, diskret</p> <p>auch: Gruppen- & Komponentendiagramm 2. stetiges Merkmal auf der y-Achse</p>	<p>Stabdiagramm</p> <p>Geschlecht der Studierenden</p> <table border="1"><thead><tr><th>Geschlecht</th><th>Absolute Werte</th></tr></thead><tbody><tr><td>Fehrend</td><td>4</td></tr><tr><td>weiblich</td><td>30</td></tr><tr><td>männlich</td><td>36</td></tr></tbody></table>	Geschlecht	Absolute Werte	Fehrend	4	weiblich	30	männlich	36																
Geschlecht	Absolute Werte																								
Fehrend	4																								
weiblich	30																								
männlich	36																								
<p><i>siehe Stabdiagramm!</i></p> <p>Besonders geeignet für nominale Skalierungen. Beschriftung ist sehr gut lesbar und wird manchmal nach Häufigkeiten sortiert. Auch als Gruppen- & Komponenten-diagramm.</p>	<p>Balkendiagramm</p> <p>Fachrichtung</p> <table border="1"><thead><tr><th>Fachrichtung</th><th>weiblich</th><th>männlich</th><th>Gesamt</th></tr></thead><tbody><tr><td>Allgemeine Dokumenta...</td><td>4</td><td>4</td><td>8</td></tr><tr><td>Wirtschaftsinformati...</td><td>2</td><td>6</td><td>8</td></tr><tr><td>E-Technik</td><td>1</td><td>7</td><td>8</td></tr><tr><td>Jura</td><td>10</td><td>2</td><td>12</td></tr><tr><td>Geschichte & Politik</td><td>4</td><td>4</td><td>8</td></tr></tbody></table> <p>Absolute Werte</p>	Fachrichtung	weiblich	männlich	Gesamt	Allgemeine Dokumenta...	4	4	8	Wirtschaftsinformati...	2	6	8	E-Technik	1	7	8	Jura	10	2	12	Geschichte & Politik	4	4	8
Fachrichtung	weiblich	männlich	Gesamt																						
Allgemeine Dokumenta...	4	4	8																						
Wirtschaftsinformati...	2	6	8																						
E-Technik	1	7	8																						
Jura	10	2	12																						
Geschichte & Politik	4	4	8																						
<p>Skalierung: stetig</p> <p>verwendet einen Teil der Datenwerte als Stem / Stamm, um Gruppen / Klassen zu bilden, und den anderen Teil als Leaf / Blatt, um die Anzahl der Häufigkeiten graphisch zu verdeutlichen. Vorteil: Man sieht die aktuellen Datenwerte, die gleichzeitig graphisch dargestellt sind.</p>	<p>Stem & Leaf Für die erste Zeile ALTERSANGABE: 40, 41, 41, 42, 43, 44</p> <table><tr><td>4</td><td>011234</td></tr><tr><td>4</td><td>566677</td></tr><tr><td>5</td><td>1122233344</td></tr><tr><td>5</td><td>55566677899</td></tr><tr><td>6</td><td>00111222333455</td></tr><tr><td>6</td><td>55677889999999</td></tr></table>	4	011234	4	566677	5	1122233344	5	55566677899	6	00111222333455	6	55677889999999												
4	011234																								
4	566677																								
5	1122233344																								
5	55566677899																								
6	00111222333455																								
6	55677889999999																								

Skalierung:
stetig

Grafik, in der mehrere Maßzahlen dargestellt werden: unteres Quartil, Median, oberes Quartil, Maximum, Minimum. In manchen Versionen werden Extremwerte extra gekennzeichnet.

Anhand des Box-Plots lässt sich auf die Verteilungsart der Daten schließen.

Box-Plot

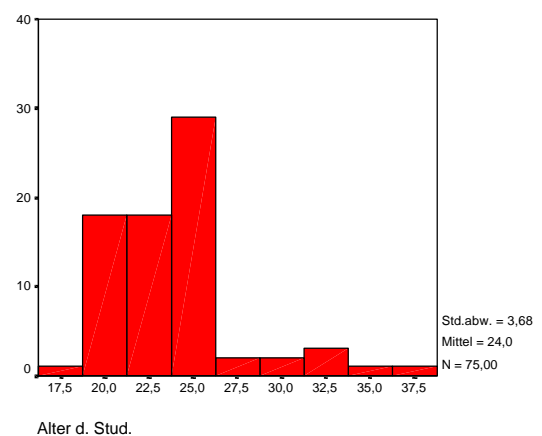


Skalierung:
stetig

Grafische Darstellung einer Häufigkeitsverteilung eines stetigen klassierten Merkmals. Das Maß für die Häufigkeit ist in der Regel die Fläche.

