FernUniversität in Hagen

Formelsammlung und Glossar zum Kurs „Statistik“ (Kurs 33209) mit Konzeptpapier – Stand: 5. Juni 2012

Inhaltsübersicht

1 Beschreibende Statistik

Univariate Häufigkeitsverteilungen

Häufigkeiten

Sei X ein diskretes Merkmal mit k Ausprägungen $a_1, a_2, ..., a_k$. Dann wird die **absolute Häufigkeit** für die Ausprägung a_i mit $h_i := h(a_i)$ und die **relative Häufigkeit** mit $f_i := f(a_i)$ bezeichnet $(i = 1, 2, \ldots, k)$ und es gilt für die relativen Häufigkeiten

$$f_i = \frac{h(a_i)}{n} \qquad i = 1, 2, \ldots, k.$$

Häufigkeitsverteilungen

Sei X ein zumindest ordinalskaliertes Merkmal mit Ausprägungen $a_1, a_2, ..., a_k$. Liegen die Ausprägungen nach aufsteigender Größe (bzw. nach aufsteigendem Rang) geordnet vor, so ist die **absolute kumulierte Häufigkeitsverteilung** für X gegeben durch

$$H(x) = h(a_1) + h(a_2) + \ldots + h(a_j) = \sum_{k=1}^{j} h(x_k).$$

Dabei ist a_j die größte Ausprägung des Merkmals X, die der Bedingung $a_i \leq x$ genügt. Die **relative kumulierte Häufigkeitsverteilung** $F(x)$ resultiert, wenn man noch durch den Umfang n des Datensatzes dividiert:

$$F(x) = \frac{H(x)}{n} = \sum_{k=1}^{j} f(x_k).$$

Für die auch als **empirische Verteilungsfunktion** bezeichnete Funktion $F(x)$ gilt

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f_1 & \text{für } a_1 \leq x < a_2 \\ \vdots & \vdots \\ f_1 + f_2 + \ldots + f_{k-1} & \text{für } a_{k-1} \leq x < a_k \\ 1 & \text{für } x \geq a_k. \end{cases}$$

Sie ist eine Treppenfunktion, die in $x = a_i$ $(i = 1, 2, \ldots, k)$ jeweils um f_i springt.

Lageparameter

Ein leicht zu bestimmender Lageparameter einer empirischen Verteilung ist der **Modus** oder **Modalwert** x_{mod}. Er bezeichnet die Merkmalsausprägung mit der größten Häufigkeit. Ein weiterer Lageparameter ist der Median $\widetilde{x}$. Hat man ein zumindest ordinalskaliertes Merkmal und Daten $x_1, x_2, \ldots, x_n$ und bezeichnet man den nach aufsteigender Größe (bei ordinalskaliertem Merkmal nach aufsteigendem Rangplatz) geordneten Datensatz mit $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, so ist der **Median** definiert durch

$$\widetilde{x} := \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \cdot (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{falls } n \text{ gerade.} \end{cases}$$

Bei metrisch skalierten Merkmalen kann man auch den **Mittelwert** $\overline{x}$ errechnen. Bei gegebenen Beobachtungswerten $x_1, x_2, \ldots, x_n$ ist er durch

$$\overline{x} := \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

erklärt. Alternativ bietet sich auch die folgende äquivalente Formel an:

$$\overline{x} = a_1 \cdot f_1 + a_2 \cdot f_2 + \ldots + a_k \cdot f_k = \sum_{i=1}^{k} a_i \cdot f_i.$$

Ein einfaches Streuungsmaß für metrisch skalierte Merkmale ist die **Spannweite** R eines Datensatzes. Sie ergibt sich aus dem geordneten Datensatz $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ als Differenz aus dem größten Wert $x_{(n)}$ und dem kleinsten Wert $x_{(1)}$: Streuungsparameter

$$R := x_{(n)} - x_{(1)}.$$

Ein weiteres Maß für die Streuung eines Datensatzes ist die **Varianz** oder **Stichprobenvarianz** s^2, die auch **empirische Varianz** genannt wird. Sie ist definiert durch

$$s^2 := \frac{1}{n} \cdot \left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\right] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Äquivalent ist die Darstellung

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2.$$

Alternativ zur Varianz kann man die **Standardabweichung** oder, genauer, die **empirische Standardabweichung** verwenden. Sie ist gegeben durch

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

Häufig wird für die Varianz eine Formel verwendet, bei der vor dem Summenterm anstelle von $\frac{1}{n}$ der Term $\frac{1}{n-1}$ steht. Das dann resultierende Streuungsmaß

$$s^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1} \cdot s^2$$

wird **korrigierte Varianz** oder **korrigierte Stichprobenvarianz** genannt. Durch Wurzelziehen geht aus s^{*2} die **korrigierte Standardabweichung** s^* hervor.

Wie bei der Berechnung des Mittelwertes $\overline{x}$ kann man auch bei der Ermittlung der Varianz im Falle mehrfach auftretender Merkmalswerte auf relative Häufigkeiten zurückgreifen. Liegen für ein diskretes Merkmal X mit den Ausprägungen $a_1, \ldots, a_k$ die Beobachtungswerte $x_1, \ldots, x_n$ vor $(n > k)$, so kann man s^2 auch wie folgt errechnen:

$$s^2 = (a_1 - \overline{x})^2 \cdot f_1 + (a_2 - \overline{x})^2 \cdot f_2 + \ldots + (a_k - \overline{x})^2 \cdot f_k = \sum_{i=1}^{k} (a_i - \overline{x})^2 \cdot f_i$$

Das p-Quantil ist bei einem mindestens ordinalskalierten Merkmal definiert durch

$$x_p = \begin{cases} x_{([np]+1)} & \text{falls } np \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{(np)} + x_{(np+1)}) & \text{falls } np \text{ ganzzahlig.} \end{cases}$$

Quantile

Dabei bezeichnet $[np]$ die größte ganze Zahl, die kleiner oder gleich np ist. Die Differenz $Q := x_{0,75} - x_{0,25}$ der als **oberes Quartil** und **unteres Quartil** bezeichneten beiden Quantile $x_{0,75}$ und $x_{0,25}$ heißt **Quartilsabstand**.

Konzentrationsmessung

Für die grafische Beurteilung von Konzentrationsphänomenen lässt sich die **Lorenzkurve** verwenden. Ausgangspunkt ist eine Grundgesamtheit mit n Merkmalsträgern und nicht-negativen Merkmalsausprägungen. Die Merkmalswerte konstituieren eine Urliste $x_1 \ldots, x_n$, aus der man durch Sortieren nach aufsteigender Größe eine geordnete Liste $x_{(1)} \ldots, x_{(n)}$ erhält. Die Lorenzkurve ist ein Polygonzug, der den Nullpunkt $(0;0)$ mit den Punkten $(u_1; v_1), \ldots, (u_n; v_n)$ verbindet. Dabei sind die Abszissenwerte u_i durch $u_i := \frac{i}{n}$ definiert und die Ordinatenwerte v_i durch $v_i := \frac{p_i}{p_n}$ mit

$$p_i := x_{(1)} + x_{(2)} + \ldots + x_{(i)}; \qquad i = 1, \ldots, n.$$

Führt man noch die gewichtete Merkmalssumme

$$q_n := 1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \ldots + n \cdot x_{(n)}$$

ein, so ist der **Gini-Koeffizient** G durch

$$G = \frac{2 \cdot q_n}{n \cdot p_n} - \frac{n+1}{n} = \frac{1}{n}\left(\frac{2 \cdot q_n}{p_n} - 1\right) - 1$$

erklärt. Für ihn gilt $0 \leq G \leq \frac{n-1}{n}$, d. h. er besitzt eine von n abhängige kleinste obere Schranke $G_{max} = \frac{n-1}{n}$. Für den **normierten Gini-Koeffizienten**

$$G^* := \frac{G}{G_{max}} = \frac{n}{n-1} \cdot G$$

gilt hingegen $0 \leq G^* \leq 1$. Als Alternative zum Gini-Koeffizienten findet man auch den **Herfindahl-Index** H, für den $\frac{1}{n} \leq H \leq 1$ gilt:

$$H := \sum_{i=1}^{n} \left(\frac{x_i}{p_n}\right)^2 = \frac{1}{p_n^2} \cdot \sum_{i=1}^{n} x_i^2.$$

Index- und Verhältniszahlen

Arten von Verhältniszahlen

Wenn man zwei Maßzahlen dividiert, resultiert eine **Verhältniszahl**. Verhältniszahlen, bei denen eine Grundgesamtheit durch Anteilsbildung bezüglich *eines Merkmals* strukturiert wird, nennt man **Gliederungszahlen**. Sie sind dimensionslos. Verhältniszahlen, die durch Quotientenbildung eine Verbindung zwischen *zwei* unterschiedlichen *Merkmalen* herstellen, heißen **Beziehungszahlen**. Die Verknüpfung der beiden Merkmale muss inhaltlich Sinn geben.

Bivariate Häufigkeitsverteilungen

Es seien zwei *diskrete* Merkmale X und Y mit beliebiger Skalierung und Ausprägungen $a_1, \ldots, a_k$ resp. $b_1, \ldots, b_m$ betrachtet. Die Merkmalswerte $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ repräsentieren eine **bivariate Urliste**. Diese lässt sich z. B. in der Form $(x_1, y_1), \ldots, (x_n, y_n)$ schreiben, wobei Merkmalspaare (x_i, y_i) mehrfach auftreten können. Die **absolute Häufigkeit** für die Ausprägungskombination (a_i, b_j) wird mit

$$h_{ij} := h(a_i, b_j) \qquad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m$$

bezeichnet und die **relative Häufigkeit** für (a_i, b_j) mit

$$f_{ij} := f(a_i, b_j) \qquad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m.$$

Die $k \cdot m$ Häufigkeiten h_{ij} und f_{ij} definieren die gemeinsame **absolute Häufigkeitsverteilung** resp. **relative Häufigkeitsverteilung** der Merkmale X und Y. Wenn man diese in tabellarischer Form wiedergibt, resultiert eine als **Kontingenztafel** oder **Kontingenztabelle** bezeichnete Darstellung. Die Dimension einer Kontingenztafel wird durch die Anzahl k und m der Ausprägungen für X und Y bestimmt. Im Falle von $k \cdot m$ Ausprägungskombinationen spricht man von einer $(k \times m)$-Kontingenztabelle. Ein Spezialfall einer Kontingenztabelle ist die **Vierfeldertafel**, die sich für $k = m = 2$ ergibt.

