Kommentierte Formelsammlung der deskriptiven und induktiven Statistik für Sozialwissenschaftler

Prof. Dr. Irene Rößler

Prof. Dr. Albrecht Ungerer

Inhaltsverzeichnis

1 Grundlagen

Statistik als Methodenlehre und nicht als Zahlenergebnis verstanden ist eine wissenschaftliche Disziplin, die sich mit der Entwicklung und Anwendung von Verfahren zur Gewinnung, Beschreibung und Analyse von in Zahlen abbildbaren empirischen Befunden beschäftigt. Sie soll in einem Entscheidungsprozess informative Daten liefern; insbesondere soll sie helfen, Theorien an der Realität zu überprüfen.

Phasen einer statistischen Erhebung

- Fragestellung (Formulierung einer praktischen Entscheidung oder wissenschaftlichen Theorie so, dass eine statistische Messung möglich ist: Grundprobleme der „empirischen Sozialforschung“)
- Festlegung der statistischen (Grund-) Gesamtheit [Bestimmung der sachlichen, zeitlichen (Zeitpunkt: Bestandsmasse; Zeitraum: Bewegungsmasse) und räumlichen Identifikationsmerkmale]
- Wahl der Erhebungsmerkmale und insbesondere bei nominalen und ordinalen Merkmalen Entwurf einer Messskala
- Wahl des Erhebungsverfahrens (z.B. schriftliche bzw. mündliche Befragung, Beobachtung, Experiment; Primär- oder Sekundärerhebung; Voll- oder Teilerhebung)
- Organisation, Durchführung und Kontrolle
- Aufbereitung der Daten (Ordnen, Datenverdichtung)
- Auswertung (Datenanalyse und Interpretation der Ergebnisse bezüglich der Fragestellung unter Berücksichtigung des Einflusses der Phasen der Datenentstehung)
- Darstellung der Ergebnisse (tabellarische und grafische Darstellung)

Gestaltungsbeschränkung durch Rahmenbedingungen (z.B. rechtliche) und ein „ökonomisches Prinzip“ (Abwägung: aktuell–billig–genau).

Merkmalsarten und Skalen

Merkmalsart		Skala	Interpretation	Transformation	Beispiel
qualitativ	rein qualitativ	Nominalskala	1. Verschiedenartigkeit	eineindeutige Transformationen	Beruf, Fachrichtung, Familienstand, Geschlecht, Körpergröße(?)
	komparativ	Ordinalskala	1. Verschiedenartigkeit 2. Ordnung	streng monotone Transformationen	Note, Kreditranking, Zufriedenheitsgrad, soziale Schicht, Körpergröße(?)
quantitativ		Intervallskala	1. Verschiedenartigkeit 2. Ordnung 3. Differenzen	lineare Transformationen $y = ax + b$, $a > 0$	°Celsius, Normabweichung, Altersjahrgang, Körpergröße(?)
		Verhältnisskala	1. Verschiedenartigkeit 2. Ordnung 3. Differenzen 4. Verhältnisse	linear-homogene Transformationen $y = ax$, $a > 0$	°Kelvin, Alter in Jahren, Einkommen, Preis, Körpergröße

Regeln für die Erstellung von Tabellen

1. Jede Tabelle trägt eine Überschrift, in der die beschriebene statistische Masse sachlich, zeitlich und räumlich abzugrenzen ist.
2. Tabellenkopf und die Vorspalte enthalten die Erläuterung zum Zahlenteil. Jede Zahl im Zahlenteil ist somit charakterisiert durch die jeweilige Zeilen- (in der Vorspalte) und Spaltenbezeichnung (im Tabellenkopf). Kein Tabellenfeld sollte leer sein. Dabei bedeutet „–" genau Null, während „0" mehr als Null, aber weniger als die Hälfte der kleinsten Darstellungseinheit bedeutet (auch 0,0 oder 0,00).
3. Fußnoten enthalten Erläuterungen zum Inhalt einer Tabelle sowie Quellenhinweise.

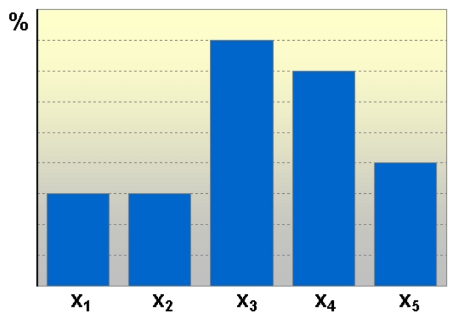
Bsp.: Tab … Wohnbevölkerung der Stadt XY am 30.02.20.. (in Tsd.)

Geschlecht	Familienstand				Insgesamt
	ledig	verheiratet	verwitwet	geschieden	
männl.	102	89	5	4	200
weibl.	109	90	15	6	220
Insgesamt	211	179	20	10	420

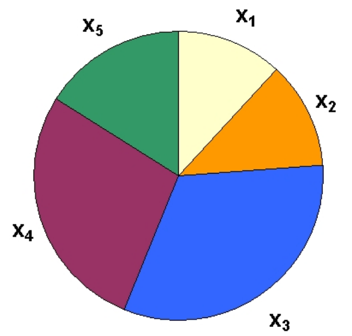
Quelle: Städtestatistisches Amt XY

Grundformen grafischer Darstellungen

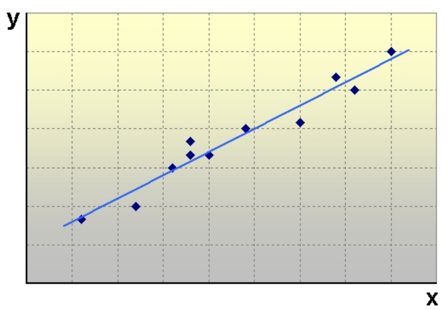
Balkendiagramm



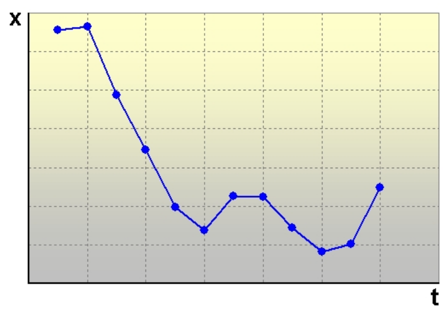
Flächendiagramm



Streuungsdiagramm



Kurvendiagramm



Aufgabe 1

Erstellen Sie ein Kreisdiagramm des Merkmals Familienstand für das obige Beispiel der Wohnbevölkerung.

Beispieldatensatz

Bei 25 Teilnehmern einer Statistik-Klausur wird eine statistische Erhebung mit den Merkmalen

- Hauptfach (Sonst. 1, Soz. 2, Pol. 3)
- Studienjahr (1, 2, 3)
- Ausgaben für Kopien im letzten Semester (Euro)
- durchschnittliches Monatseinkommen im letzten Semester (Euro)
- erwartete Leistung (unterdurchschnittlich -1, durchschnittlich 0, eher besser +1)

durchgeführt. Man erhält folgende Datenmatrix:

Stud.-Nr.	Fach	Jahr	Ausgaben für Kopien €	Einkommen €	erwartete Leistung
1	2	2	21	675	0
2	1	2	37	740	0
3	2	1	26	710	-1
4	3	2	68	860	+1
5	2	3	16	590	0
6	2	1	31	720	0
7	3	3	24	710	-1
8	3	1	6	570	+1
9	2	1	22	660	-1
10	3	3	32	760	+1
11	1	2	17	675	0
12	3	2	44	775	0
13	1	2	30	750	-1
14	2	1	12	600	+1
15	3	3	57	820	-1
16	3	2	41	805	-1
17	1	1	20	630	+1
18	2	1	19	670	0
19	3	3	47	790	-1
20	3	1	14	655	+1
21	2	2	39	745	0
22	3	1	18	660	+1
23	2	1	2	590	+1
24	2	2	10	640	0
25	1	2	27	700	+1

2 Deskriptive Statistik: Univariate Verteilungen

2.1 Darstellungsformen

Die erste Stufe einer Auswertung erhobener Daten umfasst die sinnvolle Ordnung der Merkmalswerte bzw. ihre Zusammenfassung zu Gruppen mit gleichen Merkmalsausprägungen. Die tabellarische oder grafische Darstellung der Häufigkeiten des Auftretens von Merkmalsausprägungen heißt Häufigkeitsverteilung.

Begriffe	**Symbole**
Statistische Masse (Grundgesamtheit) besteht aus statistischen Einheiten mit denselben Identifikationsmerkmalen.	Umfang: n (N) durchnummerierte (verschlüsselte, anonymisierte) statistische Einheiten: $i = 1, 2, \ldots, n\,(N)$
Urliste enthält Beobachtungswerte des Merkmals X von n statistischen Einheiten.	$a_1, \ldots, a_i, \ldots, a_n$
Merkmalsausprägungen des Merkmals X	$x_1, \ldots, x_j, \ldots, x_m$
absolute Häufigkeit der Ausprägung x_j	$h_j = h(x_j)$ mit $\sum\limits_{j=1}^{m} h_j = n$
relative Häufigkeit von x_j	$f_j = f(x_j) = \frac{h_j}{n}$ mit $\sum\limits_{j=1}^{m} f_j = 1$
relative Häufigkeitsfunktion	$f(x) = \begin{cases} f_j & \text{für } x = x_j,\ j = 1, \ldots, m \\ 0 & \text{sonst} \end{cases}$
kumulierte absolute Häufigkeit von x_j des mindestens ordinalen Merkmals X	$H_j = H(x_j)$ mit $H_j = \sum\limits_{k=1}^{j} h_k,\ x_k \leq x_j,\ H_m = n$
kumulierte relative Häufigkeit von x_j des mindestens ordinalen Merkmals X	$F_j = F(x_j)$ mit $F_j = \sum\limits_{k=1}^{j} f_k = \frac{H_j}{n},\ x_k \leq x_j,\ F_m = 1$
Empirische Verteilungsfunktion	$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ F_j & \text{für } x_j \leq x < x_{j+1},\ j = 1, \ldots, m-1 \\ 1 & \text{für } x \geq x_m \end{cases}$

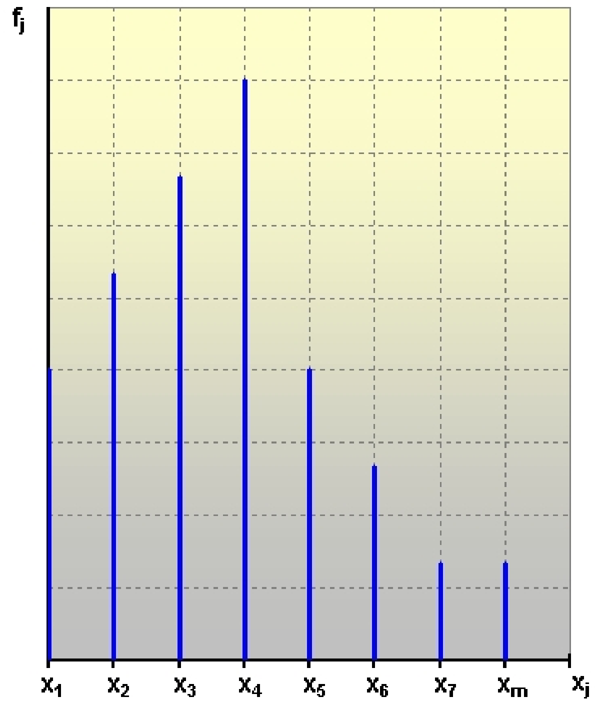
Aufgabe 2

Bei einer Erhebung stellt man folgende Personenzahl je Wohnung in den 40 Sozialwohnungen einer Stadt fest (Urliste):

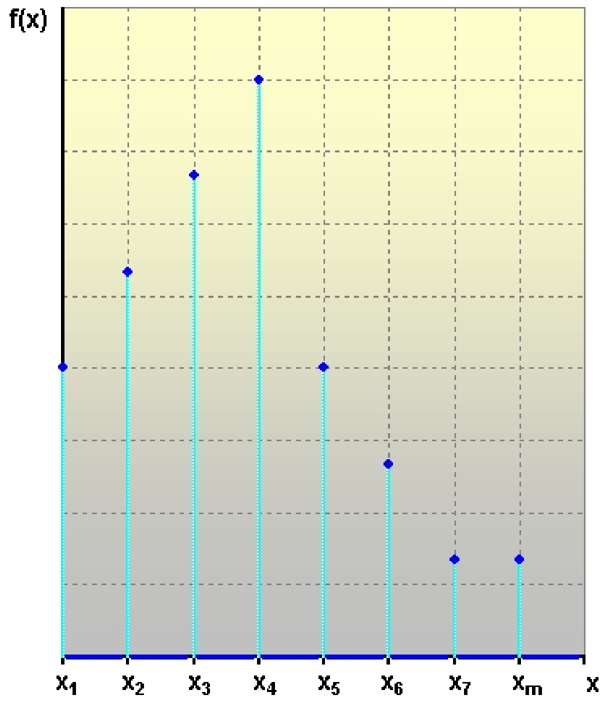
5,2,1,4,6, 3,2,4,4,7, 6,1,2,3,5, 3,3,4,3,3 0,5,2,4,3, 3,6,5,6,4, 3,5,3,4,3, 3,5,7,3,4.

Berechnen Sie in tabellarischer Form absolute und relative Häufigkeiten sowie die kumulierten Häufigkeiten. Zeichnen Sie die Häufigkeitsverteilungen.