Anwendungstipp:
Mehrere Klasseneinteilungen ausprobieren!

Histogramm

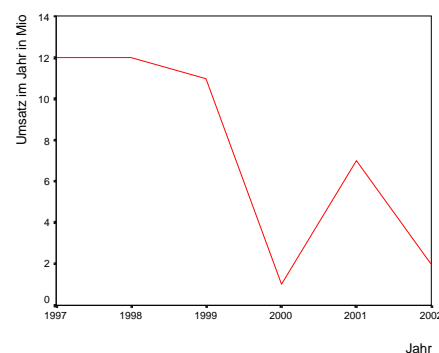


Skalierung:

y-Achse:
stetige oder diskrete mit vielen Ausprägungen.

x-Achse: i.d.R. findet sich hier der zeitliche Aspekt wieder.

Verlaufskurve



Analytische Statistik

Wahrscheinlichkeitsrechnung

Ereignisraum (\Rightarrow Menge aller möglichen Elementar - Ereignisse)

A, B, C	Ereignisse
$P(A)$	Wahrscheinlichkeit dafür, dass Ereignis A eintritt.
\bar{A}	komplementäres Ereignis zu A (Gegenereignis)
$P(A \cup B)$	Wahrscheinlichkeit dafür, dass Ereignis A oder Ereignis B oder beide eintreten (Vereinigung \rightarrow Additionssatz).
$P(A \cap B)$	Wahrscheinlichkeit dafür, dass Ereignis A und Ereignis B eintreten (Schnittmenge \rightarrow Multiplikationssatz).

Definitionen der Wahrscheinlichkeit

Klassische Definitionen (Laplace'sche Definition):

Wenn die Elementarereignisse endlich und gleichmöglich sind, gilt:

$$P(A) = \frac{\text{Anzahl_aller_günstigen_Elementarereignisse}}{\text{Anzahl_aller_gleichmöglichen_Elementarereignisse}}$$

Statistische Definition:

$P(A)$ wird als diejenige Größe definiert, der sich die relative Häufigkeit eines Ereignisses A bei unbeschränkter Versuchsserie ($n \rightarrow \infty$) nähert.

Axiomatische Definition nach Kolmogoroff:

Axiom 1: $0 \leq P(A) \leq 1$

Axiom 2: $P(S) = 1$ sicheres Ereignis S

Axiom 3: $P(E) = 0$ unmögliches Ereignis E

Additionssatz

Allgemeiner Additionssatz (für beliebige Ereignisse):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Spezieller Additionssatz (für disjunkte Ereignisse):

Disjunkte Ereignisse sind unvereinbare Ereignisse, die nicht gleichzeitig auftreten können.

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$P(A \cup \bar{A}) = P(S) = 1$$

$$P(\bar{A}) = 1 - P(A) \quad \text{Das zu einem Ereignis A komplementäres Ereignis } \bar{A} \text{ ist das Ereignis, das eintritt, wenn A nicht eintritt.}$$

Stochastische (Un-)Abhängigkeit von Ereignissen:

Gegeben seien die Ereignisse A,B,C mit $P(A)>0$, $P(B)>0$ und $P(C)>0$.

Ereignis B ist stochastisch (un-)abhängig von Ereignis A, wenn

$$P(B|A)=P(B|\bar{A})=P(B)$$

Die Ereignisse A, B, C sind voneinander stochastisch unabhängig, wenn

$$P(A|B)=P(A|C)=P(A|B \cap C)=P(A)$$

$$P(B|A)=P(B|C)=P(B|A \cap C)=P(B) \text{ und}$$

$$P(C|A)=P(C|B)=P(C|A \cap B)=P(C)$$

Bedingte Wahrscheinlichkeit

Sie ist die Wahrscheinlichkeit dafür, dass B eintritt, unter der Bedingung dass A mit $P(A) \neq 0$ bereits eingetreten ist.

$$P(B|A)=\frac{P(A \cap B)}{P(A)} \text{ mit } P(A)>0$$

$$P(A|B)=\frac{P(A \cap B)}{P(B)} \text{ mit } P(B)>0$$

Multiplikationssatz**Allgemeiner Multiplikationssatz:**

$$P(A \cap B)=P(A) \cdot P(B|A)=P(B) \cdot P(A|B)$$

$$P(A \cap B \cap C)=P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$

Spezieller Multiplikationssatz:

Für stochastisch unabhängige Ereignisse A, B und C gilt:

$$P(A \cap B)=P(A) \cdot P(B)$$

$$P(A \cap B \cap C)=P(A) \cdot P(B) \cdot P(C) \text{ für beliebig viele Ereignisse}$$

Kombinatorik

Fakultät

$$n! = 1 \cdot 2 \cdot \dots \cdot (n-1) \cdot n$$
$$0! = 1$$

Binomialkoeffizient (sprich n über k)

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{0} = 1$$

Für Berechnung des Binomialkoeffizienten mit Taschenrechner ist die Taste **nCr** zu benutzen.