Kontingenztafeln werden üblicherweise noch um je eine weitere Zeile und Spalte ergänzt, wobei die zusätzliche *Spalte* bei einer Kontingenztabelle für absolute Häufigkeiten die k Zeilensummen

$$h_{i\cdot} := h_{i1} + h_{i2} + \ldots + h_{im} = \sum_{j=1}^{m} h_{ij} \qquad i = 1, 2, \ldots, k$$

und analog bei einer Tabelle für relative Häufigkeiten die Summen

$$f_{i\cdot} := f_{i1} + f_{i2} + \ldots + f_{im} = \sum_{j=1}^{m} f_{ij} \qquad i = 1, 2, \ldots, k$$

ausweist. Die Häufigkeiten $h_{1\cdot}, h_{2\cdot}, \ldots, h_{k\cdot}$ bzw. $f_{1\cdot}, f_{2\cdot}, \ldots, f_{k\cdot}$ werden **absolute Randhäufigkeiten** resp. **relative Randhäufigkeiten** von X genannt. Sie defininieren die **Randverteilung** von X. Die zusätzliche *Zeile*, um die man eine Kontingenztafel erweitert, enthält bei Verwendung absoluter Häufigkeiten die m Spaltensummen

$$h_{\cdot j} := h_{1j} + h_{2j} + \ldots + h_{kj} = \sum_{i=1}^{k} h_{ij} \qquad j = 1, 2, \ldots, m$$

und analog im Falle relativer Häufigkeiten die Spaltensummen

$$f_{\cdot j} := f_{1j} + f_{2j} + \ldots + f_{kj} = \sum_{i=1}^{k} f_{ij} \qquad j = 1, 2, \ldots, m.$$

Diese Häufigkeiten sind die **absoluten Randhäufigkeiten** bzw. die **relativen Randhäufigkeiten** von Y. Sie konstituieren die **Randverteilung** von Y.

		Ausprägungen von Y							
		b_1	b_2	$\dots$	b_j	$\dots$	b_m		
Ausprägungen von X	a_1	h_{11}	h_{12}	$\dots$	h_{1j}	$\dots$	h_{1m}	$h_{1\cdot}$	**Randverteilung von X**
	a_2	h_{21}	h_{22}	$\dots$	h_{2j}	$\dots$	h_{2m}	$h_{2\cdot}$	
	$\vdots$			$\ddots$			$\vdots$	$\vdots$	
	a_i	h_{i1}	h_{i2}	$\dots$	h_{ij}	$\dots$	h_{im}	$h_{i\cdot}$	
	$\vdots$					$\ddots$	$\vdots$	$\vdots$	
	a_k	h_{k1}	h_{k2}	$\dots$	h_{kj}	$\dots$	h_{km}	$h_{k\cdot}$	
		$h_{\cdot 1}$	$h_{\cdot 2}$	$\dots$	$h_{\cdot j}$	$\dots$	$h_{\cdot m}$	n	
		Randverteilung von Y							

Bedingte Häufigkeiten

Dividiert man jedes der m Elemente h_{i1}, h_{i2}, $\dots$, h_{im} durch die Randhäufigkeit $h_{i\cdot}$, resultieren **bedingte relative Häufigkeiten** für Y unter der Bedingung $X = a_i$. Diese kürzt man mit $f_Y(b_j|a_i)$ ab:

$$f_Y(b_j|a_i) := \frac{h_{ij}}{h_{i\cdot}} \qquad j = 1, 2, \dots, m$$

Die m bedingten relativen Häufigkeiten $f_Y(b_1|a_i)$, $f_Y(b_2|a_i)$, $\dots$, $f_Y(b_m|a_i)$ definieren die **bedingte Häufigkeitsverteilung für** Y unter der Bedingung $X = a_i$.

Teilt man jedes der k Elemente h_{1j}, h_{2j}, $\dots$, h_{kj} durch die Randhäufigkeit $h_{\cdot j}$, so erhält man die relativen Häufigkeiten für a_1, a_2, $\dots$, a_k unter der Bedingung $Y = b_j$. Man kürzt diese **bedingten relativen Häufigkeiten** für X unter der Bedingung $Y = b_j$ mit $f_X(a_i|b_j)$ ab:

$$f_X(a_i|b_j) := \frac{h_{ij}}{h_{\cdot j}} \qquad i = 1, 2, \dots, k.$$

Die k bedingten relativen Häufigkeiten $f_X(a_1|b_j), f_X(a_2|b_j), \dots, f_X(a_k|b_j)$ konstituieren die **bedingte Häufigkeitsverteilung für** X unter der Bedingung $Y = b_j$.

Empirische Unabhängigkeit bzw. Abhängigkeit von X und Y bedeutet, dass für die Häufigkeiten h_{ij} der $(k \times m)$-Kontingenztafel

$$h_{ij} \begin{cases} = \widetilde{h}_{ij} & \text{bei fehlendem Merkmalszusammenhang} \\ \neq \widetilde{h}_{ij} & \text{bei Abhängigkeit der Merkmale} \end{cases}$$

gilt. Dabei ist

$$\widetilde{h}_{ij} := \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}.$$

Zusammenhangsmessung

Ein Zusammenhangsmaß für zwei nominalskalierte Merkmale X und Y mit den in einer $(k \times m)$-Kontingenztabelle zusammengefassten gemeinsamen Häufigkeiten h_{ij} ist der χ^2-**Koeffizient**

Nominalskalierte Merkmale

$$\chi^2 := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \widetilde{h}_{ij})^2}{\widetilde{h}_{ij}}.$$

Für diesen gilt $0 \leq \chi^2 \leq \chi^2_{max} = n \cdot (M-1)$ mit $M := \min(k; m)$. Die untere Schranke 0 wird erreicht, wenn die Merkmale empirisch unabhängig sind.

Ein aus dem χ^2-Koeffizienten abgeleitetes Zusammenhangsmaß, dessen Wert nicht mehr vom Umfang n des Datensatz abhängt, ist der durch

$$\Phi := \sqrt{\frac{\chi^2}{n}}$$

definierte **Phi-Koeffizient**. Für dieses Maß gilt $0 \leq \Phi \leq \Phi_{max} := \sqrt{M-1}$. Die obere Schranke Φ_{max} hängt von M ab. Diesen Nachteil vermeidet das Zusammenhangsmaß

$$V := \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \cdot (M-1)}},$$

das auch **Cramér's V** genannt wird und Werte zwischen 0 und 1 annimmt, also ein normiertes Zusammenhangsmaß darstellt. Mit dem Maß V lässt sich die Stärke von Merkmalszusammenhängen bei Kontingenztabellen beliebiger Dimension direkt vergleichen. Gilt $V = 1$, spricht man von vollständiger Abhängigkeit der beiden Merkmale.

Im Spezialfall einer Vierfeldertafel ($k = m = 2$) gilt

Spezialfall: Vierfeldertafel

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\cdot}h_{2\cdot}h_{\cdot 1}h_{\cdot 2}}$$

und Cramér's V stimmt hier mit dem Phi-Koeffizienten Φ überein:

$$\Phi = V = \frac{|h_{11}h_{22} - h_{12}h_{21}|}{\sqrt{h_{1\cdot}h_{2\cdot}h_{\cdot 1}h_{\cdot 2}}}.$$

Ein Zusammenhangsmaß für zwei metrisch skalierte Merkmale X und Y ist die **Kovarianz** oder **empirische Kovarianz**

Metrisch skalierte Merkmale

$$s_{xy} := \frac{1}{n} \cdot [(x_1 - \overline{x})(y_1 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Für diese gilt auch die Darstellung

$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y} = \overline{xy} - \overline{x} \cdot \overline{y}.$$

Die Kovarianz ist – wie Median, Mittelwert und Standardabweichung – maßstabsabhängig und nicht dimensionslos. Ein maßstabsunabhängiges und dimensionsloses Zusammenhangsmaß ist der **Korrelationskoeffizient nach Bravais-Pearson**

$$r := \frac{s_{xy}}{s_x \cdot s_y}.$$

Für r hat man auch die ausführlichere Formeldarstellung

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\sqrt{\overline{y^2} - \overline{y}^2}}.$$

Der Korrelationskoeffizient liegt stets zwischen -1 und $+1$.

Ordinalskalierte Merkmale

Ein Zusammenhangsmaß für ordinalskalierte Merkmale X und Y ist der **Rangkorrelationskoeffizient nach Spearman** r_{SP}. Bestimmt man für jeden Wert x_i und für jeden Wert y_i die Rangposition $rg(x_i)$ bzw. $rg(y_i)$ und zusätzlich jeweils für beide Merkmale die Mittelwerte $\overline{rg}_x$ resp. $\overline{rg}_y$ der Rangplätze, so ist r_{SP} definiert durch

$$r_{SP} = \frac{\sum_{i=1}^n (rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_{i=1}^n (rg(x_i) - \overline{rg}_x)^2} \cdot \sqrt{\sum_{i=1}^n (rg(y_i) - \overline{rg}_y)^2}}.$$

Auch für den Rangkorrelationskoeffizienten r_{SP} gilt, dass er zwischen -1 und $+1$ liegt.

Wenn kein Rangplatz mehrfach besetzt ist, vereinfacht sich die Formel für r_{SP} zu

$$r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)} \qquad d_i := rg(x_i) - rg(y_i).$$

.