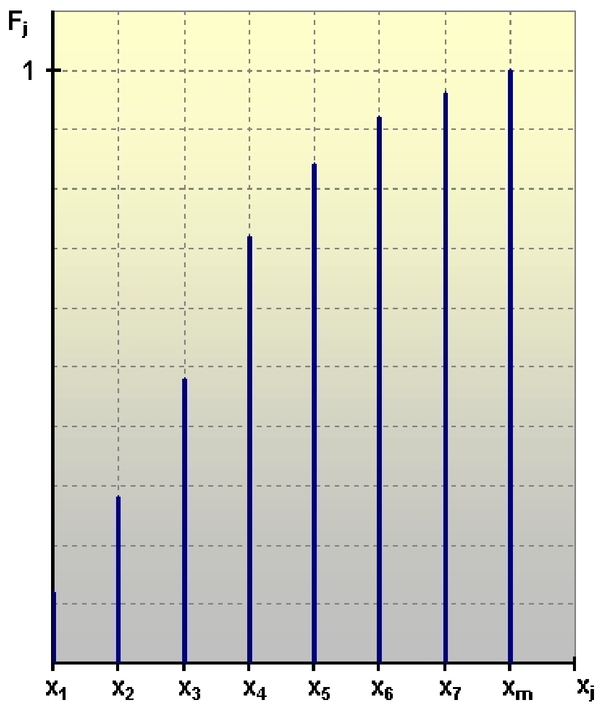
Relative Häufigkeiten



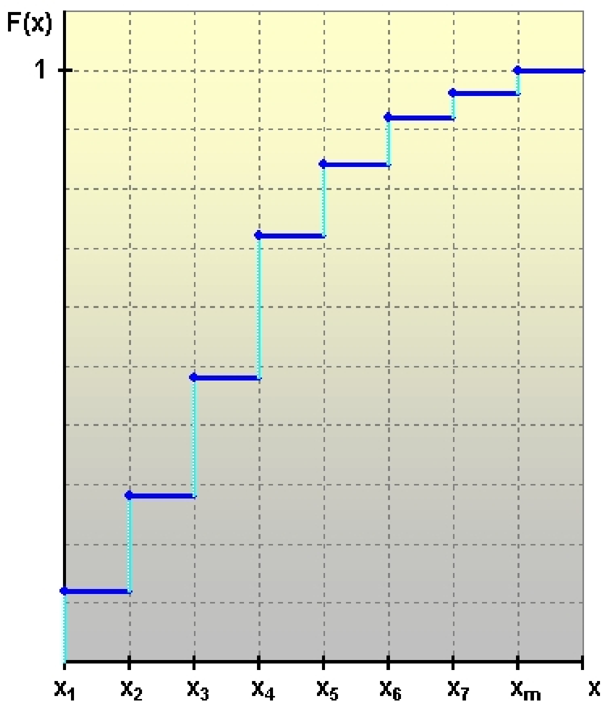
Relative Häufigkeitsfunktion



Kumulierte relative Häufigkeiten



Empirische Verteilungsfunktion



Klassierte Daten, Histogramm

Bei quantitativen Merkmalen mit sehr vielen Ausprägungen (z.B. Einkommen) oder bei stetigen Merkmalen werden zur Erhebung bzw. vor der Auszählung benachbarte Beobachtungswerte zu Klassen zusammengefasst. Die Klassengrenzen dürfen sich nicht überschneiden. Die Wahl der Klassenbreiten hängt einerseits von der Erhebbarkeit, andererseits vom gewünschten Informationsgehalt und der Klassenbesetzung ab. Weisen die Klassen eine unterschiedliche Breite auf, so werden zur Vermeidung von Missverständnissen die Klassenhäufigkeiten auf die Klassenbreiten bezogen. Als Ergebnis erhält man die besser vergleichbaren Besetzungsdichten je Klasse. Diese werden in Histogrammen auf der Ordinate abgetragen, die Häufigkeiten somit als Rechteckflächen dargestellt. Die Dichtefunktionen innerhalb der Klassen entsprechen also Rechteckverteilungen (einfachstes Modell).

Begriffe	**Symbole**
m Klassen (von ... bis unter ...)	$[a_1,b_1),\ldots,[a_j,b_j),\ldots,[a_m,b_m)$
Klassenbreite / Klassenmitte	$w_j = b_j - a_j \;\Big/\; \tilde{x}_j = \dfrac{a_j+b_j}{2}$
absolute / relative Häufigkeit	$h_j = \sum\limits_{x_i\in[a_j,b_j)} h(x_i)$ mit $\sum\limits_{j=1}^{m} h_j = n \;\Big/\; f_j = \dfrac{h_j}{n}$ mit $\sum\limits_{j=1}^{m} f_j = 1$
absolute / relative Dichte	$h_j^* = \dfrac{h_j}{w_j}$ mit $\sum\limits_{j=1}^{m} h_j^* w_j = n \;\Big/\; f_j^* = \dfrac{f_j}{w_j}$ mit $\sum\limits_{j=1}^{m} f_j^* w_j = 1$
Klassierte Häufigkeitsdichte	$f^*(x) = \begin{cases} f_j^* & \text{für } x\in[a_j,b_j),\ j=1,\ldots,m \\ 0 & \text{sonst} \end{cases}$ mit $\int\limits_{a_1}^{b_m} f^*(x)dx = 1$
kumulierte abs. / rel. Häufigk.	$H_j = \sum\limits_{k=1}^{j} h_k$ mit $H_m = n \;\Big/\; F_j = \sum\limits_{k=1}^{j} f_k = \dfrac{H_j}{n}$ mit $F_m = 1$
Klassierte Verteilungsfunktion	$F^*(x) = \int\limits_{a_1}^{x} f^*(u)du = \begin{cases} 0 & \text{für } x < a_1 \\ F_{j-1} + f_j^*(x-a_j) & \text{für } x\in[a_j,b_j),\ j=1,\ldots,m \\ 1 & \text{für } x \geq b_m \end{cases}$

Aufgabe 3

Einkommen von ... bis unter ... €	%
0 – 500	10
500 – 1.000	25
1.000 – 1.250	25
1.250 – 1.500	15
1.500 – 2.000	15
2.000 – 3.000	5
3.000 – 5.000	5

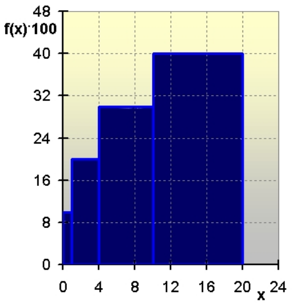
Zeichnen Sie ein Histogramm und die klassierte Verteilungsfunktion. Schätzen Sie nach der Grafik wieviel Prozent weniger als 1.150 € verdienen.

Histogramm

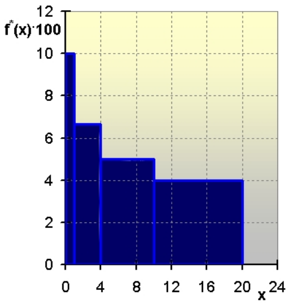
Ergebnis einer Schnellinventur

Teile von ... bis unter ... €			$f_j \cdot 100$	$f_j^* \cdot 100$
0	–	1	10	10
1	–	4	20	6,7
4	–	10	30	5
10	–	20	40	4

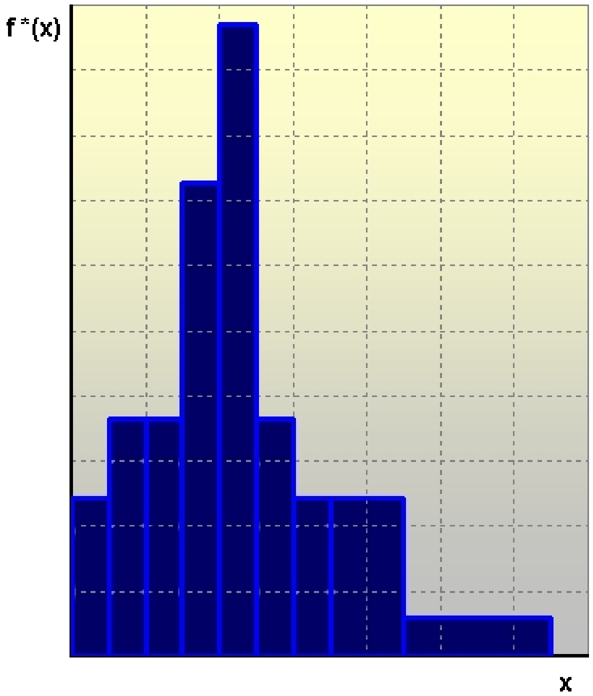
falsche Darstellung:



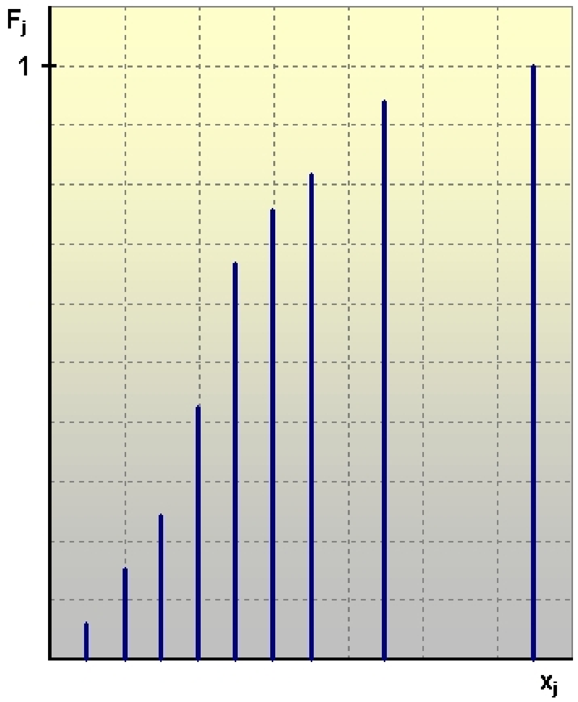
richtige Darstellung:



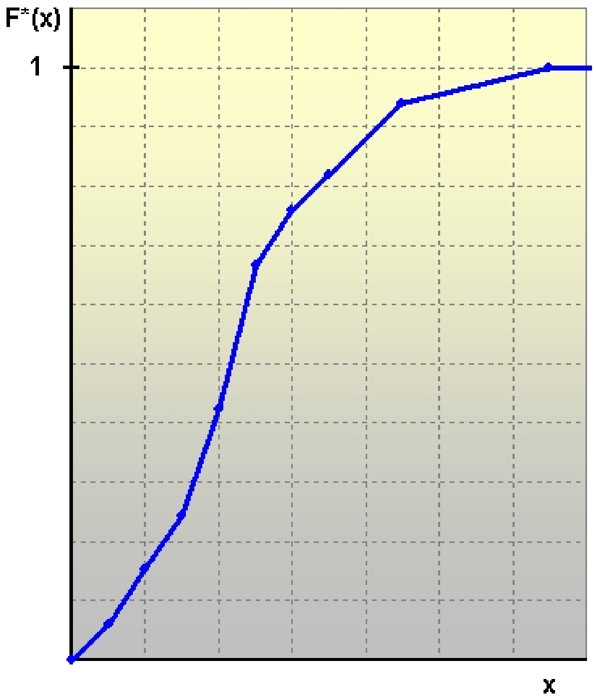
Histogramm



Kumulierte relative Häufigkeiten



Klassierte Verteilungsfunktion

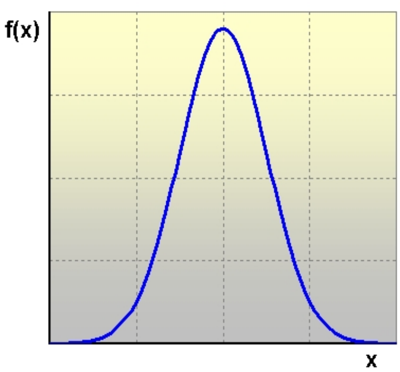
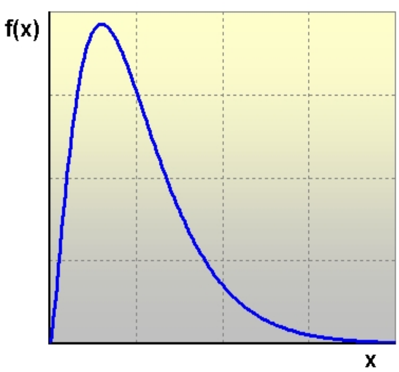
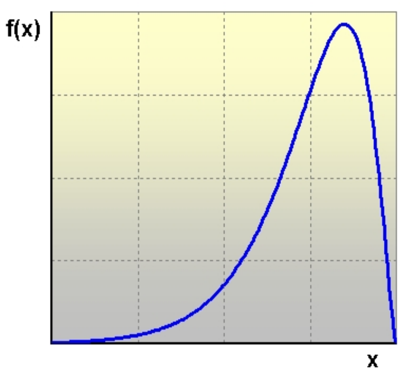


2.2 Maßzahlen der zentralen Tendenz

In der zweiten Stufe der Auswertung werden Beobachtungswerte bzw. Häufigkeitsverteilungen zu Maßzahlen verdichtet. Im Sachzusammenhang sinnvolle Maßzahlen sollen so u.a. – sofern sie nicht selbst Untersuchungsziel sind – einen übersichtlichen Vergleich verschiedener statistischer Reihen erlauben.

Mittelwerte	Symbol	Berechnung	Skalenniveau	Aussage
Modus (häufigster Wert, Dichtemittel)	D	$D = x_k$ mit $h_k = \max_j h_j$	beliebig	Die Merkmalsausprägung einer Verteilung, auf die die meisten Beobachtungswerte entfallen.
Median (Zentralwert, 2. Quartil)	Z	$Z = a_k$ mit $k = \frac{n+1}{2}$ für n ungerade und $k = \frac{n}{2}$ für n gerade, a_i der Größe nach geordnet. Für $Z = x_j$ gilt: $F(x_j) = 0,5$.	ordinal oder metrisch	Der Beobachtungswert einer der Größe nach geordneten Reihe, unterhalb dem die Hälfte aller Merkmalsträger liegt. Echte „Mitte“. Bei Verteilungen mit nur wenig Beobachtungswerten als Deskription oft nicht sinnvoll.
Arithmetisches Mittel	$\bar{x}$ (μ)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i$ $= \frac{1}{n}\sum_{j=1}^{m} h_j x_j$ $= \sum_{j=1}^{m} f_j x_j$	metrisch	Die Größe, die sich ergibt, wenn die Merkmalssumme gleichmäßig auf die Merkmalsträger aufgeteilt wird. Zur Beschreibung der „Mitte“ einer Verteilung nur bei symmetrischen Verteilungen geeignet.

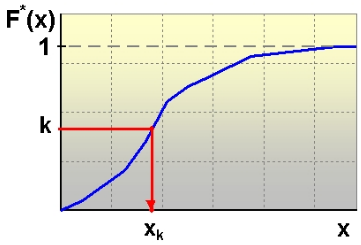
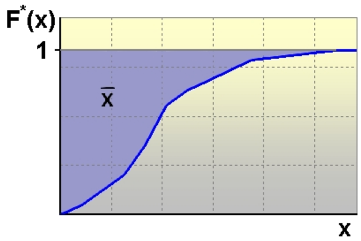
Mittelwerte und Verteilungsformen

symmetrisch	linkssteil	rechtssteil
		
$\bar{x} = D = Z$	$D < Z < \bar{x}$	$\bar{x} < Z < D$

Aufgabe 4

Berechnen Sie für die 2. Aufgabe die drei behandelten Mittelwerte.

Ergänzungen

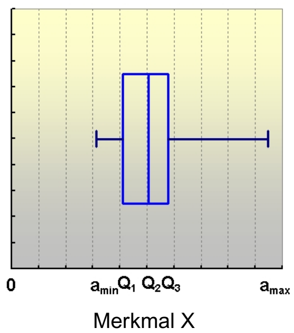
Modalklasse	$[a_D, b_D) = [a_k, b_k)$ mit $h_k^* = \max_j h_j^*$	Die am dichtesten besetzte Klasse.
Quantile z.B. • Perzentile • Dezile • Quartile Q_k	$F^*(x_k) = k$  1%-Schritte $(k \in \{1,2,\ldots,99\})$ 10%-Schritte $(k \in \{1,2,\ldots,9\})$ 25%-Schritte $(k \in \{1,2,3\})$	Die Merkmalsausprägung x_k, unterhalb der z.B. 99% der Werte (99. Perzentil) 90% der Werte (9. Dezil) 75% der Werte (3. Quartil) liegen.
Arithmetisches Mittel $\bar{x}$ • Hochrechnungseigenschaft • lineare Transformation • Arithmetisches Mittel aus arithmetischen Mitteln	 $n \cdot \bar{x} = \sum_{i=1}^{n} a_i = X$ $z_i = c + d \cdot a_i \implies \bar{z} = c + d \cdot \bar{x}$ $\bar{x} = \frac{1}{n} \sum_{j=1}^{m} n_j \bar{x}_j$ mit $n = \sum_{j=1}^{m} n_j$ $\bar{x} = \frac{1}{n} \sum_{j=1}^{m} h_j \tilde{x}_j$ mit $\tilde{x}_j = \frac{a_j + b_j}{2}$	Das arithmetische Mittel enthält als wichtigste Information die Merkmalssumme. Arithmetisches Mittel aus arithmetischen Mitteln von m Untergruppen. Schätzung des arith. Mittels bei klassierter Verteilung, falls $\bar{x}_j$ unbekannt.
Geometrisches Mittel g	$g = \sqrt[t]{\frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdots \frac{x_t}{x_{t-1}}} = \sqrt[t]{\frac{x_t}{x_0}}$ mit $\frac{x_t}{x_{t-1}}$: Messzahlen aus äquidistant gemessenen Größen	Durchschnittliche Wachstumsfaktoren wirtschaftsstatistischer Zeitreihen. [Z.B. Durchschnittsverzinsung bei Wiederanlage der Zinsen.]
Harmonisches Mittel h	$h = \frac{\sum g_i}{\sum g_i \cdot x_i^{-1}} = \left(\frac{\sum g_i \cdot x_i^{-1}}{\sum g_i}\right)^{-1}$ $\left[= \frac{\sum \mathrm{km}}{\sum \mathrm{km} \cdot \left(\frac{\mathrm{km}}{\mathrm{Std}}\right)^{-1}} = \frac{\sum \mathrm{km}}{\sum \mathrm{Std}} \right]$	Durchschnittsgrößen, wenn sich die gegebenen Gewichte auf Zählergrößen beziehen. [Z.B. Durchschnittsgeschwindigkeit, wenn die Gewichte Teilstrecken sind.]