Wahrscheinlichkeitsverteilungen

Verteilungsfunktion

⇒ definiert für alle reellen Zahlen x_i

$$F(x_i) = P(X \leq x_i)$$

F(x_i) Verteilungsfunktion

P Wahrscheinlichkeit

X Zufallsvariable (z.B. Zuweisen der Augenzahl beim Würfeln)

x_i konkrete Ausprägungen

~ verteilt wie

Binomialverteilung (anzuwenden für kleine Stichproben)

Voraussetzungen für die Anwendungen der Binomialverteilung

- Merkmal ist dichotom (besitzt genau 2 Ausprägungen: A und \bar{A})
- n unabhängige Wiederholungen des Zufallsexperiments werden durchgeführt (z.B. Ziehen von Kugeln mit Zurücklegen)
- $p = P(A)$ ist bekannt

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

P: Wahrscheinlichkeit, dass unter n (unabhängigen) Versuchen genau k mal Ereignis A auftritt

wobei:

X Zufallsvariable

k Anzahl der Treffer

p Wahrscheinlichkeit einen Treffer zu ziehen

n Anzahl der Versuche

Kurzbezeichnungsweise (sprich X verteilt wie/nach Binomialverteilung): **$X \sim B(n,p)$**

Gauß- oder Normalverteilung $N(\mu, \sigma^2)$

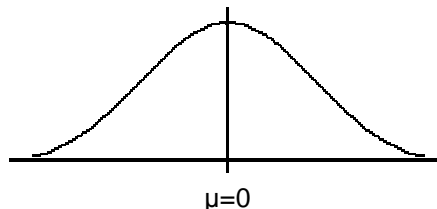
Wird eindeutig festgelegt durch:

μ Mittelwert in der GG
 σ^2 Varianz in der GG

- innerhalb der Entfernung von 1 Standardabweichung σ nach rechts und links vom Mittelwert μ befinden sich **68,26%** der Messwerte
- innerhalb der Entfernung von 2 Standardabweichungen σ nach rechts und links vom Mittelwert μ befinden sich **95,45%** der Messwerte
- innerhalb der Entfernung von 3 Standardabweichungen σ nach rechts und links vom Mittelwert μ befinden sich **99,73%** der Messwerte

Standardisierte Normalverteilung $N(0,1)$

ist vertafelt



Standardisierung einer beliebigen Zufallsvariablen $X \sim N(\mu, \sigma^2)$

$$\frac{X - \mu}{\sigma} = U \quad \text{wobei } U \sim N(0,1)$$

Transformation der Standardnormalverteilung in eine beliebige Normalverteilung

$$U \sim N(0,1) \rightarrow U \cdot \sigma + \mu \sim N(\mu, \sigma^2)$$

Parameter in der Grundgesamtheit

Zur Unterscheidung von der Stichprobe werden die Parameter der Grundgesamtheit mit griechischen Buchstaben bezeichnet:

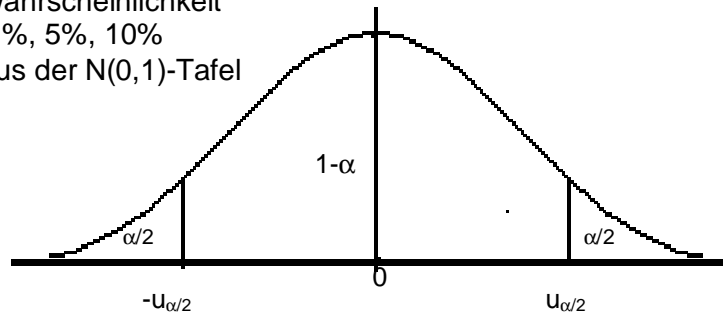
μ Mittelwert in der Grundgesamtheit
 σ Streuung in der Grundgesamtheit
 π Prozentsatz in der Grundgesamtheit

Konfidenzintervalle

$1-\alpha$ Sicherheits- oder Vertrauenswahrscheinlichkeit
üblich: 99%, 95%, 90%

α Irrtumswahrscheinlichkeit
üblich: 1%, 5%, 10%

$u_{\alpha/2}$ Werte aus der N(0,1)-Tafel



Konfidenzintervall für einen unbekannten Mittelwert μ einer Grundgesamtheit

$$P\left(\bar{x} - u_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + u_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Beispiel für 95%-ige Sicherheitswahrscheinlichkeit:

$$P\left(\bar{x} - 1,96 \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1,96 \cdot \frac{s}{\sqrt{n}}\right) = 95\%$$