2 Wahrscheinlichkeitsrechnung und schließende Statistik

Grundbegriffe der Wahrscheinlichkeitsrechnung

Grundbegriffe

Ein **Zufallsvorgang** ist ein Prozess, der zu einem von mehreren, sich gegenseitig ausschließenden Ergebnissen ω führt. Die möglichen Ergebnisse ω heißen **Elementarereignisse** und werden in der Menge $\Omega = \{\omega : \omega \text{ ist Elementarereignis}\}$ zusammengefasst, der **Ergebnismenge**. Diese kann endlich oder auch unendlich viele Elemente enthalten. Eine Teilmenge A von Ω heißt **Ereignis**. Das **Komplementärereignis** $\overline{A}$ zu A ist das Ereignis, das genau dann eintritt, wenn A nicht eintritt. Die Menge $\overline{A}$ umfasst alle Elementarereignisse, die zu Ω, nicht aber zu A gehören. Da auf jeden Fall eines der Elemente der Menge Ω als Ergebnis des Zufallsvorgangs realisiert wird, ist durch Ω ein **sicheres Ereignis** definiert. Das Komplementärereignis $\overline{\Omega}$ zum sicheren Ereignis Ω ist das **unmögliche Ereignis**, das durch die leere Menge $\emptyset$ dargestellt wird.

Durch Verknüpfung von Ereignissen entstehen zusammengesetzte Ereignisse. Diese werden häufig anhand von **Venn-Diagrammen** veranschaulicht. Letztere bestehen aus einem Rechteck, in dem die Ausgangsereignisse (Mengen A, B, ...) als Kreise oder Ellipsen dargestellt sind.

Rechenregeln für Wahrscheinlichkeiten

Die Bewertung der Chance für das Eintreten eines Ereignisses wird anhand einer Funktion P bewertet, die jedem Ereignis A eine als Wahrscheinlichkeit des Ereignisses A bezeichnete Zahl $P(A)$ zuordnet, welche den folgenden drei Bedingungen genügt (sog. Axiomensystem von Kolmogoroff): $P(A) \geq 0$, $P(\Omega) = 1$, $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \emptyset$. Hieraus lassen sich die nachstehenden Gleichungen ableiten:

$$P(\overline{A}) = 1 - P(A);\;\; P(A \cup B) = P(A) + P(B) - P(A \cap B);\;\; P(A \setminus B) = P(A) - P(A \cap B).$$

Um Wahrscheinlichkeiten berechnen zu können, benötigt man Zusatzinformationen über den jeweiligen Zufallsvorgang. Eine Zusatzinformation kann z. B. darin bestehen, dass man weiß, dass die Ergebnismenge endlich ist und die Wahrscheinlichkeiten für die n Elementarereignisse alle gleich groß sind. Ein Zufallsexperiment mit diesen Eigenschaften heißt **Laplace-Experiment**. Bei einem Laplace-Experiment lässt sich die Wahrscheinlichkeit für ein Ereignis A als Quotient aus der Anzahl der für A günstigen Fälle und der Anzahl aller möglichen Ergebnisse des Zufallsexperiments errechnen:

$$P(A) = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}.$$

Bei der Bestimmung dieses Quotienten bedient man sich der **Kombinatorik**. Dort veranschaulicht man Ergebnisse für Zufallsvorgänge mit endlicher Ergebnismenge häufig anhand des **Urnenmodells** - gedanklich ein Gefäß mit N durchnummerierten Kugeln, von denen n zufällig ausgewählt werden. Die Auswahl der Kugeln ist als Ziehung einer **Zufallsstichprobe** des Umfangs n aus einer Grundgesamtheit mit N Elementen zu interpretieren. Wenn jede denkbare Stichprobe des Umfangs n mit gleicher Wahrscheinlichkeit realisiert wird, liegt eine **einfache Zufallsstichprobe** vor.

Wieviele Möglichkeiten der Auswahl der n Elemente es gibt, hängt davon ab, ob die Elemente der Stichprobe nach der Ziehung jeweils wieder zurückgelegt werden oder ob ohne Zurücklegen ausgewählt wird (**Urnenmodell bzw. Stichprobenziehung mit / ohne Zurücklegen**). Die Anzahl hängt auch davon ab, ob es darauf ankommt, in welcher Reihenfolge die n nummerierten Kugeln gezogen werden (**Stichprobenziehung mit / ohne Berücksichtigung der Anordnung**). Formeln für die Berechnung der Anzahl der Möglichkeiten der Ziehung einer Stichprobe des Umfangs n aus einer Grundgesamtheit mit N Elementen in allen 4 Fällen sind nachstehender Tabelle zu entnehmen:

Art der Stichprobe	Ziehen *ohne* Zurücklegen	Ziehen *mit* Zurücklegen
Ziehen *mit* Berücksichtigung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
Ziehen *ohne* Berücksichtigung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

In der Tabelle treten Binomialkoeffizienten $\binom{n}{k}$ auf, die durch

$$\binom{n}{k} := \frac{n!}{(n-k)! \cdot k!}$$

erklärt sind mit $\binom{n}{0} = 1$ und $\binom{k}{1} = k$ sowie $\binom{n}{n} = 1$. Die Fakultät $k! := 1 \cdot 2 \cdot ... \cdot k$ ist das Produkt aus allen natürlichen Zahlen von 1 bis k. Ferner ist 0! durch $0! = 1$ erklärt.

Bedingte Wahrscheinlichkeiten

Bei der Berechnung von Wahrscheinlichkeiten bei Laplace-Experimenten kann man manchmal eine gegebene Zusatzinformation B nutzen. Die mit der Vorinformation B berechnete Wahrscheinlichkeit wird **bedingte Wahrscheinlichkeit** von A unter der Bedingung B genannt und mit $P(A|B)$ abgekürzt. Sie errechnet sich nach

$$P(A|B) = \frac{\text{Anzahl der für } A \cap B \text{ günstigen Ergebnisse}}{\text{Anzahl der für } B \text{ günstigen Ergebnisse}} = \frac{P(A \cap B)}{P(B)}$$

Analog lässt sich die bedingte Wahrscheinlichkeit $P(B|A)$ gemäß $P(B|A) = \frac{P(A \cap B)}{P(A)}$ errechnen. Zwischen den bedingten Wahrscheinlichkeiten $P(A|B)$ und $P(B|A)$ besteht die auch als **Satz von Bayes** bezeichnete Beziehung

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

Zwei zufällige Ereignisse A und B werden als **unabhängig** oder auch als **stochastisch unabhängig** bezeichnet, wenn das Eintreten eines Ereignisses keinen Einfluss auf das andere Ereignis hat. Dies ist gewährleistet, wenn $P(A \cap B) = P(A) \cdot P(B)$ gilt.

Unabhängigkeit von Ereignissen

Diskrete Zufallsvariablen

Hat man eine diskrete Zufallsvariable X, die k Werte $x_1, \ldots, x_k$ annehmen kann, so definieren diese Werte die **Trägermenge** der Zufallsvariablen X. Das Verhalten von X ist vollständig definiert, wenn für jede Realisation x_i die Eintrittswahrscheinlichkeit $p_i = P(X = x_i)$ bekannt ist; $i = 1, \ldots, k$. Die Funktion f, die jeder Ausprägung x_i eine Eintrittswahrscheinlichkeit p_i zuordnet, heißt **Wahrscheinlichkeitsfunktion** von X. Damit die Wahrscheinlichkeitsfunktion nicht nur auf der Trägermenge $\{x_1, \ldots, x_k\}$, sondern für alle reellen Zahlen x erklärt ist, setzt man sie Null für alle x mit $x \neq x_i$:

$$f(x) = \begin{cases} p_i & \text{für } x = x_i;\ i = 1, 2, \ldots, k \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Diskrete Gleichverteilung

Wenn alle Ausprägungen x_i die gleiche Eintrittswahrscheinlichkeit $p = \frac{1}{k}$ besitzen, spricht man von einer **diskreten Gleichverteilung** mit Parameter p.

Zur Beschreibung des Verhaltens einer diskreten Zufallsvariablen X lässt sich anstelle der Wahrscheinlichkeitsfunktion auch die **Verteilungsfunktion**

$$F(x) = P(X \leq x)$$

von X heranziehen, die man auch **theoretische Verteilungsfunktion** nennt. Für $F(x)$ gilt im Falle einer diskreten Zufallsvariablen mit der Trägermenge $\{x_1, \ldots, x_k\}$

$$F(x) = \begin{cases} 0, & \text{für } x < x_1 \\ p_1 & \text{für } x_1 \leq x < x_2 \\ \vdots & \vdots \\ p_1 + p_2 + \ldots + p_{k-1} & \text{für } x_{k-1} \leq x < x_k \\ 1 & \text{für } x \geq x_k. \end{cases}$$

Bernoulli-Verteilung

Ein weiterer Spezialfall einer diskreten Verteilung ist die **Bernoulli-Verteilung**. Sie liegt vor, wenn eine X eine **binäre Zufallsvariable** ist, also nur zwei Ausprägungen aufweist, etwa x_1 und x_2 oder A und $\overline{A}$. Wenn man die Ausprägungen x_1 und x_2 zu 1 und 0 umcodiert, spricht man auch von einer **Null-Eins-Verteilung**. Bezeichnet $p_1 = p$ bei einer Bernoulli-Verteilung die Eintrittswahrscheinlichkeit für den Fall $x = x_1$ und p_2 die für den Fall $x = x_2$, so ist $p_2 = 1 - p$. Für die Wahrscheinlichkeitsfunktion gilt dann

$$f(x) = \begin{cases} p & \text{für } x = x_1; \\ 1 - p & \text{für } x = x_2; \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Für die Verteilungsfunktion $F(x)$ der Bernoulli-Verteilung leitet sich daraus ab:

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{für } x < x_1; \\ p & \text{für } x_1 \leq x < x_2; \\ 1 & \text{für } x \geq x_2. \end{cases}$$

Für eine mit dem Parameter p bernoulli-verteilte Zufallsvariable X sagt man auch, dass sie $Be(p)$-verteilt sei und verwendet hierfür die Notation $X \sim Be(p)$.