Aufgabe 5

Berechnen Sie für die 3. Aufgabe die Modalklasse, die Quartile und das arithmetische Mittel.

2.3 Maßzahlen der Streuung

Maßzahlen der Streuung sollen die Variation der Einheiten in den Merkmalsausprägungen abbilden, bei quantitativen Merkmalen besonders bezüglich eines Mittelwerts. So gesehen sind sie auch eine Maßgröße für den Informationsgehalt eines Mittelwerts als Abbildungsergebnis einer statistischen Verteilung.

Streuungs-maße	Symbol	Berechnung	Skalen-niveau	Aussage
Homogeni-tätsindex	P	$P = \frac{m}{m-1}(1 - \sum_{j=1}^{m} f_j^2)$, $0 \leq P \leq 1$	beliebig	P ist bei der Gleichverteilung am größten und bei der Einpunktverteilung am geringsten.
Quartils-abstand • Box-and-Whisker Plot	QA	$QA = Q_3 - Q_1$ 	ordinal oder metrisch	QA gibt den mittleren Bereich der Beobachtungswerte einer der Größe nach geordneten Reihe an, unterhalb bzw. oberhalb dem je ein Viertel der Merkmalsträger liegt. Bei ordinalen Merkmalen nur sinnvoll, wenn nicht die Differenz ausgerechnet wird (so allerdings keine Maßzahl).
Varianz und Standard-abweichung	s^2 (σ^2) s (σ) $s = +\sqrt{s^2}$	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2$ $= \frac{1}{n}\sum_{i=1}^{n} a_i^2 - \bar{x}^2$ $s^2 = \frac{1}{n}\sum_{j=1}^{m} h_j(x_j - \bar{x})^2$ $= \frac{1}{n}\sum_{j=1}^{m} h_j x_j^2 - \bar{x}^2$	metrisch	s^2 ist ein Durchschnitt aus quadrierten Differenzen zwischen Beobachtungswert und dem arithmetischen Mittel. Größere Differenzen werden stärker gewichtet als kleine. Verschiebungssatz

Varianzzerlegung bei m Untergruppen $(j = 1, \ldots, m)$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{x})^2 = \underbrace{\frac{1}{n} \cdot \sum_{j=1}^{m}\sum_{i=1}^{n_j}(a_{ij} - \bar{x}_j)^2}_{s^2_{\text{int}}} + \underbrace{\frac{1}{n} \cdot \sum_{j=1}^{m} n_j(\bar{x}_j - \bar{x})^2}_{s^2_{\text{ext}}} = \frac{1}{n} \cdot \sum_{j=1}^{m} n_j \cdot s_j^2 + s^2_{\text{ext}} = s^2_{\text{int}} + s^2_{\text{ext}}$$

Die Gesamtvarianz lässt sich bei Einteilung einer Gesamtheit in Gruppen so zerlegen, dass ein Teil die Streuung der Einzelwerte innerhalb der Gruppen (s^2_{int}), der andere Teil die Streuung zwischen den Mittelwerten der Gruppen (s^2_{ext}) abbildet.

Aufgabe 6 Berechnen Sie für die 2. Aufgabe den Quartilsabstand und die Standardabweichung.

Aufgabe 7 Nehmen Sie eine Varianzzerlegung für das Hauptfach ($j = 1, 2, 3$) und die Ausgaben für Kopien (a_{ij}) des Beispieldatensatzes Seite 3 vor.

Ergänzungen

Spannweite R	$R = a_{\max} - a_{\min}$	Differenz zwischen größtem und kleinstem Beobachtungswert, z.B. bei Preis-/Kursentwicklungen.
Durchschnittliche (mittlere absolute) Abweichung d_A • Minimum-eigenschaft von d_Z	$d_A = \frac{1}{n}\sum_{i=1}^{n} \lvert a_i - A\rvert$ $= \frac{1}{n}\sum_{j=1}^{m} h_j \lvert x_j - A\rvert$ $= \sum_{j=1}^{m} f_j \lvert x_j - A\rvert,\ A = \bar{x}, Z, \ldots$ $d_A = \frac{1}{n}\sum_{i=1}^{n} \lvert a_i - A\rvert = \min$ für $A = Z$	Da $\sum_i (a_i - \bar{x}) = 0$ gilt (Schwerpunkteigenschaft des arith. Mittels), bildet man das arith. Mittel der Absolutbeträge der Abweichungen der Beobachtungswerte vom arith. Mittel ($A = \bar{x}$). Als Bezugspunkt der Abweichungen der Beobachtungswerte kann auch der Median Z oder ein anderer Mittelwert gewählt werden.
Varianz • Minimum-eigenschaft	$s_A^2 = \frac{1}{n}\sum_i (a_i - A)^2 = \min$ für $A = \bar{x}$ $s_A^2 = \frac{1}{n}\sum_i (a_i - \bar{x})^2 + (\bar{x} - A)^2$	Die mittlere quadratische Abweichung bezogen auf das arith. Mittel ist stets kleiner als die mittlere quadratische Abweichung bezogen auf einen beliebigen Wert A.
• lineare Transformation	$z_i = c + d \cdot a_i \implies s_Z^2 = d^2 \cdot s_X^2$ mit $s_X^2 = \frac{1}{n}\sum_i (a_i - \bar{x})^2$	
• z-Transformation (Standardisierung)	$z_i = \frac{a_i - \bar{x}}{s} \implies \bar{z} = 0$ und $s_Z^2 = 1$	Aus rechnerischen Gründen bzw. wegen des Vergleichs zwischen verschiedenen Merkmalen werden Daten oft z-transformiert.
• Varianz bei klassierten Daten	$s^2 = \underbrace{\sum_{j=1}^{m} f_j \frac{w_j^2}{12}}_{s_{\text{int}}^2} + \underbrace{\sum_{j=1}^{m} f_j (\bar{x}_j - \bar{x})^2}_{s_{\text{ext}}^2}$	Dabei getroffene Annahme: Rechteckverteilung innerhalb einer Klasse. Falls $\bar{x}_j$ unbekannt ist, wird $\tilde{x}_j$ verwendet.
Variationskoeffizient V	$V = \frac{s}{\bar{x}},\ x_j \geq 0,\ j = 1, \ldots, m$ und $\bar{x} > 0$	Relatives Streuungsmaß (dimensionslos): Die Standardabweichung wird auf das arithmetische Mittel bezogen.

Aufgabe 8

Berechnen Sie für die 2. Aufgabe den Variationskoeffizienten und für die 3. Aufgabe den Quartilsabstand und die Standardabweichung.

3 Deskriptive Statistik: Bivariate Verteilungen

3.1 Darstellungsformen

Werden an einem Merkmalsträger i zwei Beobachtungswerte a_i und b_i der Merkmale X und Y festgestellt, so kann untersucht werden, ob ein rechnerischer Zusammenhang zwischen diesen Merkmalen besteht. In tabellarischer Form geschieht dies bei Häufungen von gleichen Beobachtungspaaren durch eine Häufigkeitstabelle (Assoziations-, Kontingenz-, Korrelationstabelle), sonst durch eine der Größe (eines Merkmals) nach geordnete Reihe der Beobachtungspaare (nicht bei nominalen Merkmalen möglich). Die Auswertung erfolgt im ersten Fall durch Spalten- bzw. Zeilenvergleich, im zweiten Fall (vor allem grafisch) durch Reihenfolgenvergleich.

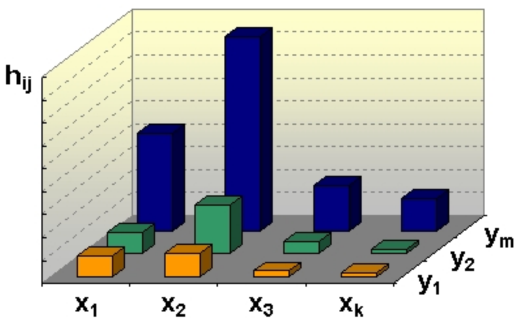
Häufigkeitsverteilung

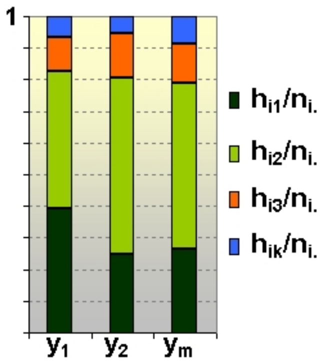
Zweidimensionale Häufigkeitstabelle Notation: x_j mit $j=1,\dots,k$ y_i mit $i=1,\dots,m$	Bedingte Verteilungen Zeilenvergleich (y_i festgehalten)	Bedingte Verteilungen Spaltenvergleich (x_j festgehalten)

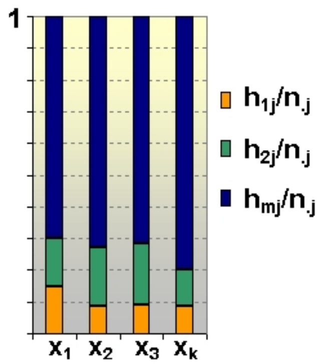
	x_1	$\dots$	x_j	$\dots$	x_k	Σ
y_1	h_{11}	$\dots$	h_{1j}	$\dots$	h_{1k}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	h_{i1}	$\dots$	h_{ij}	$\dots$	h_{ik}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_m	h_{m1}	$\dots$	h_{mj}	$\dots$	h_{mk}	$n_{m\bullet}$
Σ	$n_{\bullet 1}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet k}$	n

x_j	$\frac{h_{1j}}{n_{1\bullet}}$	$\frac{h_{ij}}{n_{i\bullet}}$	$\frac{h_{mj}}{n_{m\bullet}}$
x_1 x_2 $\vdots$ x_k			
	1	1	1

y_i	$\frac{h_{i1}}{n_{\bullet 1}}$	$\frac{h_{ij}}{n_{\bullet j}}$	$\frac{h_{ik}}{n_{\bullet k}}$
y_1 y_2 $\vdots$ y_m			
	1	1	1







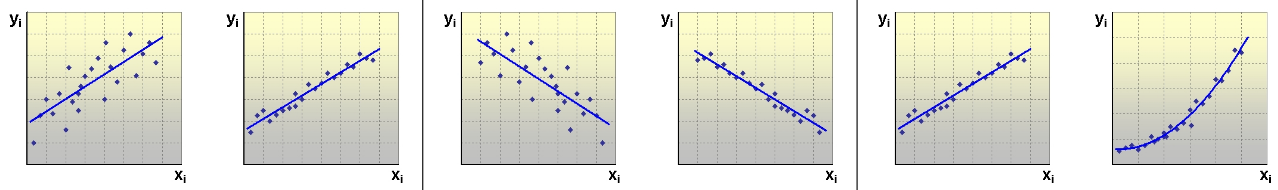
Statistische Unabhängigkeit

Besteht kein rechnerischer Zusammenhang zwischen den Merkmalen in der betrachteten Gesamtheit, so ergeben sich in den Spalten bzw. Zeilen dieselben relativen Häufigkeiten, wenn als Bezugsgröße jeweils die Spalten- bzw. Zeilensumme verwendet wird (bedingte Verteilung). Die absoluten Häufigkeiten in den Tabellenfeldern h^e_{ij} lassen sich dann als normiertes Produkt der Randhäufigkeiten errechnen:

$$h^e_{ij} = \frac{n_{\bullet j} \cdot n_{i\bullet}}{n}$$

Korrelation

Korrelationsrechnung bei ordinalen oder metrischen Merkmalen: Messung der Stärke und Richtung des rechnerischen Zusammenhangs zwischen Merkmalen, der einseitig ($x \longrightarrow y$), gegenseitig ($x \longleftrightarrow y$) oder über ein drittes Merkmal (oder einen Merkmalskomplex) ($z \longrightarrow (x,y)$) bewirkt sein kann. Die Korrelation ist an der Form der tabellarischen oder grafischen Anordnung erkennbar.



schwache	starke	schwache	starke	linearer	nicht linearer
positive Korrelation		negative Korrelation		statistischer Zusammenhang	

Es wird ab jetzt nicht mehr in den Symbolen zwischen Beobachtungswert und Merkmalsausprägung unterschieden, sondern sowohl die Beobachtungswerte als auch die Merkmalsausprägungen des Merkmals X werden mit x_i bzw. des Merkmals Y mit y_i bezeichnet. Bei $i = 1,\dots,n$ handelt es sich um Beobachtungswerte und bei $i = 1,\dots,m(k)$ um Merkmalsausprägungen.

Aufgabe 9

200 erwerbstätige Wähler werden nach der Stellung im Beruf (x_j mit x_1: Arbeiter, x_2: Angestellte/Beamte, x_3: Selbständige) und ihrer Wahlentscheidung bei den letzten Landtagswahlen (y_i mit y_1: CDU, y_2: SPD, y_3: FDP, y_4: Grüne) befragt. Man erhält folgendes Ergebnis:

	x_1	x_2	x_3
y_1	30	51	9
y_2	44	32	4
y_3	2	11	7
y_4	4	6	–

Berechnen Sie die Randverteilungen, die (sieben) bedingten Verteilungen sowie die absoluten Häufigkeiten der Assoziationstabelle bei statistischer Unabhängigkeit der betrachteten Merkmale in dieser Gesamtheit.
Wie hoch ist der Anteil

- der Angestellten/Beamten, die die SPD wählen?
- der Angestellten/Beamten unter den Wählern der SPD?
- der Wähler der SPD unter den Angestellten/Beamten?

Aufgabe 10

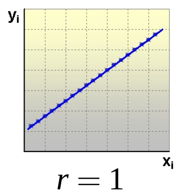
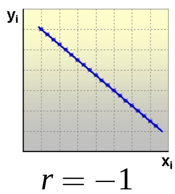
In einem Betrieb werden für die letzten zwölf Quartale die Zahl der Arbeitslosen im zugehörigen Arbeitsamtsbezirk (x in Hdrt.) und die Zahl der Krankmeldungen (y in Hdrt.) verglichen:

x_i	70	80	90	120	130	150	150	170	70	60	60	50
y_i	8	7	10	7	6	4	3	2	13	14	16	18

Zeichnen Sie ein Streuungsdiagramm. Interpretation?