Sonderregel:

Umfasst die Stichprobe mehr als 5% der Grundgesamtheit (große Stichprobe), benutze

$$\text{statt } \frac{s}{\sqrt{n}} \quad \text{besser } \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Konfidenzintervall für eine unbekannte Proportion π einer Grundgesamtheit

$$P\left(p - u_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + u_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Voraussetzung: $n \cdot p \geq 5$ wenn $p \leq 0,5$
 $n \cdot (1-p) \geq 5$ wenn $p > 0,5$

Beispiel für 95%-ige Sicherheitswahrscheinlichkeit:

$$P\left(p - 1,96 \cdot \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1,96 \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 95\%$$

Sonderregel:

Umfasst die Stichprobe mehr als 5% der Grundgesamtheit (große Stichprobe), benutze

$$\text{statt } \sqrt{\frac{p(1-p)}{n}} \quad \text{besser } \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Fallzahlbestimmung

Bestimmung des erforderlichen (Mindest-) Stichprobenumfangs bei Schätzung eines (prozentualen) Anteils π

$$n \geq \frac{u_{\alpha/2}^2 \cdot p \cdot (1-p)}{e^2}$$

n	die erforderliche Mindestfallzahl
1- α bzw. α	Sicherheits- bzw. Irrtumswahrscheinlichkeit
$u_{\alpha/2}$	aus Tabelle von N(0,1)
p	„Erfahrungswert“ für ungefähre Größenordnung des %-Anteils existiert kein Wert, setze p=0,5
e	Exaktheit Abstand zwischen Stichprobenschätzung p und wahren π der Grundgesamtheit $e = \pi - p $

Bestimmung der erforderlichen (Mindest-)Fallzahl n bei der Schätzung eines (unbekannten) Mittelwertes μ

$$n \geq \frac{u_{\alpha/2}^2 \cdot \hat{\sigma}^2}{e^2}$$

n	erforderliche Mindestfallzahl
1- α bzw. α	Sicherheits- bzw. Irrtumswahrscheinlichkeit
e	Exaktheit $e = \mu - \bar{x} $ Unterschied zwischen Stichprobenmittelwert \bar{x} und wahren μ der Grundgesamtheit
$\hat{\sigma}^2$	Schätzung der Varianz des betrachteten stetigen Merkmals falls keine Schätzung existiert, berechne $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$

A	
Additionssatz	17
Allgemeiner Additionssatz	17
Allgemeiner Multiplikationssatz	18
Analytische Statistik Definition	4
Analytische Statistik	17
Arithmetisches Mittel	7

B	
Balkendiagramm	15
Bedingte Wahrscheinlichkeit	18
Binomialkoeffizient	19
Binomialverteilung	19
Box-Plot	16

D	
Deskriptive Statistik Definition	4
Deskriptive Statistik	7
Diagramme	15
Diskrete Variablen	6

E	
Ereignisraum	17

F	
Fakultät	19
Fallzahlbestimmung	22

G	
Gaußverteilung	18
Geometrisches Mittel	7
Grafiken	15
Grundbegriffe der Statistik	5
Grundgesamtheit	5

H	
Häufigkeiten	12
Häufigkeitstabelle	12
Histogramm	16

I	
Inhaltsverzeichnis	2f

K	
Klassenbildung	14
Kombinatorik	19
Konfidenzintervalle	21
Kreisdiagramm	15
Kreuztabelle	13

M	
Median	9
Merkmal	5
Merkmalsausprägungen	6
Merkmalsträger	5
Mittelwerte	7

Modus	9
Multiplikationssatz	18

N	
Nominale Variablen	6
Normalverteilung	20

O	
Ordinale Variablen	6

P	
Parameter in der Grundgesamtheit	18

Q	
Quantitative Merkmale	6
Quantitative Merkmale	6
Quartile	11
Quartilsabstand	11

S	
Schätzung des arithmetischen Mittels für klassierte Merkmale	8
Spannweite	10
Spezieller Additionssatz	17
Spezieller Multiplikationssatz	18
Stabdiagramm	15
Standardabweichung	10
Standardisierte Normalverteilung	20
Stem & Leaf	15
Stetige / metrische Variablen	6
Stichprobe	5
Stichprobenumfangs	22
Stochastische (Un-)Abhängigkeit	18
Streuungsmaße	10

T	
Transformation der Standardnormalverteilung	20

U	
Urliste	5

V	
Varianz	10
Variationskoeffizient	11
Verlaufskurve	16
Verteilungsfunktion	19
Vollerhebung	5

W	
Wahrscheinlichkeit	17
Wahrscheinlichkeitsrechnung	17
Wahrscheinlichkeitsverteilungen	19

Z	
Zufallsvariable	20