Kenngrößen einer diskreten Zufallsvariablen

Der **Erwartungswert** $E(X)$ einer diskreten Zufallsvariablen mit der Trägermenge $\{x_1, \ldots, x_k\}$ ist gegeben durch

$$\mu := E(X) = \sum_{i=1}^{k} x_i p_i.$$

Für die mit $V(X)$ oder σ^2 abgekürzte **Varianz** $V(X) = E\left[(X-\mu)^2\right]$ gilt, wenn X wieder als diskret spezifiziert ist mit der Trägermenge $\{x_1, \ldots, x_k\}$, die Darstellung

$$\sigma^2 := V(X) = \sum_{i=1}^{k} (x_i - \mu)^2 p_i.$$

Die **Standardabweichung** σ von X ist definiert durch $\sigma = \sqrt{V(X)}$. Für die Varianz ist manchmal die Darstellung $\sigma^2 = E(X^2) - \mu^2$ nützlich, die nicht nur im diskreten Fall gilt und auch als **Verschiebungssatz** angesprochen wird.

Für Erwartungswert und Varianz der Null-Eins-Verteilung gilt $\mu = 1 \cdot p + 0 \cdot (1-p) = p$ resp. $\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1-p)$.

Operationen mit Zufallsvariablen

Unterzieht man eine Zufallsvariable X mit Erwartungswert $\mu = E(X)$ einer Lineartransformation $Y = aX + b$, so ergeben sich Erwartungswert und Varianz nach

$$E(aX+b) = a \cdot E(X) + b; \quad V(aX+b) = a^2 \cdot V(X).$$

Für den Erwartungswert und die Varianz der Summe zweier unabhängiger Zufallsvariablen X und Y gilt ferner $E(X+Y) = E(X) + E(Y)$ sowie $V(X+Y) = V(X) + V(Y)$.

Quantile als weitere Kenngrößen

Wie bei empirischen Verteilungen kann man auch bei theoretischen Verteilungen **Quantile** zur Charakterisierung heranziehen. Das p-**Quantil** einer Verteilung ist durch

$$F(x_p) = p \quad (0 < p < 1)$$

definiert, also durch den Wert x_p der Verteilungsfunktion $F(x)$, an dem $F(x)$ den Wert p annimmt. Der **Median** $\widetilde{x} = x_{0,5}$ sowie das **untere Quartil** $x_{0,25}$ und das **obere Quartil** $x_{0,75}$ einer theoretischen Verteilung sind spezielle Quantile, die sich bei Wahl von $p = 0,5$ resp. von $p = 0,25$ und $p = 0,75$ ergeben.

Die Binomialverteilung

Hat man ein Bernoulli-Experiment mit den möglichen Ausgängen $x_1 = A$ und $x_2 = \overline{A}$ und den Eintrittswahrscheinlichkeiten $P(A) = p$ bzw. $P(\overline{A}) = 1 - p$ mehrfach und unabhängig voneinander durchgeführt, so interessiert man sich oft dafür, wie oft eine der beiden Realisationen auftritt, etwa A. Ist n die Anzahl der unabhängig durchgeführten Bernoulli-Experimente und bezeichnet X die Anzahl der Ausgänge A, so ist die Zählvariable X eine diskrete Zufallsvariable mit den Ausprägungen $i \quad (i = 0, 1, \ldots, n)$. Wenn man den Ausgang jedes der n Bernoulli-Experimente anhand einer Indikatorvariablen

$$X_i = \begin{cases} 1 & \text{bei Eintritt von } x_1 = A \\ 0 & \text{bei Eintritt von } x_2 = \overline{A} \end{cases}$$

beschreibt (null-eins-verteilte Zufallsvariable), so lässt sich X als Summe

$$X = \sum_{i=1}^{n} X_i$$

der n voneinander unabhängigen Indikatorvariablen schreiben. Die Verteilung der Zählvariablen X heißt **Binomialverteilung**. Die Bernoulli-Verteilung ist ein Spezialfall der Binomialverteilung ($n = 1$).

Für die Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ der Binomialverteilung gilt

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für } x = 0, 1, \ldots, n \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

und für ihre **Verteilungsfunktion** $F(x) = P(X \leq x)$ auf der Trägermenge $\{0, 1, \ldots, n\}$.

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \qquad x = 0, 1, \ldots, n.$$

Für den Erwartungswert $\mu = E(X)$ und die Varianz $\sigma^2 = V(X)$ einer binomialverteilten Variablen X verifiziert man die Darstellungen

$$\mu = n \cdot p; \quad \sigma^2 = n \cdot p(1-p).$$

Die Binomialverteilung beschreibt das Zufallsverhalten einer Zählvariablen X bei einem n-fach durchgeführten Bernoulli-Experiment, wobei die einzelnen Experimente voneinander unabhängig sind. Die Zählvariable weist aus, wie häufig einer der beiden möglichen Ausgänge $x_1 = A$ und $x_2 = \overline{A}$ und $P(A) = p$ bzw. $P(\overline{A}) = 1 - p$ auftrat. Die Binomalverteilung lässt sich durch das Urnenmodell *mit* Zurücklegen veranschaulichen.

Hypergeometrische Verteilung

Wenn man hingegen einer Urne mit N Kugeln, von denen M rot und die restlichen $N - M$ schwarz sind, nacheinander n Kugeln *ohne* Zurücklegen entnimmt, so repräsentiert die Ziehung jeder Kugel zwar weiterhin ein Bernoulli-Experiment, die Einzelexperimente sind aber nicht mehr unabhängig. Die Eintrittswahrscheinlichkeit für das interessierende Ereignis wird jetzt nicht nur von M, sondern auch vom Umfang N der Grundgesamtheit beeinflusst. Die Verteilung der Zählvariablen X ist bei einer Stichprobenentnahme ohne Zurücklegen nicht mehr durch eine Biomialverteilung gegeben, sondern durch die **hypergeometrische Verteilung**. Letztere ist durch drei Parameter beschrieben, nämlich durch N, M und n, und man schreibt $X \sim H(n; M; N)$.

Die Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ der hypergeometrischen Verteilung besitzt die Darstellung

$$f(x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{für } x \in T \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Für die Verteilungsfunktion $F(x) = P(X \leq x)$ gilt dann auf der Trägermenge

$$F(x) = \sum_{k=0}^{x} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad x \in T.$$

Da die Wahrscheinlichkeitsfunktion für $x \notin T$ stets 0 ist, bleibt $F(x)$ zwischen zwei benachbarten Elementen der Trägermenge auf dem Niveau des kleineren Werts, um dann in $x_{max} = min(n; M)$ den Endwert 1 anzunehmen (Treppenfunktion).

Erwartungswert $\mu = E(X)$ und Varianz $\sigma^2 = V(X)$ der hypergeometrischen Verteilung sind gegeben durch

$$\mu = n \cdot \frac{M}{N}; \quad \sigma^2 = n \cdot \frac{M}{N}(1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}.$$

Stetige Zufallsvariablen

Diskrete Zufallsvariablen sind dadurch gekennzeichnet, dass man die Anzahl ihrer Ausprägungen abzählen kann. Das Zufallsverhalten einer diskreten Zufallsvariablen X mit k Ausprägungen x_i $(i = 1, \ldots, k)$ und den Eintrittswahrscheinlichkeiten $p_i = P(X = x_i)$ lässt sich vollständig durch die Wahrscheinlichkeitsfunktion $f(x)$ oder die Verteilungsfunktion $F(x)$ charakterisieren.

Bei *stetigen* Zufallsvariablen ist die Trägermenge, also die Menge der möglichen Realisationen, ein *Intervall*. Das Verhalten einer stetigen Zufallsvariablen X lässt sich wie im diskreten Fall durch die **Verteilungsfunktion**

$$F(x) = P(X \leq x)$$

vollständig charakterisieren. Anstelle der Wahrscheinlichkeitsfunktion verwendet man hier die **Dichtefunktion**, kurz auch **Dichte** genannt. Diese Funktion $f(x)$ nimmt nur nicht-negative Werte an und hat die Eigenschaft, dass sich jeder Wert $F(x)$ der Verteilungsfunktion durch Integration der Dichte bis zur Stelle x ergibt:

$$F(x) = \int_{-\infty}^{x} f(t)dt \text{ für alle reellen } x.$$

Für alle Werte x, bei denen die Dichtefunktion $f(x)$ stetig ist, stimmt sie mit der Ableitung $F'(x)$ der Verteilungsfunktion überein:

$$F'(x) = f(x).$$

Für die Differenz $F(b) - F(a)$ von Werten der Verteilungsfunktion gilt

$$F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_{a}^{b} f(t)dt.$$

Die Gesamtfläche unter der Dichtekurve besitzt den Wert 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Eine einfache stetige Verteilung ist die **Rechteckverteilung**, auch **stetige Gleichverteilung** genannt. Man nennt eine stetige Zufallsvariable *rechteckverteilt* oder *gleichverteilt* über dem Intervall $[a, b]$, wenn sie die Dichtefunktion

Rechteckverteilung

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{für alle sonstigen } x \end{cases}$$

besitzt. Für die Verteilungsfunktion $F(x)$ einer über $[a, b]$ rechteckverteilten Zufallsvariablen X gilt

$$F(x) = \begin{cases} 0 & \text{für } x < a; \\ \frac{x-a}{b-a} & \text{für } a \leq x \leq b; \\ 1 & \text{für } x > b. \end{cases}$$

Der **Erwartungswert** $E(X)$ einer stetigen Zufallsvariablen ist gegeben durch

Kenngrößen einer stetigen Zufallsvariablen

$$\mu := E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

und die **Varianz** $V(X) = E[(X - \mu)^2]$ durch

$$\sigma^2 := V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx.$$

Die Standardabweichung σ (lies: *sigma*) ist wieder durch $\sigma = \sqrt{V(X)}$ erklärt.