3.2 Maßzahlen des rechnerischen Zusammenhangs

Kenngrößen bivariater Verteilungen, die die Stärke des rechnerischen Zusammenhangs zwischen den beiden Merkmalen in der untersuchten Gesamtheit abbilden, heißen Assoziations- oder Kontingenzmaße (wenn eines der Merkmale nominal skaliert ist) bzw. Korrelationskoeffizienten (wenn keines der Merkmale nominal skaliert ist).

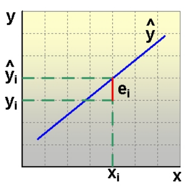
Bezeichnung	Symbol	Berechnung	Skalenniveau	Aussage
Chi-Quadrat-Koeffizient Pearson's Kontingenzkoeffizient Korrigierter Kontingenzkoeffizient	χ^2 C C^*	$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(h_{ij}-h_{ij}^e)^2}{h_{ij}^e}$ $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$ $C^* = \frac{C}{C_{\max}}$ mit $C_{\max} = \sqrt{\frac{\min(k,m)-1}{\min(k,m)}}$	beliebig	Es ist $\chi^2 > 0$, wenn ein Zusammenhang besteht. Eine Richtung des Zusammenhangs ist nicht interpretierbar. Viele Assoziationsmaße beruhen auf der Größe χ^2, die den Unterschied zwischen den tatsächlichen Häufigkeiten und den bei Unabhängigkeit geltenden Häufigkeiten abbildet.
Rangkorrelationskoeffizient von Spearman	R_{sp}	$R_{\text{sp}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ mit d_i: Differenz der Rangplätze der Beobachtungswerte x_i und y_i,	beide Merkmale mindestens ordinal	Je größer R_{sp} ist, desto stärker ist der Zusammenhang zwischen den Rangfolgen. Rangplätze werden allerdings als intervallskaliert angenommen. Es gilt: $-1 \leq R_{\text{sp}} \leq 1$ (bei eindeutigen Rängen).
Korrelationskoeffizient von Bravais-Pearson	r (ρ)	$r = \frac{s_{XY}}{s_X \cdot s_Y}$ mit der Kovarianz $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ $= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$	beide Merkmale metrisch	r misst die Stärke des linearen Zusammenhangs. Es gilt: $-1 \leq r \leq 1$.   $r = 1$ $r = -1$
Eta-Quadrat-Koeffizient	η^2	$\eta^2 = \frac{s_{\text{ext}}^2}{s^2} = 1 - \frac{s_{\text{int}}^2}{s^2}$	beeinflussendes Merkmal beliebig, beeinflusstes Merkmal metrisch	η^2 gibt an, welcher Anteil der Streuung durch die Gruppenzugehörigkeit erklärt werden kann. Es gilt: $0 \leq \eta^2 \leq 1$.

Aufgabe 11 Berechnen Sie für die Aufgaben 7, 9 und 10 sinnvolle Maßzahlen des rechnerischen Zusammenhangs.

1. Ergänzung: Messung von Zusammenhängen

<table>
<tr><th>Bezeichnung</th><th>Symbol</th><th>Berechnung</th><th>Skalenniveau</th><th>Aussage</th></tr>
<tr><td></td><td></td><td>Bei binären Merkmalen, d.h. 2×2-Tabellen:
$$\begin{array}{|c|c|c|}\hline & x_1 & x_2 \\ \hline y_1 & a & b \\ \hline y_2 & c & d \\ \hline\end{array} \qquad a = h_{11} \text{ etc.}$$</td><td rowspan="3">beliebig</td><td>Beide Maße beruhen auf dem Unterschiedsbetrag des Produkts der Diagonalenhäufigkeiten.</td></tr>
<tr><td>Prozentsatzdifferenz</td><td>$d\%$</td><td>$$d\% = \frac{|ad-bc|}{(a+c)(b+d)} \cdot 100 = \left|\frac{h_{11}}{n_{.1}} - \frac{h_{12}}{n_{.2}}\right| \cdot 100$$</td><td>$0 \le d\% \le 100$.</td></tr>
<tr><td>Phi-Koeffizient</td><td>Φ</td><td>$$\Phi = \sqrt{\frac{\chi^2}{n}} = \frac{|ad-bc|}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$</td><td>$0 \le \Phi \le 1$.</td></tr>
<tr><td>Kendalls Tau-b

Somers' d</td><td>τ_b

d_y</td><td>$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c+n_d+T_x)(n_c+n_d+T_y)}}$$ bei quadratischen Tabellen.
$$d_y = \frac{n_c - n_d}{n_c + n_d + T_y}$$ $$d_y = \frac{ad-bc}{(a+c)(b+d)} \quad \text{bei } 2 \times 2\text{-Tabellen.}$$ n_c: Zahl der konkordanten Paare
n_d: Zahl der diskordanten Paare
n_c, n_d: eindeutige Paarreihungen
T_x, T_y, (T_{xy}): „Ties“: Zahl der Paare, die sich nicht bzgl. beider Merkmale unterscheiden</td><td>beide Merkmale mindestens ordinal</td><td>Die Maße beruhen auf Paarvergleichen. Bei n Einheiten gibt es $\frac{n(n-1)}{2}$ mögliche Paare. n_c ist z.B. die Anzahl der Paare, bei der eine Einheit bzgl. beider Merkmale einen höheren Rang hat als die Partnereinheit.
$-1 \le \tau_b, d_y \le 1$.</td></tr>
<tr><td colspan="3">Regressionsrechnung</td><td rowspan="5">beide Merkmale metrisch</td><td rowspan="4">Abbildung des rechnerischen (linearen) Einflusses einer erklärenden Variablen x auf eine Zielvariable y für einen bestimmten Datensatz.
a: Schätzwert für y, wenn $x = 0$ ist.
b: Schätzwert für die Zunahme von y, wenn x um eine Einheit zunimmt.</td></tr>
<tr><td>Regressionsfunktion</td><td>$\hat{y} = f(x)$</td><td>$\hat{y} = a + b \cdot x$</td></tr>
<tr><td>Methode der Kleinsten Quadrate, Fehler e_i</td><td>e_i</td><td>

$$\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 \stackrel{!}{=} \min$$</td></tr>
<tr><td>Regressionskoeffizienten</td><td>a

b</td><td>$a = \bar{y} - b\bar{x}$
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{s_{XY}}{s_X^2}$$</td></tr>
<tr><td>Bestimmtheitsmaß</td><td>r^2</td><td>$$r^2 = \left(\frac{s_{XY}}{s_X\, s_Y}\right)^2 = \frac{s_{\hat{Y}}^2}{s_Y^2}, \qquad 0 \le r^2 \le 1$$</td><td>Anteil der durch die Regressionsgerade „erklärten“ Varianz der Zielvariablen.</td></tr>
</table>

2. Ergänzung: PRE-Maße (Proportional Reduction in Error)

PRE-Maße sollen eine Interpretation der Stärke des Einflusses der unabhängigen auf die abhängige Variable erlauben.

$$PRE = \frac{E_1 - E_2}{E_1} \quad \text{„proportionale Abnahme des Vorhersagefehlers“}$$

E_1: *„Fehler“ bzgl. der Vorhersage der abhängigen Variablen Y aufgrund ihrer Verteilung.*

E_2: *„Fehler“ bzgl. der Vorhersage der abhängigen Variablen Y bei Kenntnis des Einflusses der unabhängigen Variablen X.*

Die PRE-Maße unterscheiden sich je nach „Fehler“-Definition und verwendetem Vorhersagewert.

Bezeichnung	Symbol	Berechnung	Skalenniveau	Aussage
Goodmans und Kruskals Lambda	λ_y	$\lambda_y = \frac{\sum_j \max_i h_{ij} - \max_i n_{i\bullet}}{n - \max_i n_{i\bullet}}$ $\lambda_y = \frac{E_1 - E_2}{E_1}$ mit $E_1 = n - \max_i n_{i\bullet}$ $E_2 = \sum_j (n_{\bullet j} - \max_i h_{ij})$	beliebig	Man würde den häufigsten Wert vorhersagen, also ist E_1 die Zahl der falschen Voraussagen. Entsprechend E_2: Man würde die häufigsten Werte der bedingten Verteilungen voraussagen, also ist E_2 die Anzahl der falschen Voraussagen. Es gilt: $0 \le \lambda_y \le 1$.
Goodmans und Kruskals Gamma	γ	$\gamma = \frac{n_c - n_d}{n_c + n_d}$ (bei wenig Ties) $\gamma = \frac{E_1 - E_2}{E_1}$ mit $E_1 = 0.5(n_c + n_d)$ $E_2 = \min(n_c, n_d)$ für $n_c < n_d$: $\gamma < 0$ für $n_c > n_d$: $\gamma > 0$	beide Merkmale mindestens ordinal	Wenn man „nichts“ weiß außer der Zahl Paare mit eindeutiger Reihenfolge, würde man E_1 tippen. (Prinzip des unzureichenden Grundes) γ ist größer null, wenn die Zahl der konkordanten Paare überwiegt und γ ist kleiner null, wenn die Zahl der diskordanten Paare überwiegt. Es gilt: $-1 \le \gamma \le 1$.
Bestimmtheitsmaß	r^2	$r^2 = \frac{s^2_{\hat{Y}}}{s^2_Y} = 1 - \frac{s^2_e}{s^2_Y}$ $r^2 = \frac{E_1 - E_2}{E_1}$ mit $E_1 = s^2_Y,\quad E_2 = s^2_e$	beide Merkmale metrisch	E_1 ist der als Varianz berechnete Prognosefehler, wenn man $\bar{y}$ als Vorhersagewert für jedes y_i verwenden würde. E_2 ist der Prognosefehler, wenn man $\hat{y}_i$ als Vorhersagewert verwendet. Es gilt: $0 \le r^2 \le 1$.
Eta-Quadrat-Koeffizient	η^2	$\eta^2 = \frac{s^2_{\text{ext}}}{s^2} = 1 - \frac{s^2_{\text{int}}}{s^2}$ $\eta^2 = \frac{E_1 - E_2}{E_1}$ mit $E_1 = s^2,\quad E_2 = s^2_{\text{int}}$	unabh. Merkmal beliebig, abh. Merkmal metrisch	E_1 ist der als Varianz berechnete Prognosefehler, wenn man $\bar{y}$ als Vorhersagewert für jedes y_{ij} verwenden würde. E_2 ist der Prognosefehler, wenn man bei $j = 1, \ldots, m$ Untergruppen $\bar{y}_j$ als Vorhersagewert verwendet. Es gilt: $0 \le \eta^2 \le 1$.

4 Aufgaben zur Wiederholung

Aufgabe 12

Aus einer Erhebung bei 2 000 Erwerbstätigen einer Region erhält man folgende Verteilung der Ausgaben für den öffentlichen Nahverkehr:

Ausgaben von … bis unter … €	Erwerbstätige	Ausgabensumme je Klasse (T€)
0 – 10	300	0,9
10 – 20	400	4,8
20 – 25	400	8,8
25 – 30	300	8,1
30 – 40	300	10,2
40 – 50	200	9,2
50 – 100	100	8,0

a) Zeichnen Sie ein Histogramm, die Verteilungsfunktion und bestimmen Sie die Quartile.

b) Berechnen Sie das arithmetische Mittel, die Varianz (aus externer und interner) und den Variationskoeffizienten.
Kritisieren Sie für dieses Beispiel die Annahme einer Rechteckverteilung bei den grafischen Darstellungen und bei der Berechnung der internen Varianz.

Lösung: a) $Q_1 = 15,\ Q_2 = 23,75,\ Q_3 = 33$, b) $\bar{x} = 25,\ s^2 = 332,\ V = 0,73$

Aufgabe 13

Drei zufällig ausgewählte Gruppen A, B und C von Autofahrern wurden mit unterschiedlichen Konzepten zur Nutzung des öffentlichen Nahverkehrs beim Stadtbesuch animiert. Für den letzten Monat erhielt man folgende Ergebnisse:

Nutzung des Angebots … mal	Gruppe A	Gruppe B	Gruppe C
0	50	30	20
1	60	40	30
2	40	80	50
3	20	80	100
4	20	40	60
5	10	30	40

a) Berechnen Sie λ_y und interpretieren Sie das Ergebnis als PRE-Maß.

b) Berechnen Sie den korrigierten C-Koeffizienten.

c) Berechnen Sie den η^2-Koeffizienten und interpretieren Sie ihn als PRE-Maß.

d) Vergleichen Sie die Aussagen des C-Koeffizienten, des λ-Koeffizienten und des η^2-Koeffizienten.

Lösung: a) $\lambda_y = 0,067$, b) $C^* = 0,44$, c) $\eta^2 = 0,10$

Aufgabe 14

In der Caféteria einer Universität wurde ein neues Cola-Getränk eingeführt, das am ersten Tag kostenlos an Studenten verteilt wurde. Von den Probanden wurden je 200 Studentinnen und Studenten gebeten, in den folgenden zwei Wochen zu notieren, wie oft sie das Getränk wieder kauften. Man erhielt folgendes Ergebnis:

Käufe	weiblich	männlich
0	70	15
1	50	20
2	35	35
3	20	60
4	10	35
5	10	20
6	5	15

a) Zeichnen Sie die empirischen Verteilungsfunktionen für beide Gruppen.

b) Berechnen Sie Modus, Median und arithmetisches Mittel sowie die Varianz für jede Verteilung.

c) Wie hoch ist die Varianz der aggregierten Verteilung?

Lösung: b) $D_\mathrm{m} = 3,\ D_\mathrm{w} = 0, Z_\mathrm{m} = 3,\ Z_\mathrm{w} = 1,\ \bar{x}_\mathrm{m} = 3,\ \bar{x}_\mathrm{w} = 1,5, s^2_\mathrm{m} = 2,5,\ s^2_\mathrm{w} = 2,55,$
c) $s^2 = 3,0875$

Aufgabe 15

500 Studierende wurden nach ihrer Meinung zur beabsichtigten stärker leistungsorientierten Bezahlung der Professoren (-1: Unsinn, 0: neutral, +1: unbedingt) und einer regelmäßigen Leistungsmessung durch Befragung von Vorlesungsbesuchern (-1: Unsinn, 0: neutral, +1: unbedingt) befragt:

Bezahlung	Befragung		
	-1	0	+1
-1	80	20	40
0	10	50	80
+1	30	10	180

a) Berechnen und zeichnen Sie die bedingten Verteilungen (nur Spalten) und verbalisieren Sie das Ergebnis.

b) Berechnen Sie Kendall's τ_b. Entspricht das Ergebnis Ihrer Interpretation der bedingten Verteilungen?

Lösung: b) $n_c = 41\,700,\ n_d = 9\,500,\ T_x = 30\,000,\ T_y = 18\,400,\ T_{xy} = 25\,150,\ \tau_b = 0,428$

Aufgabe 16

Bei neun Sportstudenten wird vor der Durchführung eines Trainingsprogramms eine antropometrische Messung vorgenommen:

Student Nr. i	1	2	3	4	5	6	7	8	9
y: Gewicht (kg)	63	78,2	85,2	78	79,5	69,5	75,6	78	68
x: Größe (cm):	170	178	190	182	186	174	184	181	175

a) Zeichnen Sie ein Streuungsdiagramm.

b) Berechnen Sie eine lineare Regressionsfunktion nach der Methode der kleinsten Quadrate und zeichnen Sie das Ergebnis in das Diagramm.
Interpretieren Sie den Koeffizienten b.

c) Berechnen und vergleichen Sie die Aussagen des Korrelationskoeffizienten nach Bravais-Pearson und des Bestimmtheitsmaßes. Wodurch könnte das Bestimmtheitsmaß erhöht werden? Interpretieren Sie das Bestimmtheitsmaß als PRE-Maß.

Lösung: b) $y = -105 + x$, c) $r^2 = 0,856$

5 Induktive Statistik: Einführung

5.1 Wahrscheinlichkeitsrechnung

Bisher wurden Methoden zur zahlenmäßigen Beschreibung genau abgegrenzter statistischer Massen vorgestellt. Ziel statistischer Untersuchungen ist jedoch meist, allgemeingültigere Ergebnisse zu erhalten. Werden solche Daten als Ergebnisse von Zufallsexperimenten – z.B. Befragungsergebnisse aus einer Zufallsstichprobe von Personen – gewonnen, so ist zwar der Grad der Allgemeingültigkeit des Ergebnisses (der Induktionsschluss) unsicher, er kann aber mit Hilfe der Wahrscheinlichkeitsrechnung quantifiziert werden.

Regeln der Wahrscheinlichkeitsrechnung (am Beispiel der Aufgabe 9, Seite 13)

1. Eigenschaften des Wahrscheinlichkeitsmaßes (Axiome von Kolmogoroff)		
$P(A) \geq 0$ $P(I) = 1$ $P(A \cup B) =$ $P(A) + P(B)$	$P(\text{SPD}) = 0,4$ $P(\text{FDP}) = 0,1$ $P(\text{SPD} \cup \text{FDP}) =$ $0,4 + 0,1 = 0,5$	Die Wahrscheinlichkeit P für ein Ereignis A (Zusammenfassung möglicher Ergebnisse eines Zufallsexperiments) ist nie negativ. Die Wahrscheinlichkeit für das sichere Ereignis I ist 1. Die Wahrscheinlichkeiten für 2 sich ausschließende Ereignisse können addiert werden.
2. Additionssatz (Verknüpfung $\cup$: „entweder-oder“, Vereinigung)		
$P(A \cup B) =$ $P(A) + P(B) - P(A \cap B)$	$P(\text{SPD} \cup \text{Arbeiter})$ $= 0,4 + 0,4 - 0,22$ $= 0,58$	Schließen sich zwei Ereignisse nicht aus, so muss von der Summe der Wahrscheinlichkeiten für die Einzelergebnisse die Wahrscheinlichkeit der Schnittmenge abgezogen werden.
3. Multiplikationssatz (Verknüpfung $\cap$: „sowohl-als-auch“, Schnitt)		
$P(A \cap B) = P(A) \cdot P(B\|A)$ $= P(B) \cdot P(A\|B)$ $P(A\|B) = \dfrac{P(A \cap B)}{P(B)}$	$P(\text{SPD} \cap \text{Selbstg.})$ $= 0,4 \cdot 0,05 = 0,02$ $= 0,1 \cdot 0,2 = 0,02$	Bei (stochastischer) Unabhängigkeit zweier Ereignisse gilt: $P(A \cap B) = P(A) \cdot P(B)$ $P(A\|B) = P(A)$

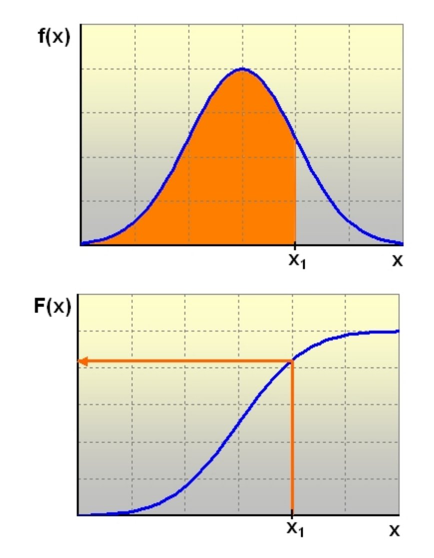
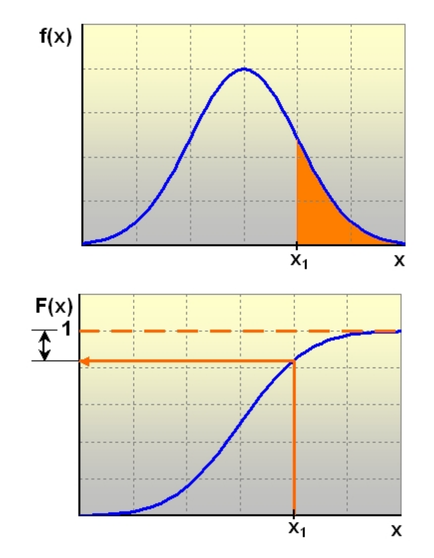
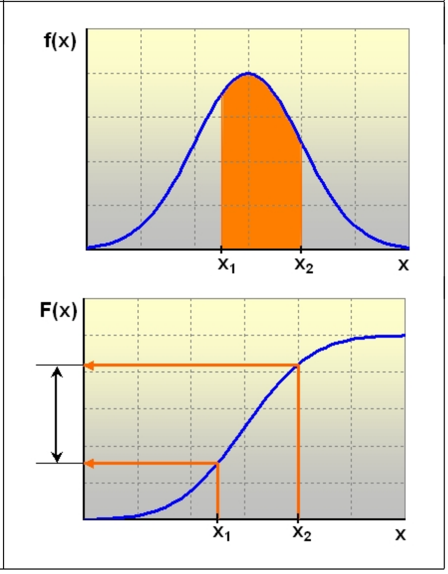
Praktische Berechnung von Wahrscheinlichkeiten

- Bei einfachen Zufallsexperimenten, deren Ergebnisse (Elementarereignisse) gleichwahrscheinlich sind, lassen sich Wahrscheinlichkeiten aus dem Verhältnis von „günstigen“ zu „möglichen“ Fällen berechnen (Glücksspiele, Urnenmodelle). Die diesem Wahrscheinlichkeitsmaß zugrundeliegende Auffassung wird auch **klassischer** Wahrscheinlichkeitsbegriff genannt.
- In den Wirtschafts- und Sozialwissenschaften wird beim „Schätzen“ und „Testen“ (vgl. Abschnitt 5.3) zumeist vom **statistischen** oder **frequentistischen** Wahrscheinlichkeitsbegriff ausgegangen: Wahrscheinlichkeit ist eine relative Häufigkeit, die in einer sehr langen Reihe unabhängiger Versuche festgestellt wurde. Der allgemeine Ursachenkomplex für die Häufigkeitsverteilung muss allerdings konstant bleiben. Beispielsweise könnte man so eine Verteilung von möglichen Ergebnissen einer Stichprobenziehung errechnen und aus dieser Verteilung dann Wahrscheinlichkeiten für ganz bestimmte Ergebnisse entnehmen.
- Insbesondere bei ökonomischen Anwendungen z.B. bei Risikoabschätzungen in Entscheidungssituationen spielt der induktive, speziell der **subjektive** Wahrscheinlichkeitsbegriff eine Rolle. Die Wahr-

scheinlichkeit wird als ein Maß für den Grad der Überzeugtheit von der Richtigkeit einer Aussage aufgefasst. Vielfach wird die Meinung vertreten, dass in praktischen Anwendungen jede Wahrscheinlichkeitsaussage subjektive Elemente enthalte.

Wahrscheinlichkeitsverteilungen

Drückt man die möglichen Ergebnisse als Zufallsvariable X aus, d.h. als eine Abbildung, die jedem Ergebnis aus der Ergebnismenge eine reelle Zahl zuordnet, so könnte man in allen drei genannten Fällen eine Verteilung von Wahrscheinlichkeiten auf die Zufallsvariable X als Funktionsgleichung erstellen. Die Funktion $F(x)$, die jedem $x \in \mathbb{R}$ die Wahrscheinlichkeit $P(X \leq x)$ zuordnet, also $F(x) = P(X \leq x)$, heißt Verteilungsfunktion von X. Die Wahrscheinlichkeiten für mögliche Realisationen x kann man dann an der Verteilungsfunktion $F(x)$ ablesen. Für die praktische Anwendung üblich sind häufig verwendete Wahrscheinlichkeits- bzw. Verteilungsfunktionen, die schon tabellarisch (in „Tafeln“) ausgewertet sind.

$P(X \leq x_1) = F(x_1)$	$P(X > x_1) = 1 - F(x_1)$	$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$
		

- In der Praxis wird nämlich zur Bestimmung von Wahrscheinlichkeiten oft so vorgegangen, dass je nach Art der Zufallsvariablen und des die Wahrscheinlichkeit erzeugenden Zufallsprozesses aus vorliegenden „theoretischen“ Verteilungen, das sind in mathematische Modelle – hier Funktionsgleichungen – abgebildete, theoretische Zufallsprozesse, eine „passende“ ausgewählt wird. Eine so zustandekommende Wahrscheinlichkeitsaussage ist dann natürlich selbst mit einer gewissen Unsicherheit (nämlich die der richtigen Modellauswahl) behaftet, ohne dass diese Unsicherheit quantifiziert werden könnte.
- Für derartige Verteilungen lassen sich normalerweise Kenngrößen wie in der deskriptiven Statistik (Erwartungswert, Varianz) berechnen. Günstig ist es, wenn diese Kenngrößen auch eine Funktion der Parameter der Verteilung sind. Beispielsweise sind bei der Gauß’schen Normalverteilung die Kenngrößen μ und σ^2 selbst Parameter der Verteilung (vgl. Abschnitt 5.2).

Aufgabe 17

a) Berechnen Sie die Wahrscheinlichkeitsverteilung für das Ereignis „Zahl der Arbeiter“ in einer Stichprobe von 3 Personen aus den 200 der Aufgabe 9, Seite 13.

b) Angenommen, wir ziehen aus der Einkommensverteilung von Aufgabe 3, Seite 6, eine Stichprobe vom Umfang $n = 1$. Wie groß ist die Wahrscheinlichkeit, jemanden zu ziehen, dessen Einkommen weniger als 1000, 2000 und mehr, zwischen 1250 und unter 3000 beträgt?

5.2 Normalverteilung, Stichprobenverteilung

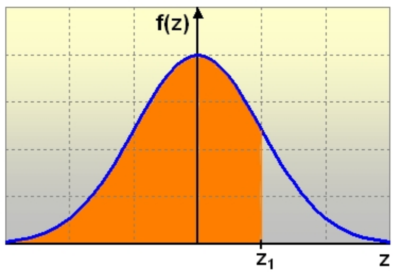
Die am häufigsten eingesetzte theoretische Verteilung ist die Gauß'sche Normalverteilung. Die Zufallsvariable kann hier als Summe „sehr vieler" voneinander unabhängigen Einflussvariablen interpretiert werden, also z.B. als arithmetisches Mittel bei der Ziehung von einfachen, unabhängigen Zufallsstichproben. Die Normalverteilung ist dann die Verteilung aller möglichen Ziehungsergebnisse.

Die Parameter der Normalverteilung sind die (auch deshalb schon in der deskriptiven Statistik häufig verwendeten) Größen μ und σ^2. Für $X \sim N(\mu, \sigma^2)$ gilt:

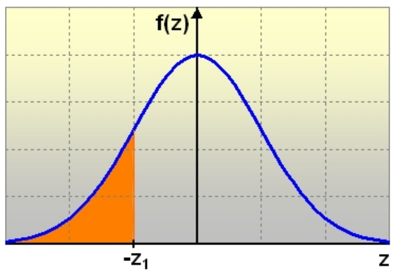
$$P(X \le x) = F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du.$$

In der Praxis bestimmt man diese Wahrscheinlichkeit bei bekannten μ und σ so, dass man die Differenz $x - \mu$ als Vielfaches z von σ ausdrückt, also $x = \mu + z \cdot \sigma$ bzw. $z = \frac{x-\mu}{\sigma}$ berechnet. Die zu z gehörende Wahrscheinlichkeit kann in Tafeln zur Standardnormalverteilung abgelesen werden.

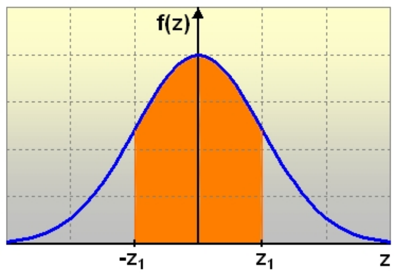
$P(Z \le z_1) = F(z_1)$



$P(Z \le -z_1) = F(-z_1) = 1 - F(z_1)$



$P(-z_1 \le Z \le z_1) = 2F(z_1) - 1$



z	0,00	0,25	0,50	0,75	1,00	1,25	1,50	1,75	2,00	2,50	3,00
F(z)	0,500	0,5987	0,6915	0,7734	0,8413	0,8944	0,9332	0,9599	0,9772	0,9938	0,9987

Bei der Ziehung unabhängiger Zufallsstichproben vom Umfang n aus einer **beliebigen** Grundgesamtheit mit arithmetischem Mittel μ und Standardabweichung σ gilt für die Verteilung aller möglichen arithmetischen Mittel:

- Der Erwartungswert („Durchschnitt") aller möglichen Stichprobenergebnisse für das arithmetische Mittel ist das arithmetische Mittel der Grundgesamtheit, d.h. $E(\overline{X}) = \mu$.
- Die Streuung aller möglichen Durchschnitte hängt von der Streuung in der Grundgesamtheit und dem Stichprobenumfang ab, d.h. $E(\overline{X} - \mu)^2 = Var(\overline{X}) = \sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}$ (bzw. $\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$ ohne Zurücklegen; für N gegenüber n genügend groß kann der Korrekturfaktor $(N-n)/(N-1)$ vernachlässigt werden).
- Bei „großen" (Praxis: $n > 100$) Stichprobenumfängen kann die Verteilung der Stichprobenergebnisse durch eine Normalverteilung mit den Parametern μ und $\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}$ approximiert werden (zentraler Grenzwertsatz, vgl. Beispiel mit excel2000 www.prof-roessler.de/Dateien/Statistik/zgs.xls).