Für den Erwartungswert und die Varianz der stetigen Gleichverteilung über $[a, b]$ gilt

$$\mu = E(X) = \frac{a+b}{2}; \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

Neben dem Erwartungswert und der Varianz bzw. der Standardabweichung kann man noch die **Quantile** x_p heranziehen, die durch $F(x_p) = p$ definiert sind.

Quantile

Hat man eine Zufallsvariable X mit Erwartungswert μ und Standardabweichung σ, so bezeichnet man die spezielle Lineartransformation

$$Z := \frac{X - \mu}{\sigma}$$

als **Standardisierung** oder auch als **z-Transformation**. Man verifiziert für die standardisierte Variable Z, dass $E(Z) = 0$ und $V(Z) = 1$.

Eine Zufallsvariable X folgt einer **Normalverteilung**, wenn ihre Dichte die Gestalt

Normalverteilung und Standardnormalverteilung

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{für alle reellen } x$$

besitzt. Hierfür wird oft die Notation $X \sim N(\mu; \sigma^2)$ verwendet. Die Verteilungsfunktion der Normalverteilung ist gegeben durch

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

Unterzieht man eine $N(\mu;\sigma^2)$-verteilte Zufallsvariable X einer Lineartransformation $Y = aX + b$, so ist auch Y normalverteilt:

$$X \sim N(\mu;\sigma^2), Y = aX + b \longrightarrow Y \sim N(a\mu + b; a^2\sigma^2)$$

Für den Erwartungswert und die Varianz der Summe zweier unabhängiger normalverteilter Zufallsvariablen X und Y gilt

$$X \sim N(\mu_X;\sigma_X^2), Y \sim N(\mu_Y;\sigma_Y^2), X \text{und } Y \text{unabhängig } \rightarrow X + Y \sim N(\mu_X + \mu_Y; \sigma_X^2 + \sigma_Y^2).$$

Operationen mit normalverteilten Zufallsvariablen

Hat man eine beliebig normalverteilte Zufallsvariable $X \sim N(\mu;\sigma^2)$, so kann man diese stets der speziellen Lineartransformation $Z := \frac{X-\mu}{\sigma}$ unterziehen. Für die resultierende Zufallsvariable Z gilt $Z \sim N(0,1)$:

$$X \sim N(\mu;\sigma^2) \quad \xrightarrow{\text{Transformation von } X \text{ in } Z=(X-\mu)/\sigma} \quad Z \sim N(0,1)$$

Für die Dichtefunktion der Standardnormalverteilung hat sich anstelle von $f(..)$ eine spezielle Notation eingebürgert, nämlich $\phi(..)$:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Für die Verteilungsfunktion der Standardnormalverteilung hat sich die Bezeichnung $\Phi(..)$ etabliert. Sie ist erklärt durch

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt.$$

Da die Dichtefunktion $\phi(z) = \Phi'(x)$ der Standardnormalverteilung symmetrisch zum Nullpunkt ist, gilt

$$\Phi(-z) = 1 - \Phi(z).$$

Mit den Werten $\Phi(z)$ kann man Werte $F(x)$ der Verteilungsfunktion *jeder* beliebigen Normalverteilung bestimmen und zwar gemäß

$$F(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Man leitet hieraus die folgenden Darstellungen ab:

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right); \quad P(X > a) = 1 - P\left(X \leq a\right) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right);$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Das p-Quantil der Normalverteilung ist der eindeutig bestimmte Wert x_p, an dem die Verteilungsfunktion $F(x)$ den Wert p erreicht. Insbesondere sind also die **p-Quantile**

der Standardnormalverteilung durch $\Phi(z_p) = p$ definiert. Da die Dichte der Standardnormalverteilung symmetrisch zum Nullpunkt ist, gilt $z_p = -z_{1-p}$.

Aus der Normalverteilung lassen sich einige Verteilungen ableiten. Es sind dies vor allem die χ^2-Verteilung, die t-Verteilung und die F-Verteilung. Geht man von n unabhängigen standardnormalverteilten Variablen $Z_1, Z_2, \ldots, Z_n$ aus und bildet die Summe

$$X := Z_1^2 + Z_2^2 + ... + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

der quadrierten Variablen, so sagt man, dass die Verteilung der Variablen X einer **χ^2-Verteilung** mit n Freiheitsgraden folgt und verwendet die Kurznotation $X \sim \chi_n^2$. Für den Erwartungswert und die Varianz einer χ_n^2-verteilten Variablen X lässt sich ableiten: χ^2-Verteilung

$$E(X) = n; \quad V(X) = 2n.$$

Die **Quantile** einer **χ^2-Verteilung** mit n Freiheitsgraden werden mit $\chi_{n;p}^2$ abgekürzt.

Aus der Standardnormalverteilung und der χ^2-Verteilung leitet sich die **t-Verteilung** ab. Sind X und Z unabhängige Zufallsvariablen mit $X \sim \chi_n^2$ und $Z \sim N(0;1)$, dann folgt die Zufallsvariable

$$T := \frac{Z}{\sqrt{\frac{X}{n}}}$$

einer t-Verteilung mit n Freiheitsgraden und man schreibt $T \sim t_n$. Für den Erwartungswert und die Varianz einer t_n-verteilten Variablen T lässt sich zeigen, dass

$$E(T) = 0; \quad V(T) = \frac{n}{n-2},$$

wobei die letzte Gleichung für $n \geq 3$ gilt. Die Funktionsdarstellungen für Dichte- und Verteilungsfunktion werden wie bei der χ^2-Verteilung nicht weiter benötigt. Die Dichte der t-Verteilung ist wie die der Standardnormalverteilung symmetrisch zum Nullpunkt. Mit zunehmender Anzahl n der Freiheitsgrade nähert sich aber die Dichte der t-Verteilung der der Standardnormalverteilung an. Für die **Quantile** $t_{n;p}$ der t-Verteilung gilt die Symmetriebeziehung $t_{n;p} = -t_{n;1-p}$. t-Verteilung

Aus der χ^2-Verteilung leitet sich die **F-Verteilung** ab. Sind X_1 und X_2 zwei unabhängige Zufallsvariablen mit $X_1 \sim \chi_m^2$ und $X_2 \sim \chi_n^2$, so folgt die Zufallsvariable F-Verteilung

$$Y := \frac{X_1/m}{X_2/n}$$

einer F-Verteilung mit m und n Freiheitsgraden und man schreibt $Y \sim F_{m;n}$. Mit $Y \sim F_{m;n}$ gilt $\frac{1}{Y} \sim F_{n;m}$. Bezeichnet man die p-**Quantile** einer $F_{m;n}$-verteilten Zufallsvariablen mit $F_{m;n;p}$ bzw. $F_{n;m;p}$, so gilt $F_{m;n;p} = \frac{1}{F_{n;m;1-p}}$. Man kann sich daher bei der Tabellierung von Quantilen der F-Verteilung auf Quantile $F_{m;n;p}$ mit $m \leq n$ beschränken.

Bivariate Verteilungen von Zufallsvariablen

Eine Zufallsvariable X, gleich ob diskret oder stetig, lässt sich durch die Verteilungsfunktion $F(x) = P(X \leq x)$ beschreiben. Hat man *zwei* Zufallsvariablen X und Y, so lässt sich deren Verteilung analog durch die **gemeinsame Verteilungsfunktion**

$$F(x; y) := P(X \leq x; Y \leq y)$$

charakterisieren. Sind $F_X(x) = P(X \leq x)$ und $F_Y(y) = P(Y \leq y)$ die Verteilungsfunktionen der Einzelvariablen, so sind X und Y **unabhängig** oder **stochastisch unabhängig**, wenn sich $F(x; y)$ von X und Y als Produkt

$$F(x; y) = F_X(X \leq x) \cdot F_Y(Y \leq y)$$

von $F_X(x)$ und $F_Y(y)$ darstellen lässt. Neben $F(x; y)$ lässt sich zur Charakterisierung der gemeinsamen Verteilung zweier Zufallsvariablen X und Y auch die Wahrscheinlichkeitsfunktion (diskreter Fall) resp. die Dichtefunktion (stetiger Fall) heranziehen.

Wichtige Stichprobenfunktionen

Zieht man aus einer Grundgesamtheit eine n-elementige Stichprobe, so wird diese in der schließenden Statistik durch Zufallsvariablen $X_1, X_2, ..., X_n$ modelliert, für die man dann Realisationen $x_1, x_2, ..., x_n$ hat und verwertet. Die Zufallsvariablen $X_1, X_2, ..., X_n$ werden meist anhand einer **Stichprobenfunktion** aggregiert:

$$X_1, X_2, ..., X_n \xrightarrow{\text{Verdichtung der Stichprobeninformation}} g(X_1, X_2, .., X_n)$$

Eine besonders wichtige Stichprobenfunktion ist der **Stichprobenmittelwert**

$$\overline{X} := \frac{1}{n} \cdot (X_1 + X_2 + \ldots + X_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i.$$

Eine weitere Stichprobenfunktion ist die **Stichprobenvarianz**

$$S^2 := \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$$

bzw. die **korrigierte Stichprobenvarianz**

$$S^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} \cdot S^2.$$

Verteilung des Stichprobenmittelwerts

Wenn die Stichprobenvariablen $X_1, X_2, ..., X_n$ alle unabhängig $N(\mu; \sigma^2)$-verteilt sind, so gilt für den Stichprobenmittelwert $\overline{X}$

$$\overline{X} \sim N(\mu; \sigma^2_{\overline{X}}) \quad \text{mit} \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}.$$