Aufgabe 18

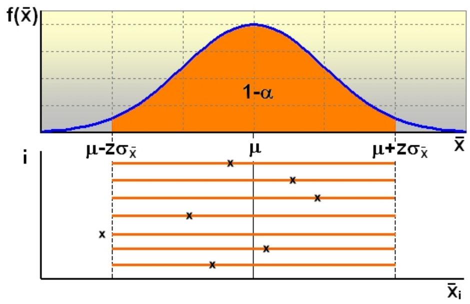
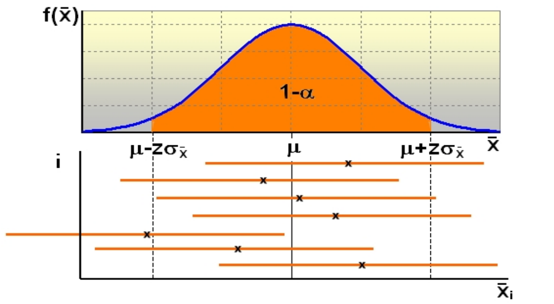
Angenommen, die Körpergröße von Männern in Deutschland sei normalverteilt mit $\mu = 178$cm und $\sigma = 10$cm.

a) Wie groß ist die Wahrscheinlichkeit bei zufälliger Auswahl eines Mannes, eine Körpergröße aa) $x \le 193$cm ab) $x > 168$cm ac) 158cm$< x \le 198$cm zu erhalten?

b) Angenommen, man ziehe eine Stichprobe mit Zurücklegen vom Umfang $n = 100$ (1000). Wie groß ist die Wahrscheinlichkeit, als arithmetisches Mittel einen Wert ba) $\bar{x} > 177$cm bb) $\bar{x} \le 180$cm bc) $175 < \bar{x} \le 181$ zu erhalten?

5.3 Grundlagen des Schätzens und Testens

Ist die Verteilung möglicher Stichprobenergebnisse bekannt – also z.B. eine bestimmte theoretische Verteilung oder eine durch Simulationsstudien näherungsweise abgeleitete Verteilung – so können schon vor einer speziellen Stichprobenziehung Wahrscheinlichkeitsaussagen zu erwarteten Ergebnissen getroffen (Inklusionsschluss) oder ein notwendiger Stichprobenumfang, der eine „Mindestgenauigkeit“ gewährleistet, bestimmt werden. Auch könnten von einem gegebenen Stichprobenergebnis aus quantifizierte Mutmaßungen über den „wahren“ Wert in der Grundgesamtheit angestellt werden (Repräsentationsschluss). Ist die Stichprobenverteilung die Normalverteilung $N(\mu,\sigma_{\overline{x}}^2)$, so lässt sich die Vorgehensweise für z.B. symmetrische Intervalle wie folgt veranschaulichen.

Inklusionsschluss	Repräsentationsschluss
$P(\mu - z\cdot\sigma_{\overline{x}} \le \overline{X} \le \mu + z\cdot\sigma_{\overline{x}}) = 1-\alpha$	$P(\overline{X} - z\cdot\sigma_{\overline{x}} \le \mu \le \overline{X} + z\cdot\sigma_{\overline{x}}) = 1-\alpha$
	
ca. 95% der $\overline{x}_i$ liegen im Bereich $\pm 2\cdot\sigma_{\overline{x}}$ um μ	ca. 95% der Intervalle der Länge $\pm 2\cdot\sigma_{\overline{x}}$ um die $\overline{x}_i$ überdecken den Wert μ – „Vertrauensintervalle“

Die Größe $|e| = z\cdot\sigma_{\overline{x}}$ ist der sog. Stichprobenfehler. Sind e, z und σ gegeben, so kann ein „notwendiger“ Stichprobenumfang berechnet werden: $n \ge z^2\cdot\frac{\sigma^2}{e^2}$.

- Beim **Repräsentationsschluss** wird bei vorgegebenem z und $\sigma_{\overline{x}}$ ein Intervall berechnet, das mit einer Wahrscheinlichkeit von $(1-\alpha)$ den unbekannten Wert μ überdeckt. σ ist jedoch meist unbekannt und wird dann aus der Stichprobe geschätzt: $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum(x_i-\overline{x})^2$ (weil $E(s^2) = \sigma^2$, d.h. s^2 erwartungstreue Schätzfunktion für σ^2. (N groß: kein Korrekturfaktor, n groß: $Z = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim$ N(0,1).))
- Beim **Hypothesentest** wird überprüft, ob ein bestimmtes Stichprobenergebnis zu den (nach dem **Inklusionsschluss**) wahrscheinlichen Ergebnissen gehört. Wenn nicht, gilt die Hypothese als widerlegt.
- Beim **Rückschluss** von einem bestimmten „repräsentativen“ Stichprobenergebnis auf die unbekannte Grundgesamtheit – die übliche Anwendung in der Markt- und Meinungsforschung – wird die Güte des Ergebnisses durch die Angabe eines **Vertrauensintervalls** (Repräsentationsschluss), des Stichprobenfehlers oder wenigstens des Stichprobenumfangs dokumentiert.
- Ist X eine **0,1-Variable** und p (bzw. π) der Anteil der 1-Träger in der Stichprobe (Grundgesamtheit), so ist $\overline{x} = p$ (bzw. $\mu = \pi$) und $s^2 = \frac{n}{n-1}p(1-p)$ (bzw. $\sigma^2 = \pi(1-\pi)$).

Aufgabe 19

a) Es wird behauptet, deutsche Männer seien im Durchschnitt 178cm groß bei einer Standardabweichung von 10cm. Wir überprüfen die Behauptung durch Zufallsstichproben vom Umfang $n = 100$ (1000) und erhalten jeweils $\overline{x} = 179$. Ist die Behauptung bei einer Wahrscheinlichkeit von $(1-\alpha) = 0{,}9545$ haltbar (also bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}0455$ widerlegbar)?

b) Durch eine einfache Zufallsstichprobe von 900 Haushalten aus den ca 39 Mio. Haushalten in Deutschland sollen die Durchschnittsausgaben für Nachrichtenübermittlung erfasst werden. Wir erhalten $\sum_{i=1}^{900} x_i = 45\,000$ und $\sum_{i=1}^{900} x_i^2 = 9\,531\,900$. Wie „genau“ ist das Ergebnis?

6 Induktive Statistik: Anwendungen

6.1 Zufallsstichproben

Der Repräsentationsschluss ist ein Rückschluss vom eingetroffenen Stichprobenergebnis auf den unbekannten, aber festen Parameter in der Grundgesamtheit. Da nach der Realisation keine Wahrscheinlichkeitsaussagen mehr möglich sind, spricht man in frequentistischer Betrachtungsweise von einer Konfidenzaussage: Die bzgl. des Stichprobenfehlers getroffene Aussage (das Intervall) wäre bei einer großen Zahl unabhängiger Stichprobenziehungen in z.B. 95,45% (Konfidenzniveau) der Fälle richtig. Als interessierende Ergebnisse aus Zufallsstichproben werden hier arithmetische Mittel bzw. Merkmalssummen betrachtet. Bei gegebenem Konfidenzniveau – also gegebenem z, sofern die Gauß'sche Normalverteilung als Stichprobenverteilung verwendet werden darf, – hängt der Stichprobenfehler von der Streuung der möglichen Stichprobenergebnisse, also hier von der Standardabweichung $\sigma_{\bar{x}}$ ab, die in der Praxis geschätzt werden muss.

Einfache Zufallsstichproben

Bei einfachen Zufallsstichproben (simple random sampling) hat vor der ersten zufälligen Auswahl jede Einheit in der Grundgesamtheit dieselbe Auswahlwahrscheinlichkeit. Es kann mit (m.Z.) oder ohne (o.Z.) Zurücklegen gezogen werden.

Vorgehensweise	m.Z.	o.Z.
1. Genauigkeitsvorgabe, d.h. gwünschte Genauigkeit entweder absolut (e') oder relativ ($e'_r = \frac{e'}{\mu'}$) bei vermutetem μ'		
2. Abschätzung der Varianz (aus anderen, z.B. früheren Erhebungen, Pilotstudien, „Annahmen“ bzw. der Stichprobenrealisation selbst)	σ'^2	σ'^2
3. Bestimmung des notwendigen Stichprobenumfangs n $\left[\frac{N-n}{N-1} \approx 1 - \frac{n}{N} \quad \text{mit} \quad \frac{n}{N}: \text{„Auswahlsatz“}\right]$	$n \geq z^2 \frac{\sigma'^2}{e'^2}$ $n \geq z^2 \frac{V'^2}{e_r'^2}$	$n \geq N\left(1 + \frac{Ne'^2}{z^2\sigma'^2}\right)^{-1}$ $n \geq N\left(1 + \frac{Ne_r'^2}{z^2V'^2}\right)^{-1}$
4. Zufallsauswahl (vollständige Auswahlliste!) und Erhebung x_i		
5. Hochrechnung	$\hat{\mu} = \bar{x}$	$\hat{\mu} = \bar{x}$
6. Fehlerrechnung $\hat{e}$	$\lvert\hat{e}\rvert = z\frac{s}{\sqrt{n}}$	$\lvert\hat{e}\rvert = z\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$
7. Konfidenzintervalle	$\bar{x} - \hat{e} \leq \mu \leq \bar{x} + \hat{e}$ $N \cdot \bar{x} - N \cdot \hat{e} \leq N \cdot \mu \leq N \cdot \bar{x} + N \cdot \hat{e}$	

Geschichtete Zufallsstichproben

Um die Streuung der möglichen Ergebnisse zu verringern, versucht man in der Praxis durch Nutzung von Zusatzinformationen die Gesamtheit in – bezüglich der Varianz des zu erhebenden (bzw. eines mit ihm hoch korrelierten) Merkmals – homogene Untergruppen zu schichten (stratified sampling). Wir gehen davon aus, dass die Zahl der Schichten und die Schichtgrenzen schon festgelegt sind, die Gesamtstichprobe n proportional zu den Schichtumfängen N_h der L Schichten ($h = 1, \ldots, L$) aufgeteilt wird und die Stichproben je Schicht n_h m.Z. ausgewählt werden. ($\sum n_h = n$, $\sum N_h = N$)

Vorgehensweise	
1. Genauigkeitsvorgabe e'	
2. Abschätzung der Varianzen	$\sigma_h'^2$
3. Notwendiger Stichprobenumfang n	$n \geq z^2 \dfrac{\sum N_h \sigma_h'^2}{N \cdot e'^2}$
4. Proportionale Aufteilung	$n_h = n \dfrac{N_h}{N}$
5. Zufallsauswahl m.Z. je Schicht und Berechnung $\bar{x}_h$	
6. Hochrechnung	$\hat{\mu} = \bar{x} = \dfrac{1}{N} \sum N_h \bar{x}_h$
7. Fehlerrechnung $\hat{e}$	$\lvert\hat{e}\rvert = z \dfrac{1}{N} \sqrt{\sum N_h^2 \dfrac{s_h^2}{n_h}} = z \sqrt{\dfrac{1}{n} \sum \dfrac{N_h}{N} s_h^2}$
8. Konfidenzintervalle	$\bar{x} - \hat{e} \leq \mu \leq \bar{x} + \hat{e}$ $N \cdot \bar{x} - N \cdot \hat{e} \leq N \cdot \mu \leq N \cdot \bar{x} + N \cdot \hat{e}$

Aufgabe 20

Aus einer früheren Erhebung zu den monatlichen Ausgaben für ein Kind hat man für eine Grundgesamtheit von Haushalten mit Kindergeldansprüchen folgende Daten:

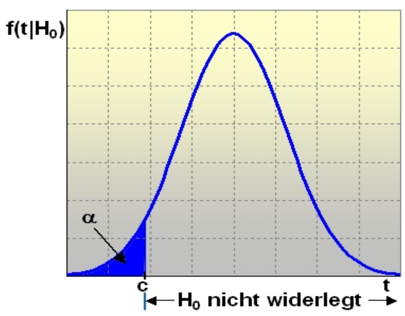
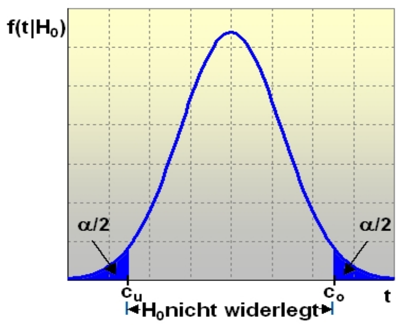
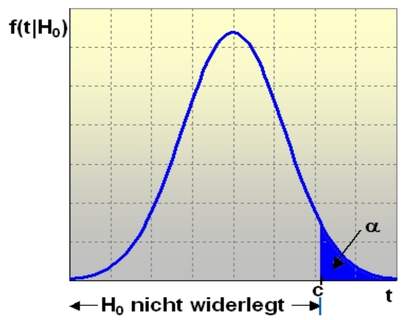
Schicht Nr.	Anzahl der Haushalte (Mio)	Gesamtausgaben je Schicht (Mio €)	Summe der quadrierten Einzelausgaben je Schicht (Mio €)
1	5	750	125 000
2	3	900	280 800
3	2	1 000	512 800

Man berechne für eine geplante neue Erhebung der Durchschnittsausgaben den notwendigen Stichprobenumfang bei uneingeschränkter und bei geschichteter Zufallsauswahl (Aussagewahrscheinlichkeit 95,45%, zulässiger absoluter Zufallsfehler 5,- €).

6.2 Hypothesenprüfung

Die sog. **Nullhypothese** *(H_0) ist die mathematische Formulierung einer aus der Theorie oder Erfahrung oder Güteforderung etc. sich ergebenden Hypothese so, dass eine Überprüfung durch einen statistischen Test möglich ist. Dazu gehören eine adäquate empirische Messung und deren Umsetzung in eine statistische Kenngröße (***Testfunktion** *T als Zufallsvariable) so, dass bei bekanntem Zufallsprozess eine Verteilung möglicher Ergebnisse angegeben werden kann. So lassen sich Regeln ableiten, die mögliche Stichprobenergebnisse als mit einer Hypothese verträglich oder nicht verträglich einzuordnen erlauben.*

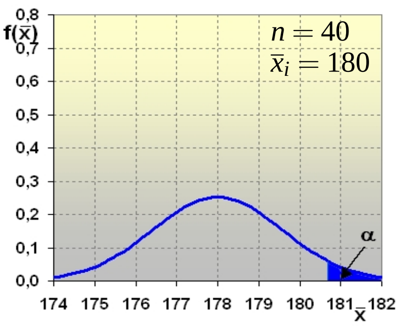
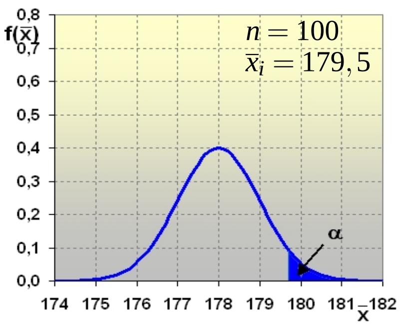
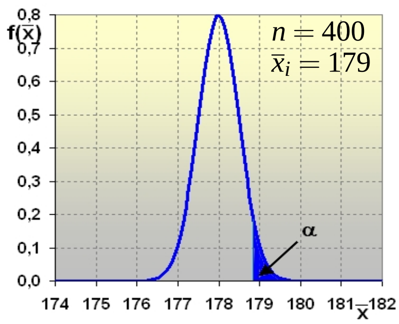
Signifikanztest

$H_0: t \geq t_0$ Bereichshypothese	$H_0: t = t_0$ Punkthypothese	$H_0: t \leq t_0$ Bereichshypothese
		

Zur Entscheidung, ob eine Hypothese vorläufig aufrechterhalten werden kann oder durch eine Stichprobe als widerlegt gilt, wird eine Verteilung der möglichen Ergebnisse t einer Testfunktion betrachtet, die sich bei wahrer Hypothese ergeben hätte $[f(t|H_0)]$. Ist das eingetroffene Ergebnis als „unwahrscheinlich" einzustufen, so gilt die Hypothese als widerlegt. Je unwahrscheinlicher das Ergebnis wäre, d.h. je stärker die Widerlegung ausfällt, desto höher ist die **Signifikanz**.

Beispiel: Aufgabe 19, Seite 22: Stichprobenergebnisse $\bar{x}_i$. Sind Männer größer als 178cm oder nicht?

$$\mu_0 = 178,\ \sigma = 10,\ \alpha = 0{,}0446,\ T = \overline{X},\ H_0:\ \mu \leq \mu_0 = 178,\ f(t|H_0) = N(\mu, \tfrac{\sigma^2}{n})$$

		
für $\bar{x}_i$ keine Widerlegung von H_0	für $\bar{x}_i$ keine Widerlegung von H_0	für $\bar{x}_i$ Widerlegung von H_0

Hinweis zur Interpretation

Ein Ergebnis, das „signifikant" oder gar „hochsignifikant" ist (vgl. „purer" Signifikanztest, Seite 26), bedeutet nun nicht, dass es in der Sache wesentlich sei, sondern nur, dass der Verfahrenseinfluss vermutlich gering ist. Dies kann einfach z.B. durch einen großen Stichprobenumfang erreicht werden. Nichtsignifikanz, also kein Widerspruch zur Hypothese bedeutet ebensowenig, dass die Hypothese sachlich gerechtfertigt oder gar bestätigt wurde – sie wurde nur nicht mit der gewählten Verfahrensweise widerlegt.

Fehlermöglichkeiten bei Tests

Bei der geschilderten Vorgehensweise der Hypothesenprüfung – nämlich sehr unwahrscheinliche Ergebnisse (am Rand der Testverteilung) als Widerlegung aufzufassen –, geht man natürlich das Risiko ein, fälschlicherweise zu widerlegen. Das Risikomaß hierfür ist die **Irrtumswahrscheinlichkeit** α*, d.h. der Anteil all derjenigen Ergebnisse für* t*, die man als unwahrscheinlich bezeichnen würde.*

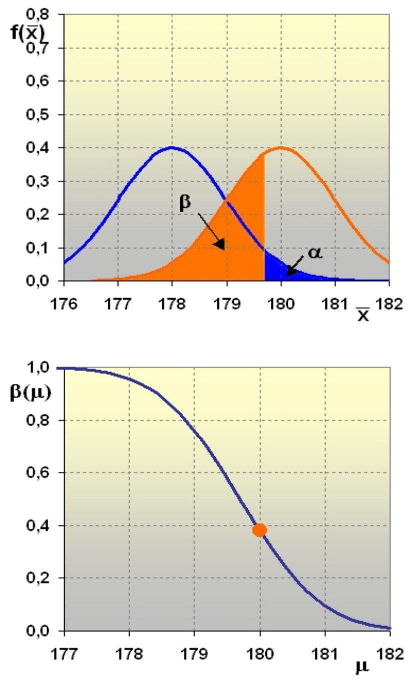
Testentscheidung	tatsächlicher Zustand	
	H_0 richtig	H_0 falsch
H_0 nicht verworfen	richtige Entscheidung	Fehler 2. Art (Wahrscheinlichkeit β)
H_0 verworfen	Fehler 1. Art (Wahrscheinlichkeit α)	richtige Entscheidung

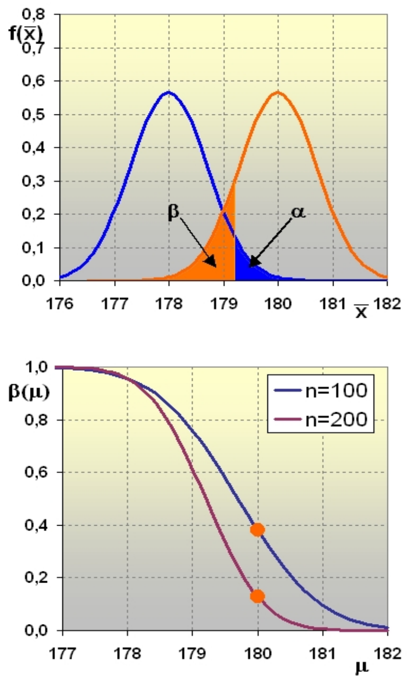
α wird beim klassischen Signifikanztest vorgegeben. Bei gegebener Testfunktion und ihrer Verteilung ist damit der Ablehnungsbereich für H_0 festgelegt. Manchmal wird erst nach der Stichprobenauswertung ein α berechnet, zu dem H_0 gerade noch nicht verworfen wird („purer" Signifikanztest). Je geringer dann α ausfällt, desto stärker ist die Widerlegung von H_0, d.h. desto höher ist die Signifikanz.

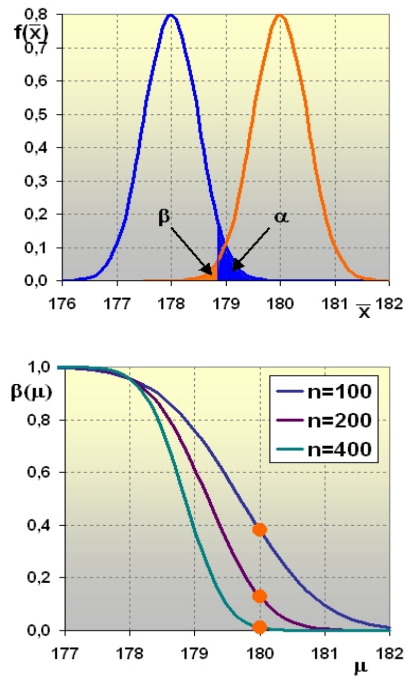
β hängt von einer Alternativhypothese H_1 ab, die in wissenschaftlichen Anwendungen selten als Punkthypothese (klassischer Alternativtest) formulierbar ist. $(1-\beta)$ wird als „Macht" – β als „Operationscharakteristik" – eines Tests bezeichnet und gilt als Auswahlkriterium: Hat man bei vorgegebenem α die Wahl zwischen verschiedenen Testverfahren, so wird man jenes mit der größten Macht wählen.

Beispiel: Aufgabe 19, Seite 22: $\mu_0 = 178$, $\sigma = 10$, $\alpha = 0{,}0446$, $H_0: \ \mu \le \mu_0 = 178 \implies z = 1{,}7$

$n = 100$	$n = 200$	$n = 400$
$n = 100$: $\sigma_{\overline{x}} = 1$, $c = 179{,}7$	$n = 200$: $\sigma_{\overline{x}} = 0{,}707$, $c = 179{,}2$	$n = 400$: $\sigma_{\overline{x}} = 0{,}5$, $c = 178{,}85$
$H_1: \mu \ge \mu_1 = 180$: $\beta = 0{,}3821$	$H_1: \mu \ge \mu_1 = 180$: $\beta = 0{,}1292$	$H_1: \mu \ge \mu_1 = 180$: $\beta = 0{,}0107$

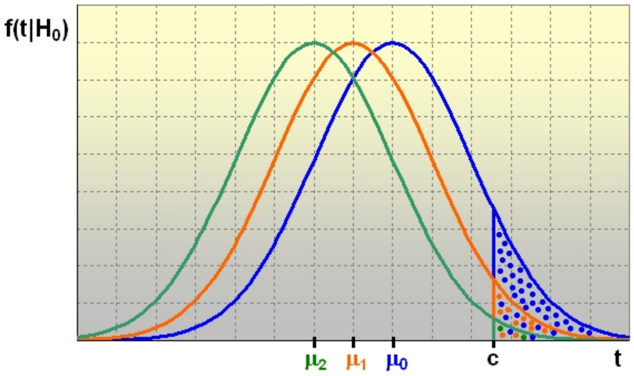






Praktische Vorgehensweise beim klassischen Signifikanztest

Eine Testentscheidung bzw. die Angabe eines Signifikanzniveaus wird getroffen auf der Grundlage einer Testverteilung bei Gültigkeit der Nullhypothese. Widerlegt man die H_0, dann wäre auch die Testverteilung und damit die so berechnete Irrtumswahrscheinlichkeit α falsch. Man wird deshalb die zu prüfende Hypothese bei einer Bereichshypothese als Bereichsgegenhypothese H_1 bzw. bei einer Punkthypothese als Bereichsgegenhypothesen H_1 und H_2 formulieren. Die Irrtumswahrscheinlichkeit erreicht dann höchstens α, auch wenn H_0 nicht zutrifft.



Bsp.: Bei $H_0: \ \mu \leq \mu_0$ ist für jedes $\mu < \mu_0$ das zugehöige α kleiner als bei $H_0: \ \mu = \mu_0$.

Z.B.:
- ● $\alpha = 0,1$, $\mu_0 = 178$, $\sigma = 10$, $n = 100$
- ● $\alpha_1 = 0,0374$, $\mu_1 = 177,5$
- ● $\alpha_2 = 0,0113$, $\mu_2 = 177$

$\Longrightarrow \quad c = 179,3$

Da H_0 also nie bestätigt, sondern höchstens nicht widerlegt werden kann, bedeutet damit eine Widerlegung von H_0 indirekt eine Bestätigung (und nicht nur Nicht-Widerlegung) von H_1.

Schritte	Beispiel 1	Beispiel 2
1. Formulierung von H_0	$H_0: \ \mu \leq \mu_0$ $H_1: \ \mu > \mu_0$	$H_0: \ \mu_1 \leq \mu_2$ $H_1: \ \mu_1 > \mu_2$
2. Wahl der Testfunktion	$T = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$	$T = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
3. Testverteilung, Wahl von α und Bestimmung des Ablehnungsbereichs	$N(0,1): \ z_{1-\alpha}$	$N(0,1): \ z_{1-\alpha}$ für $n_1 + n_2 - 2 > 30$
4. Stichprobenziehung und Berechnung von t		
5. Testentscheidung, d.h. Widerlegung von H_0 bei	$t > z_{1-\alpha}$	$t > z_{1-\alpha}$
Für weitere Tests vgl. „Häufig angewandte Testverfahren“.		

Auf-gabe 21

Deutsche Männer sind im Durchschnitt 178cm groß bei einer Streuung von $\sigma = 10$cm. 10% sind blond. Eine Stichprobe von 100 Managern in höheren Positionen ergab eine durchschnittliche Körpergröße von $\bar{x} = 175$cm. 13 Manager waren blond. Prüfen Sie bei einer Irrtumswahrscheinlichkeit von $\alpha = 0,0446$

a) die „Napoleon“-Hypothese: Im Beruf erfolgreiche Männer sind im Durchschnitt kleiner als alle,

b) die „Teutonen“-Hypothese: Unter den im Beruf erfolgreichen Männern gibt es mehr Blonde.

Häufig angewandte Testverfahren, α vorgegeben

(Hypothetische) Frage, die durch das Verfahren beantwortet werden soll	Zu vergleichende statistische Kenngrößen (Verteilungsvoraussetzung)	Nullhypothese H_0	Testfunktion T	Testverteilung T/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α, z.B. α = 0,05
Kann eine Stichprobe gemessen am arithmetischen Mittel aus einer bestimmten Grundgesamtheit stammen?	$\overline{X}$ und μ_0 bei bekanntem σ $(X \sim N(\mu,\sigma^2))$	$H_0:\ \mu = \mu_0$ $(H_0:\ \mu \le \mu_0$ $H_0:\ \mu \ge \mu_0)$	$\dfrac{\overline{X}-\mu_0}{\sigma}\sqrt{n}$	$N(0,1)$	$\lvert t\rvert > z_{1-\alpha/2}$ $(t > z_{1-\alpha}$ $t < -z_{1-\alpha})$
	$\overline{X}$ und μ_0 bei unbekanntem σ $(n \le 30:\ X \sim N(\mu,\sigma^2)$ $n > 30:\ X$ bel. vert.)	$H_0:\ \mu = \mu_0$ $(H_0:\ \mu \le \mu_0$ $H_0:\ \mu \ge \mu_0)$	$\dfrac{\overline{X}-\mu_0}{S}\sqrt{n}$	$t(n-1)$ bei $n > 30$ $N(0,1)$	$\lvert t\rvert > t_{1-\alpha/2}$ $\lvert t\rvert > z_{1-\alpha/2}$ $(t > t_{1-\alpha},\ t > z_{1-\alpha}$ $t < -t_{1-\alpha},\ t < -z_{1-\alpha})$
Unterscheiden sich zwei Stichproben oder stammen sie aus derselben Grundgesamtheit? (i = 1,2)	$\overline{X}_1$ und $\overline{X}_2$ mit $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, aber unbekannt $(n_i \le 30:\ X_i \sim N(\mu_i,\sigma_i^2)$ $n_i > 30:\ X_i$ bel. vert.)	$H_0:\ \mu_1 = \mu_2$ $(H_0:\ \mu_1 \le \mu_2$ $H_0:\ \mu_1 \ge \mu_2)$	$\dfrac{\overline{X}_1-\overline{X}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $\hat{\sigma}^2 = \dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$t(n_1+n_2-2)$ bei $n_1,n_2 > 30$ $N(0,1)$	$\lvert t\rvert > t_{1-\alpha/2}$ $\lvert t\rvert > z_{1-\alpha/2}$ $(t > t_{1-\alpha},\ t > z_{1-\alpha}$ $t < -t_{1-\alpha},\ t < -z_{1-\alpha})$
Unterscheiden sich mindestens zwei Stichproben beim Vergleich von r Stichproben? (i = 1, ..., r)	$\overline{X}_1,\overline{X}_2,\dots,\overline{X}_r$ mit $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2$, aber unbekannt $(X_i \sim N(\mu_i,\sigma_i^2))$	$H_0:\ \mu_1 = \mu_2 = \dots = \mu_r$	$\dfrac{\dfrac{\sum_{i=1}^{r} n_i(\overline{X}_i-\overline{X})^2}{r-1}}{\dfrac{\sum_{i=1}^{r}\sum_{k=1}^{n_i}(X_{ik}-\overline{X}_i)^2}{n-r}}$	$f(r-1,n-r)$ mit $n=\sum_{i=1}^{r} n_i$	$t > f_{1-\alpha}$
Kann eine Stichprobe gemessen an der Varianz aus einer beliebigen Grundgesamtheit stammen?	S^2 und σ_0^2 mit μ unbekannt $(X \sim N(\mu,\sigma^2))$	$H_0:\ \sigma^2 = \sigma_0^2$	$\dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$	$t > \chi^2_{1-\alpha}$
Unterscheiden sich zwei Stichproben bezüglich der Varianz?	S_1^2 und S_2^2 $(X_1 \sim N(\mu_1,\sigma_1^2)$ $X_2 \sim N(\mu_2,\sigma_2^2))$	$H_0:\ \sigma_1^2 = \sigma_2^2$	$\dfrac{S_1^2}{S_2^2}$	$f(n_1-1,n_2-1)$	$t > f_{1-\alpha}$
Sind zwei Merkmale statistisch verbunden?	h_{ij} und h_{ij}^e in einer Kreuztabelle mit m Zeilen und k Spalten	$H_0:\ \pi_{ij} = \pi_{ij}^e$	$\sum_{i=1}^{m}\sum_{j=1}^{k}\dfrac{(h_{ij}-h_{ij}^e)^2}{h_{ij}^e}$	$\chi^2([m-1][k-1])$	$t > \chi^2_{1-\alpha}$ (h_{ij}^e sollte größer als 5 sein)

7 Aufgaben zur Wiederholung

Aufgabe 22

Je 200 zufällig ausgewählte Politologen und Soziologen werden danach befragt, wieviele Klausuren sie zur Erlangung des Statistikscheines benötigten. Ergebnis:

Klausuren	1	2	3	4	5
Soziologen	80	60	40	20	–
Politologen	112	44	24	12	8

Benötig(t)en die Politologen weniger Anläufe?