Wenn man den Stichprobenmittelwert standardisiert, folgt

$$\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} \sim N(0; 1).$$

Für die aus n unabhängigen $N(\mu; \sigma^2)$-verteilten Stichprobenvariablen X_i gebildete Stichprobenvarianz lässt sich eine Beziehung zur χ^2-Verteilung ableiten. Auch die Variablen X_i kann man zunächst standardisieren. Für die Summe der Quadrate der resultierenden standardnormalverteilten Variablen Z_i gilt, dass sie χ^2_n-verteilt ist:

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n.$$

Hieraus kann man ableiten, dass die mit dem Faktor $\frac{n}{\sigma^2}$ multiplizierte Stichprobenvarianz S^2 bzw. – äquivalent – die mit $\frac{n-1}{\sigma^2}$ multiplizierte korrigierte Stichprobenvarianz S^{*2} einer χ^2-Verteilung mit $n-1$ Freiheitsgraden folgt:

Verteilung der Stichprobenvarianz

$$\frac{n \cdot S^2}{\sigma^2} = \frac{(n-1) \cdot S^{*2}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \sim \chi^2_{n-1}.$$

Ferner lässt sich zeigen, dass eine Ersetzung von σ durch die als Schätzung für σ verwendete **korrigierte Stichprobenstandardabweichung** $S^* := \sqrt{S^{*2}}$ zu einer t-Verteilung mit $n-1$ Freiheitsgraden führt:

$$\frac{\overline{X} - \mu}{S} \cdot \sqrt{n-1} = \frac{\overline{X} - \mu}{S^*} \cdot \sqrt{n} \sim t_{n-1}.$$

Hat man *zwei* Zufallsvariablen X und Y mit Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$ und Varianzen $\sigma_X^2 = V(X)$ und $\sigma_Y^2 = V(Y)$, so kann man einen linearen Zusammenhang zwischen X und Y anhand der mit $Cov(X;Y)$ abgekürzten **Kovarianz** von X und Y messen (nicht-normiertes Zusammenhangsmaß). Letztere ist der Erwartungswert von $(X - \mu_X)(Y - \mu_Y)$:

Kovarianz und Korrelation

$$Cov(X;Y) := E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X) \cdot E(Y).$$

Wenn X und Y unabhängig sind, hat ihre Kovarianz stets den Wert 0, d. h. es gilt

$$X \text{ und } Y \text{ sind unabhängig } \rightarrow Cov(X;Y) = 0.$$

Sind X und Y zwei Zufallsvariablen mit der Kovarianz $Cov(X;Y)$, so gilt

$$V(X+Y) = V(X) + V(Y) + 2 \cdot Cov(X;Y).$$

Wie die empirische Kovarianz ist auch die theoretische Kovarianz maßstabsabhängig. Sie hat daher keine untere oder obere Schranke. Eine Normierung wird durch Verwendung des **Korrelationskoeffizienten** ρ erreicht. Dieser ist definiert durch

$$\rho = \frac{Cov(X;Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}.$$

Der Korrelationskoeffizient ρ liegt wie sein empirisches Analogon r stets zwischen -1 und 1. Im Falle $\rho = 0$ spricht man von **Unkorreliertheit**, im Falle $\rho \neq 0$ von **Korreliertheit** der Variablen X und Y. Unabhängigkeit von X und Y impliziert stets Unkorreliertheit:

$$X \text{ und } Y \text{ sind unabhängig } \rightarrow \rho = 0.$$

Schätzung von Parametern

Wenn man für ein stochastisches Merkmal X ein geeignetes Verteilungsmodell spezifiziert hat, sind die Parameter der Verteilung zunächst noch unbekannt und müssen geschätzt werden. Dabei kommen die Punkt- und die Intervallschätzung in Betracht. Mit einer **Punktschätzung** will man einen unbekannten Parameter möglichst gut treffen, während eine **Intervallschätzung** einen als **Konfidenzintervall** bezeichneten Bereich festlegt, in dem der unbekannte Parameter mit einer Wahrscheinlichkeit von mindestens $1-\alpha$ liegt, wobei α eine vorgegebene kleine Irrtumswahrscheinlichkeit ist.

Will man für einen unbekannten Parameter θ – z. B. den Erwartungswert oder die Varianz – eine Punktschätzung anhand von Stichprobendaten $x_1, x_2, ..., x_n$ gewinnen, verwendet man die Realisation einer **Stichprobenfunktion** $g(x_1, x_2, ..., x_n)$ als Schätzwert. Da die Stichprobendaten als Ausprägungen von Zufallsvariablen $X_1, X_2, ..., X_n$ interpretiert werden, ist auch der aus ihnen errechnete Schätzwert eine Realisation einer Zufallsvariablen $g(X_1, X_2, ..., X_n)$, die **Schätzstatistik**, **Schätzfunktion** oder kurz **Schätzer** genannt wird.

Eigenschaften von Schätzfunktionen

Ein Gütekriterium für eine Schätzfunktion ist die **Erwartungstreue** oder **Unverzerrtheit**. Diese beinhaltet, dass der Schätzer „im Mittel" den zu schätzenden Wert θ genau trifft, d. h. $E(\widehat{\theta}) = \theta$. Wenn ein Schätzer $\widehat{\theta}$ nicht erwartungstreu ist, heißt die Differenz

$$B(\widehat{\theta}) := E(\widehat{\theta}) - \theta = E(\widehat{\theta} - \theta)$$

Verzerrung oder **Bias**. Ein Schätzer $\widehat{\theta}$ heißt **asymptotisch erwartungstreu** oder **asymptotisch unverzerrt** wenn er zwar verzerrt ist, die Verzerrung aber gegen Null strebt, wenn der Stichprobenumfang n gegen ∞ (unendlich) konvergiert:

$$\lim_{n\to\infty} E(\widehat{\theta}) = \theta.$$

Ein Gütemaß für Schätzer, das sowohl die Verzerrung als auch die Streuung berücksichtigt, ist der mit **MSE** abgekürzte **mittlere quadratische Fehler**

$$MSE(\widehat{\theta}) := E\left[\left(\widehat{\theta} - \theta\right)^2\right] = V(\widehat{\theta}) + B(\widehat{\theta})^2.$$

Punktschätzung von Erwartungswerten

Will man den Erwartungswert μ einer Zufallsvariablen anhand der Ausprägungen unabhängiger Stichprobenvariablen $X_1, X_2, ..., X_n$ schätzen, verwendet man den Stichprobenmittelwert $\overline{X}$. Da man die Erwartungswertbildung auf die Stichprobenvariablen einzeln anwenden kann, gilt

$$E(\overline{X}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \ldots + E(X_n)] = \frac{1}{n} \cdot n \cdot \mu = \mu.$$

Wenn die Stichprobenvariablen $X_1, X_2, ..., X_n$ die Varianz σ^2 haben, gilt für $V(\overline{X}) = \sigma^2_{\overline{X}}$:

$$V(\overline{X}) = \frac{\sigma^2}{n}.$$

Verwendet man zur Schätzung der Varianz σ^2 einer Zufallsvariablen die **Stichprobenvarianz** S^2, so ist diese Schätzung verzerrt: Punktschätzung der Varianz

$$E(S^2) = \frac{n-1}{n} \cdot \sigma^2.$$

Eine *unverzerrte* Schätzung für σ^2 resultiert, wenn man anstelle von S^2 zur Varianzschätzung die **korrigierte Stichprobenvarianz** S^{*2} heranzieht:

$$E(S^{*2}) = \frac{n}{n-1} \cdot E(S^2) = \sigma^2.$$

Wenn man ein Bernoulli-Experiment n-mal durchführt, kann man den Ausgang der n Einzelexperimente anhand einer Folge unabhängiger null-eins-verteilter Stichprobenvariablen $X_1, X_2, ..., X_n$ modellieren.Verwendet man den hieraus gebildeten Stichprobenmittelwert $\overline{X}$ zur Schätzung des Erwartungswerts p der Null-Eins-Verteilung, so gilt Punktschätzung von Anteilswerten

$$E(\widehat{p}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \ldots + E(X_n)] = \frac{1}{n} \cdot n \cdot p = p.$$

Für die Varianz $V(\widehat{p})$ des Schätzers $\widehat{p}$ erhält man

$$V(\widehat{p}) = \frac{p \cdot (1-p)}{n}.$$

Bei einer **Intervallschätzung** wird anhand der Daten ein Intervall bestimmt, das den zu schätzenden Parameter θ mit einer Wahrscheinlichkeit von mindestens $1-\alpha$ enthält. Das Intervall soll eine möglichst geringe Länge aufweisen. Konfidenzintervalle für Erwartungswerte

Am einfachsten ist der Fall der Intervallschätzung des Erwartungswerts $\mu = E(X)$ eines $N(\mu;\sigma^2)$-verteilten Merkmals X, wenn die Varianz $\sigma^2 = V(X)$ bekannt ist. Die Zufallsvariable $Z := \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ ist dann standardnormalverteilt und liegt folglich mit Wahrscheinlichkeit $1-\alpha$ in dem durch die Quantile $z_{\alpha/2} = -z_{1-\alpha/2}$ und $z_{1-\alpha/2}$ begrenzten Intervall $\left[-z_{1-\alpha/2}; z_{1-\alpha/2}\right]$. Hieraus leitet man ab, dass

$$P\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Für den unbekannten Verteilungsparameter μ hat man also die Wahrscheinlichkeitsaussage, dass dieser mit Wahrscheinlichkeit $1-\alpha$ im hier mit KI bezeichneten Intervall

$$KI = \left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

liegt. Dies ist das **Konfidenzintervall** zum **Konfidenzniveau** $1-\alpha$ für μ, das eine Intervallschätzung für μ repräsentiert. Die Länge des Konfidenzintervalls ist durch

$$\text{Länge}(KI) = 2 \cdot z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

gegeben. Die vorstehenden Ableitungen sind leicht zu modifizieren, wenn man die Varianz σ^2 nur in Form einer Schätzung $\widehat{\sigma}^2$ kennt. Man erhält mit $\nu := n-1$

$$KI = \left[\overline{X} - t_{\nu;1-\alpha/2}\frac{S^*}{\sqrt{n}}; \overline{X} + t_{\nu;1-\alpha/2}\frac{S^*}{\sqrt{n}}\right].$$

Statistische Testverfahren

Klassifikationen für Tests

Wenn man für die Teststatistik die Kenntnis des Verteilungstyps in der Grundgesamtheit voraussetzt, liegt ein **parametrischer Test** vor, andernfalls ein **verteilungsfreier** oder **nicht-parametrischer Test**. Man kann Tests auch danach klassifizieren, worauf sich die Hypothesen beziehen. So gibt es **Tests für Erwartungswerte**, **Tests für Varianzen** oder **Tests für Anteile** von Populationen. Für die drei genannten Fälle gibt es Ein- und Mehrstichproben-Tests. Mit **Anpassungstests** wird untersucht, ob eine Zufallsvariable einer bestimmten Verteilung folgt, z. B. der Normalverteilung. Mit **Unabhängigkeitstests** will testen, ob zwei Zufallsvariablen stochastisch unabhängig sind.

Häufig werden statistische Tests, deren Prüfstatistik einer bestimmten diskreten oder stetigen Verteilung folgt, zu einer Gruppe zusammengefasst. So gibt es ganz unterschiedliche Tests, die mit einer χ^2-, t- oder F-verteilten Testgröße operieren. Diese Tests werden dann als **χ^2-Tests**, **t-Tests** resp. als **F-Tests** angesprochen. Ein Test mit normalverteilter Prüfstatistik wird auch als **Gauß-Test** bezeichnet.

Bei der Prüfung von Hypothesen über Parameter kann es darauf ankommen, Veränderungen nach beiden Seiten zu entdecken oder auch nur in eine Richtung. Man spricht dann von einem **zweiseitigen Test** bzw. einem **einseitigen Test**. Wenn zwei Hypothesen direkt aneinandergrenzen, wie etwa im Falle $H_0 : \mu = \mu_0$ und $H_1 : \mu \neq \mu_0$, spricht man von einem **Signifikanztest**. Andernfalls, etwa im Falle $H_0 : \mu = \mu_0$ und $H_1 : \mu = \mu_1$ ($\mu_0 < \mu_1$), liegt ein **Alternativtest** vor.

Grundbegriffe und Tests für Erwartungswerte

Die anhand eines Tests zu untersuchende Fragestellung wird in Form einer Nullhypothese H_0 und einer Alternativhypothese H_1 formuliert. Die **Nullhypothese** H_0 beinhaltet eine bisher als akzeptiert geltende Aussage über den Parameter einer Grundgesamtheit. Die **Alternativhypothese** H_1 beinhaltet die eigentliche Forschungshypothese.

Ein Test basiert auf einer **Prüfvariablen**, auch **Teststatistik** genannt, deren Ausprägung sich im Ein-Stichprobenfall aus einer Stichprobe $x_1, x_2, .., x_n$ ergibt. Letztere wird als Realisation von Stichprobenvariablen $X_1, X_2, .., X_n$ interpretiert. Die Stichprobenvariablen werden nicht direkt verwendet; man aggregiert sie vielmehr anhand einer Stichprobenfunktion $g(X_1, X_2, .., X_n)$, z. B. anhand des Stichprobenmittelwerts $\overline{X}$ oder der Stichprobenvarianz S^2 bzw. S^{*2}. Da die Stichprobenvariablen Zufallsvariablen sind, gilt dies auch für die Teststatistik. Die Testentscheidung hängt also von der Ausprägung $g(x_1, x_2, .., x_n)$ der herangezogenen Stichprobenfunktion ab.

Zweiseitiger Test für den Erwartungswert

Bei einem *zweiseitigen* Test für den Erwartungswert μ einer normalverteilten Variablen lauten die zu testenden Hypothesen

$$H_0 : \ \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \ \mu \neq \mu_0.$$

Wenn die Varianz σ^2 von X bekannt ist, gilt unter H_0, also für $\mu = \mu_0$, die Aussage $\overline{X} \sim N(\mu_0; \sigma^2_{\overline{X}})$ mit $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. Ein mit einer normalverteilten Prüfgröße operierender Test wird auch **Gauß-Test** genannt. Der mit $\overline{X}$ bzw. mit der standardisierten Prüfvariablen

$$Z := \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n}$$

operierende Test der obigen Hypothesen ist demnach ein *zweiseitiger* Gauß-Test. Für diesen gilt, dass eine Ausprägung z mit Wahrscheinlichkeit $1 - \alpha$ im Intervall $\left[z_{\alpha/2}; z_{1-\alpha/2}\right]$ liegt, wobei $z_{\alpha/2} = -z_{1-\alpha/2}$ und $z_{1-\alpha/2}$ Quanitle der Standardnormalverteilung sind. Das Intervall heißt **Annahmebereich** für H_0. Der Bereich außerhalb des genannten Intervalls definiert den **Ablehnungsbereich** für die Nullhypothese. Die Grenzen des Intervalls werden **kritische Werte** genannt. Im Falle der Verwerfung von H_0 ist die Alternativhypothese H_1 statistisch „bewiesen" in dem Sinne, dass ihre Gültigkeit mit einer Irrtumswahrscheinlichkeit α als gesichert angenommen werden kann. Die fälschliche Zurückweisung der Nullhypothese wird als **Fehler 1. Art** oder auch als **α-Fehler** bezeichnet. Die Wahrscheinlichkeit α ist das **Signifikanzniveau** des Tests.

Die Nullhypothese $H_0 : \mu = \mu_0$ wird beim zweiseitigen Gauß-Test mit Irrtumswahrscheinlichkeit α verworfen, wenn sich für die aus der Stichprobenfunktion $\widehat{\mu} = \overline{X}$ durch Standardisierung hervorgegangene Variable Z eine Realisation ergibt, die außerhalb des Intervalls $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$ liegt, wenn also $|z| > z_{1-\alpha/2}$ gilt.

Einseitiger Test für den Erwartungswert

Beim *einseitigen* Hypothesentest für den Erwartungswert μ besteht die Nullhypothese nicht nur aus einem einzigen Wert, sondern aus allen Werten unterhalb oder oberhalb eines bestimmten Wertes des zu testenden Parameters. Man testet nun entweder

$$H_0: \ \mu \leq \mu_0 \qquad \text{gegen} \qquad H_1: \ \mu > \mu_0 \qquad \textit{(rechtsseitiger Test)}$$

oder

$$H_0: \ \mu \geq \mu_0 \qquad \text{gegen} \qquad H_1: \ \mu < \mu_0 \qquad \textit{(linksseitiger Test)}.$$

Die Testentscheidung beim einseitigen Hypothesentest orientiert sich allein an der Verteilung der Prüfgröße im Grenzfall $\mu = \mu_0$. Das Signifikanzniveau α ist bei einem einseitigen Test als *obere Schranke* für den Eintritt eines Fehlers 1. Art zu interpretieren. Beim Übergang von einem zweiseitigen zu einem einseitigen Hypothesentest bleibt die Testgröße unverändert, aber die Bedingungen für die Ablehnung der Nullhypothese ändern sich. Beim *rechtsseitigen* Gauß-Test wird die Nullhypothese $H_0 : \mu \leq \mu_0$ verworfen, wenn die Bedingung $z > z_{1-\alpha}$ erfüllt ist. Beim *linksseitigen* Test mit $H_0 : \mu \geq \mu_0$ lautet die entsprechende Bedingung $z < z_\alpha$.

Fehlerarten beim Testen

Ein statistischer Test kann also zur Ablehnung der Nullhypothese H_0 führen (Entscheidung für H_1) oder zur Nicht-Verwerfung von H_0 (Beibehaltung von H_0 mangels Evidenz für H_1). Jede der beiden Testentscheidungen kann richtig oder falsch sein. Es gibt somit insgesamt vier denkbare Fälle, von denen zwei falsche Entscheidungen darstellen. Neben dem **Fehler 1. Art** oder **α-Fehler**, der fälschlichen Verwerfung der Nullhypothese, kann auch eine Nicht-Verwerfung einer nicht zutreffenden Nullhypothese eintreten. Diese Fehlentscheidung heißt **Fehler 2. Art** oder **β-Fehler**.

Testentscheidung	**tatsächlicher Zustand**	
	Nullhypothese richtig	Nullhypothese falsch
Nullhypothese nicht verworfen	richtige Entscheidung	**Fehler 2. Art** (β-Fehler)
Nullhypothese verworfen	**Fehler 1. Art** (α-Fehler)	richtige Entscheidung

Die genannten Fehlerwahrscheinlichkeiten sind bedingte Wahrscheinlichkeiten:

$$P(\text{Fehler 1. Art}) = P(\text{Ablehnung von } H_0 | H_0 \text{ ist wahr})$$
$$P(\text{Fehler 2. Art}) = P(\text{Nicht-Verwerfung von } H_0 | H_1 \text{ ist wahr}).$$

Bewertung der Leistungsfähigkeit eines Tests

Zur Beurteilung eines Tests für den Erwartungswert μ zieht man die **Gütefunktion**

$$G(\mu) = P(\text{Ablehnung von } H_0 | \mu)$$

des Tests heran. Diese gibt für jeden möglichen Wert des Erwartungswerts μ des normalverteilten Merkmals X die Wahrscheinlichkeit für die Verwerfung der Nullhypothese an, spezifiziert also die Ablehnungswahrscheinlichkeit für H_0 als Funktion von μ. Beim *zweiseitigen* Gauß-Test ist die Gütefunktion durch

$$G(\mu) = \Phi\left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right)$$

gegeben, während man für die *einseitigen* Testvarianten die nachstehenden Formeldarstellungen ableiten kann:

$$G(\mu) = 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) \qquad \textit{(rechtsseitiger Fall)}$$
$$G(\mu) = \Phi\left(-z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) \qquad \textit{(linksseitiger Fall).}$$

Vorgehensweise bei unbekannter Varianz

Der Test für den Erwartungswert μ einer normalverteilten Variablen ist zu modifizieren, wenn die Varianz σ^2 nur in Form einer Schätzung vorliegt. Die Prüfstatistik lautet nun

$$T := \frac{\overline{X} - \mu_0}{S^*} \cdot \sqrt{n}.$$

Diese Testvariable ist nicht mehr standardnormalverteilt, sondern t-verteilt mit $\nu := n-1$ Freiheitsgraden. Der Annahmebereich für den mit der obigen Prüfstatistik arbeitenden **t-Test** ist im *zweiseitigen* Fall durch $[-t_{\nu;1-\alpha/2}; t_{\nu;1-\alpha/2}]$ gegeben. Die Nullhypothese wird also bei Verwendung der Prüfstatistik T mit Irrtumswahrscheinlichkeit α verworfen, wenn die t_ν-verteilte Prüfgröße außerhalb des Intervalls $[-t_{\nu;1-\alpha/2}; t_{\nu;1-\alpha/2}]$ liegt, wenn also $|t| > t_{n-1;1-\alpha/2}$ gilt. Dieses Intervall ist stets breiter als das Intervall $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$, das den Annahmebereich des zweiseitigen Gauß-Tests repräsentiert. Die Unterschiede nehmen aber mit zunehmendem Wert von $\nu = n-1$ ab.