Lösung: $t = 1{,}885$, also „ja“, sofern $\alpha > 0{,}0297$ $(\eta^2 = 0{,}009!)$

Aufgabe 23

Aus einer früheren Erhebung zu Bücherausgaben von Studenten hat man folgendes Ergebnis:

Wert von … bis unter … €	Anzahl der Studenten
0 – 10	500
10 – 30	500
30 – 60	500
60 – 100	500
100 – 150	500

a) Zeichnen Sie ein Histogramm und die Verteilungsfunktion. Bestimmen Sie die Quartile.

b) Berechnen Sie die Durchschnittsausgaben, die Varianz und den Variationskoeffizienten.

c) Eine neue Zufallsstichprobe ist geplant. Berechnen Sie den notwendigen Stichprobenumfang, wenn der relative Stichprobenfehler bei einer Aussagewahrscheinlichkeit von 95,45% nicht höher als 2% sein soll. Erläutern Sie, wie durch eine Schichtung ein geringerer Stichprobenumfang erreicht werden kann.

Lösung: a) $Q_1 = 15$, $Z = 45$, $Q_3 = 90$, b) $\hat{\bar{x}} = 55$, $s^2_{\text{int}} = 91{,}7$, $s^2_{\text{ext}} = 1\,870$, $s^2 = 1\,961{,}7$, $V = 0{,}805$, c) $n \geq 6\,800$ ($n_{\text{prop}} \geq 304$, sofern Schichtung entsprechend Klassierung und Klassenbesetzung)

Aufgabe 24

Auf die Frage „Haben Sie den Eindruck, dass die Euroeinführung zu Preiserhöhungen missbraucht wurde?“ antworteten 1000 zufällig ausgewählte Bürger „Eurolands“ wie folgt:

Antwort	AT	BeNeLux	DE	ES	FI	FR	GR	IE	IT	PT
Ja	5	50	180	50	5	140	5	5	140	20
Nein	20	50	50	85	15	65	30	10	60	15

a) Berechnen Sie Pearsons korrigierten C-Koeffizienten und interpretieren Sie das Ergebnis. Führen Sie einen χ^2-Test durch.

b) Berechnen Sie λ_y und interpretieren Sie das Ergebnis auch im Vergleich zum korrigierten C-Koeffizienten.

Lösung: a) $\chi^2 = 141{,}8$, $C = 0{,}35$, $C^* = 0{,}5$, $\chi^2_{1-\alpha}(9) = \chi^2_{0{,}95}(9) = 16{,}9$, $t > 16{,}9 \Longrightarrow H_0$ ablehnen b) $\lambda_y = 0{,}225$

Aufgabe 25

Leistungstest bei 250 Schülern:

a) In einem Test bei 250 zufällig ausgewählten 15jährigen Schülern in einem Bundesland wurde die Fähigkeit, Texte zu interpretieren mit der Fähigkeit, Textaufgaben in Mathematik zu lösen, verglichen (-1: unteres Drittel, 0: mittleres Drittel, +1: oberes Drittel):

Mathelösung	Texterfassung -1	0	+1
-1	50	20	10
0	20	50	30
+1	10	20	40

Berechnen Sie Kendall's τ_b und interpretieren Sie das Ergebnis.

b) Bei den drei Gruppen (-1: Gruppe 1 etc.) mit unterschiedlicher Texterfassungskompetenz wurde außerdem die Zeit (Std.) erfasst, die die Schüler pro Woche fernsehen:

Gruppe i	$\bar{x}_i$	s_i^2	n_i
1	30	100	80
2	25	80	90
3	20	60	80

Berechnen Sie den η^2-Koeffizienten und führen Sie einen F-Test ($\alpha = 0,05$) durch. Interpretieren Sie beide Ergebnisse im Zusammenhang.

Lösung: a) $n_c = 11\,600,\ n_d = 3000,\ T_x = 6000,\ T_y = 6200,\ T_{xy} = 4325,\ \tau_b = 0,415$
b) $\bar{x} = 25,\ \eta^2 = 0,1\bar{6},\ s^2 = 96,\ t = 24,7 > f_{1-\alpha}(2,247) = 2,995$

Aufgabe 26

Für sechzehn Arbeitslose ergibt sich folgender Zusammenhang zwischen dem Alter, dem Geschlecht und der seitherigen Dauer der Arbeitslosigkeit in Monaten:

Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Geschl.	m	w	m	m	m	w	m	m	w	w	m	m	m	w	w	m
Alter	26	42	34	40	28	52	42	54	46	36	38	48	46	30	38	40
Dauer	3	12	8	10	4	16	7	10	14	6	4	10	6	5	7	6

a) Stellen Sie in einem Streuungsdiagramm den Zusammenhang zwischen den Merkmalen Alter und Arbeitslosigkeitsdauer für diese Gruppen dar. Berechnen Sie eine lineare Regression nach der Methode der kleinsten Quadrate, das Bestimmtheitsmaß und interpretieren Sie es als PRE-Maß.

b) Berechnen Sie den η^2-Koeffizienten und interpretieren Sie ihn als PRE-Maß für den Einfluss des Geschlechts auf die Arbeitslosigkeitsdauer.

Lösung: a) $\hat{y} = -5,659 + 0,341x,\ r^2 = 0,5737,$ b) $\bar{y}_w = 10,\ \bar{y}_m = 68,\ \eta^2 = 0,185$

Anhang: Tafeln zu einigen wichtigen Verteilungen

A Standardnormalverteilung

Vertafelt sind die Werte der Verteilungsfunktion $F(z) = P(Z \leq z)$ für $z \geq 0$.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	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
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

B t-Verteilung

Vertafelt sind die Werte von t zu gegebenen Werten der Verteilungsfunktion für n Freiheitsgrade. Für $t_{1-\alpha}(n)$ gilt $F(t_{1-\alpha}(n)) = 1-\alpha$.

n	$1-\alpha$									
	0,600	0,700	0,750	0,800	0,900	0,950	0,975	0,990	0,995	0,999
1	0,325	0,727	1,000	1,376	3,078	6,314	12,706	31,821	63,656	318,289
2	0,289	0,617	0,816	1,061	1,886	2,920	4,303	6,965	9,925	22,328
3	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841	10,214
4	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604	7,173
5	0,267	0,559	0,727	0,920	1,476	2,015	2,571	3,365	4,032	5,894
6	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707	5,208
7	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499	4,785
8	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355	4,501
9	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250	4,297
10	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169	4,144
11	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106	4,025
12	0,259	0,539	0,695	0,873	1,356	1,782	2,179	2,681	3,055	3,930
13	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012	3,852
14	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977	3,787
15	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947	3,733
16	0,258	0,535	0,690	0,865	1,337	1,746	2,120	2,583	2,921	3,686
17	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898	3,646
18	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878	3,610
19	0,257	0,533	0,688	0,861	1,328	1,729	2,093	2,539	2,861	3,579
20	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845	3,552
21	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831	3,527
22	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819	3,505
23	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807	3,485
24	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,492	2,797	3,467
25	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787	3,450
26	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779	3,435
27	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771	3,421
28	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763	3,408
29	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756	3,396
30	0,256	0,530	0,683	0,854	1,310	1,697	2,042	2,457	2,750	3,385
40	0,255	0,529	0,681	0,851	1,303	1,684	2,021	2,423	2,704	3,307
50	0,255	0,528	0,679	0,849	1,299	1,676	2,009	2,403	2,678	3,261
100	0,254	0,526	0,677	0,845	1,290	1,660	1,984	2,364	2,626	3,174
150	0,254	0,526	0,676	0,844	1,287	1,655	1,976	2,351	2,609	3,145
∞	0,253	0,524	0,674	0,842	1,282	1,645	1,960	2,326	2,576	3,090

C Chi-Quadrat-Verteilung

Vertafelt sind die Werte von χ^2 zu gegebenen Werten der Verteilungsfunktion für n Freiheitsgrade. Für $\chi^2_{1-\alpha}(n)$ gilt $F(\chi^2_{1-\alpha}(n)) = 1-\alpha$. Approximation für $n > 35$: $\chi^2_{1-\alpha}(n) \approx \frac{1}{2}(z_{1-\alpha} + \sqrt{2n-1})^2$.

n	$1-\alpha$									
	0,600	0,700	0,800	0,900	0,950	0,975	0,980	0,990	0,995	0,999
1	0,708	1,074	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	1,833	2,408	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	2,946	3,665	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,266
4	4,045	4,878	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,466
5	5,132	6,064	7,289	9,236	11,070	12,832	13,388	15,086	16,750	20,515
6	6,211	7,231	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	7,283	8,383	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,321
8	8,351	9,524	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,124
9	9,414	10,656	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	10,473	11,781	13,442	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	11,530	12,899	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	12,584	14,011	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	13,636	15,119	16,985	19,812	22,362	24,736	25,471	27,688	29,819	34,527
14	14,685	16,222	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,124
15	15,733	17,322	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,698
16	16,780	18,418	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252
17	17,824	19,511	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,791
18	18,868	20,601	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	19,910	21,689	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,819
20	20,951	22,775	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,314
21	21,992	23,858	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,796
22	23,031	24,939	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	24,069	26,018	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	25,106	27,096	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,179
25	26,143	28,172	30,675	34,382	37,652	40,646	41,566	44,314	46,928	52,619
26	27,179	29,246	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,051
27	28,214	30,319	32,912	36,741	40,113	43,195	44,140	46,963	49,645	55,475
28	29,249	31,391	34,027	37,916	41,337	44,461	45,419	48,278	50,994	56,892
29	30,283	32,461	35,139	39,087	42,557	45,722	46,693	49,588	52,335	58,301
30	31,316	33,530	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,702
31	32,349	34,598	37,359	41,422	44,985	48,232	49,226	52,191	55,002	61,098
32	33,381	35,665	38,466	42,585	46,194	49,480	50,487	53,486	56,328	62,487
33	34,413	36,731	39,572	43,745	47,400	50,725	51,743	54,775	57,648	63,869
34	35,444	37,795	40,676	44,903	48,602	51,966	52,995	56,061	58,964	65,247
35	36,475	38,859	41,778	46,059	49,802	53,203	54,244	57,342	60,275	66,619

D F-Verteilung

Vertafelt sind die Werte von f zu gegebenen Werten der Verteilungsfunktion für (n_1, n_2) Freiheitsgrade. Für $f_{1-\alpha}(n_1, n_2)$ gilt $F(f_{1-\alpha}(n_1, n_2)) = 1-\alpha$.

n_1	$1-\alpha$	n_2										
		40	50	60	70	80	90	100	120	150	200	∞
1	0,900	2,835	2,809	2,791	2,779	2,769	2,762	2,756	2,748	2,739	2,731	2,706
1	0,950	4,085	4,034	4,001	3,978	3,960	3,947	3,936	3,920	3,904	3,888	3,841
1	0,975	5,424	5,340	5,286	5,247	5,218	5,196	5,179	5,152	5,126	5,100	5,024
1	0,990	7,314	7,171	7,077	7,011	6,963	6,925	6,895	6,851	6,807	6,763	6,635
2	0,900	2,440	2,412	2,393	2,380	2,370	2,363	2,356	2,347	2,338	2,329	2,303
2	0,950	3,232	3,183	3,150	3,128	3,111	3,098	3,087	3,072	3,056	3,041	2,996
2	0,975	4,051	3,975	3,925	3,890	3,864	3,844	3,828	3,805	3,781	3,758	3,689
2	0,990	5,178	5,057	4,977	4,922	4,881	4,849	4,824	4,787	4,749	4,713	4,605
3	0,900	2,226	2,197	2,177	2,164	2,154	2,146	2,139	2,130	2,121	2,111	2,084
3	0,950	2,839	2,790	2,758	2,736	2,719	2,706	2,696	2,680	2,665	2,650	2,605
3	0,975	3,463	3,390	3,343	3,309	3,284	3,265	3,250	3,227	3,204	3,182	3,116
3	0,990	4,313	4,199	4,126	4,074	4,036	4,007	3,984	3,949	3,915	3,881	3,782
4	0,900	2,091	2,061	2,041	2,027	2,016	2,008	2,002	1,992	1,983	1,973	1,945
4	0,950	2,606	2,557	2,525	2,503	2,486	2,473	2,463	2,447	2,432	2,417	2,372
4	0,975	3,126	3,054	3,008	2,975	2,950	2,932	2,917	2,894	2,872	2,850	2,786
4	0,990	3,828	3,720	3,649	3,600	3,563	3,535	3,513	3,480	3,447	3,414	3,319
5	0,900	1,997	1,966	1,946	1,931	1,921	1,912	1,906	1,896	1,886	1,876	1,847
5	0,950	2,449	2,400	2,368	2,346	2,329	2,316	2,305	2,290	2,274	2,259	2,214
5	0,975	2,904	2,833	2,786	2,754	2,730	2,711	2,696	2,674	2,652	2,630	2,566
5	0,990	3,514	3,408	3,339	3,291	3,255	3,228	3,206	3,174	3,142	3,110	3,017
6	0,900	1,927	1,895	1,875	1,860	1,849	1,841	1,834	1,824	1,814	1,804	1,774
6	0,950	2,336	2,286	2,254	2,231	2,214	2,201	2,191	2,175	2,160	2,144	2,099
6	0,975	2,744	2,674	2,627	2,595	2,571	2,552	2,537	2,515	2,494	2,472	2,408
6	0,990	3,291	3,186	3,119	3,071	3,036	3,009	2,988	2,956	2,924	2,893	2,802
7	0,900	1,873	1,840	1,819	1,804	1,793	1,785	1,778	1,767	1,757	1,747	1,717
7	0,950	2,249	2,199	2,167	2,143	2,126	2,113	2,103	2,087	2,071	2,056	2,010
7	0,975	2,624	2,553	2,507	2,474	2,450	2,432	2,417	2,395	2,373	2,351	2,288
7	0,990	3,124	3,020	2,953	2,906	2,871	2,845	2,823	2,792	2,761	2,730	2,639
8	0,900	1,829	1,796	1,775	1,760	1,748	1,739	1,732	1,722	1,712	1,701	1,670
8	0,950	2,180	2,130	2,097	2,074	2,056	2,043	2,032	2,016	2,001	1,985	1,938
8	0,975	2,529	2,458	2,412	2,379	2,355	2,336	2,321	2,299	2,278	2,256	2,192
8	0,990	2,993	2,890	2,823	2,777	2,742	2,715	2,694	2,663	2,632	2,601	2,511
9	0,900	1,793	1,760	1,738	1,723	1,711	1,702	1,695	1,684	1,674	1,663	1,632
9	0,950	2,124	2,073	2,040	2,017	1,999	1,986	1,975	1,959	1,943	1,927	1,880
9	0,975	2,452	2,381	2,334	2,302	2,277	2,259	2,244	2,222	2,200	2,178	2,114
9	0,990	2,888	2,785	2,718	2,672	2,637	2,611	2,590	2,559	2,528	2,497	2,407