Beim *rechtsseitigen* t-Test wird die Nullhypothese $H_0 : \mu \leq \mu_0$ im Falle $t > t_{n-1;1-\alpha}$ verworfen,gilt, beim *linksseitigen* t-Test mit $H_0 : \mu \geq \mu_0$ für $t < t_{n-1;\alpha} = -t_{n-1;1-\alpha}$.

Es gibt eine Alternative für die Durchführung von Hypothesentests, bei der die Testentscheidung nicht auf dem Vergleich von Testvariablenwerten und kritischen Werten beruht, sondern auf dem Vergleich eines vorgegebenen Signifikanzniveaus α mit dem auch als **empirisches Signifikanzniveau** bezeichneten ***p*-Wert**. Dieser gibt bei gegebenem Stichprobenbefund das Niveau α' an, bei dem die Nullhypothese bei Verwendung des jeweiligen Datensatzes *gerade noch* verworfen würde.

p-Wert

Die Ausführungen über das Testen zwei- und einseitiger Hypothesen für Erwartungswerte bei normalverteiltem Merkmal lassen sich auf Hypothesen für Varianzen übertragen. Die Hypothesen im *zweiseitigen* Fall lauten nun

Tests für Varianzen

$$H_0: \ \sigma^2 = \sigma_0^2 \qquad \text{gegen} \qquad H_1: \ \sigma^2 \neq \sigma_0^2.$$

Der Test wird durchgeführt mit der Prüfstatistik

$$T := \frac{n \cdot S^2}{\sigma_0^2} = \frac{(n-1) \cdot S^{*2}}{\sigma_0^2},$$

die bei Gültigkeit von H_0 einer χ^2-Verteilung mit $\nu = n-1$ Freiheitsgraden folgt: $T \sim \chi^2_{n-1}$. Die Nullhypothese wird bei diesem **χ^2-Test** mit Irrtumswahrscheinlichkeit α verworfen, wenn die Realisation t der Prüfgröße entweder kleiner als $\chi^2_{n-1;\alpha/2}$ oder größer als $\chi^2_{n-1;1-\alpha/2}$ ist, wenn also der für die Testgröße berechnete Wert t außerhalb des Intervalls $[\chi^2_{\nu;\alpha/2}; \chi^2_{\nu;1-\alpha/2}]$ liegt mit $\nu = n-1$. Für den *einseitigen* Fall hat man

$$H_0: \ \sigma^2 \leq \sigma_0^2 \qquad \text{gegen} \qquad H_1: \ \sigma^2 > \sigma_0^2 \qquad \textit{(rechtsseitiger Test)}$$

resp.

$$H_0: \ \sigma^2 \geq \sigma_0^2 \qquad \text{gegen} \qquad H_1: \ \sigma^2 < \sigma_0^2 \qquad \textit{(linksseitiger Test).}$$

Beim rechtsseitigen Test wird H_0 mit einer Irrtumswahrscheinlichkeit von höchstens α verworfen, wenn für die Realisation t der Testgröße T die Bedingung $t > \chi^2_{n-1;1-\alpha}$ erfüllt ist. Die Ablehnbedingung fürH_0 beim linksseitigen Test lautet entsprechend $t < \chi^2_{n-1;\alpha}$.

Oft will man klären, ob es Niveauunterschiede zwischen zwei Teilpopulationen gibt, für die man je eine Stichprobe des Umfangs n_1 resp. n_2 hat. Formal interpretiert man in solchen Fällen die Daten aus beiden Stichproben als Ausprägungen zweier Zufallsvariablen X_1 und X_2. Letztere werden als unabhängig angenommen. Anhand eines **Zweistichproben-Tests** wird dann untersucht, ob sich die Erwartungswerte $\mu_1 := E(X_1)$ und $\mu_2 := E(X_2)$ signifikant unterscheiden. Getestet wird hier (*zweiseitiger* Fall)

Zwei-Stichproben-Tests für Erwartungswerte

$$H_0: \ \mu_1 = \mu_2 \qquad \text{gegen} \qquad H_1: \ \mu_1 \neq \mu_2.$$

Wie bei den Einstichproben-Tests für Erwartungswerte wird auch bei Zweistichproben-Tests i. a. Normalverteilung unterstellt, also $X_1 \sim N(\mu_1; \sigma_1^2)$ und $X_2 \sim N(\mu_2; \sigma_2^2)$. Man unterscheidet auch hier zwischen den Fällen bekannter und geschätzter Varianzen σ_1^2 und σ_2^2. In beiden Fällen geht man bei der Prüfvariablenkonstruktion von der Differenz

Prüfvariablen-konstruktion

$$D := \overline{X}_1 - \overline{X}_2$$

der Stichprobenmittelwerte aus. Wegen $\overline{X}_1 \sim N(\mu_1; \sigma_1^2)$ und $\overline{X}_2 \sim N(\mu_2; \sigma_2^2)$ und der vorausgesetzten Unabhängigkeit von X_1 und X_2 gilt

$$D \sim N(\mu_D; \sigma_D^2) \qquad \text{mit} \qquad \mu_D = \mu_1 - \mu_2 \text{ und } \sigma_D^2 = \sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Bei Gültigkeit von H_0 ist $\mu_D = 0$, also $D \sim N(0; \sigma_D^2)$, so dass man unter der Voraussetzung bekannter Varianzen σ_1^2 und σ_2^2 den Test anhand der Prüfgröße

$$Z = \frac{D}{\sigma_D} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

durchführen kann, die standardnormalverteilt ist. Haben die beiden Varianzen denselben Wert, etwa $\sigma^2 := \sigma_1^2 = \sigma_2^2$, vereinfacht sich die Testgröße zu

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}.$$

Die Nullhypothese wird bei diesem **Zweistichproben-Gauß-Test** mit Irrtumswahrscheinlichkeit α verworfen, wenn $|z| > z_{1-\alpha/2}$ gilt. Diese Aussage gilt unabhängig davon, ob die Varianzen übereinstimmen oder nicht.

Bei unbekannten Varianzen ist σ_D^2 zu schätzen. Bezeichnet man die korrigierten Varianzen der beiden Stichproben mit S_1^{*2} resp. S_2^{*2}, so liefert

$$\widehat{\sigma}_D^2 := \frac{(n_1 - 1) \cdot S_1^{*2} + (n_2 - 1) \cdot S_2^{*2}}{(n_1 - 1) + (n_2 - 1)}$$

eine erwartungstreue Schätzung für σ_D^2, die die beiden Stichprobenvarianzen mit dem Umfang der Stichprobenumfänge gewichtet. Dies führt zur Prüfstatistik

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_1-1) \cdot S_1^{*2} + (n_2-1) \cdot S_2^{*2}}{n_1+n_2-2}}}$$

des **Zweistichproben-t-Tests**. Für diese Prüfvariable kann man zeigen, dass sie bei Gleichheit der beiden Stichprobenvarianzen t-verteilt ist mit $\nu = n_1 + n_2 - 2$ Freiheitsgraden. Man verwirft dann H_0 zum Signifikanzniveau α, falls für die Prüfgröße die Bedingung $|t| > t_{n_1+n_2-2;1-\alpha/2}$ zutrifft.

Test auf Unabhängigkeit

Erwähnt sei noch, dass der für zwei nominalskalierte Merkmale X und Y als Zusammenhangsmaß verwendbare χ^2-Koeffizient als Testgröße beim Test der Hypothesen

$$H_0: \ X \text{ und } Y \text{ sind unabhängig} \quad \text{gegen} \quad H_1: \quad X \text{ und } Y \text{ sind abhängig}$$

herangezogen werden kann (**Unabhängigkeitstest**). Die Testvariable $T = \chi^2$ ist unter schwachen Voraussetzungen unter H_0 approximativ χ^2-verteilt mit $\nu := (k-1) \cdot (m-1)$ Freiheitsgraden. Die Nullhypothese H_0 wird dann mit einer Irrtumswahrscheinlichkeit α verworfen, wenn für die Realisation χ^2 der Testgröße gilt:

$$\chi^2 > \chi^2_{(k-1)\cdot(m-1);1-\alpha}.$$

Regressionsanalyse (einfaches Regressionsmodell)

Das **einfache lineare Regressionsmodell** ist definiert durch

$$y_i = \alpha + \beta x_i + u_i \qquad i = 1, \ldots, n,$$

wobei $(x_1, y_1), \ldots, (x_n, y_n)$ Datenpaare für zwei Merkmale X und Y sind und u_i die Ausprägung einer von Beobachtungsperiode zu Beobachtungsperiode variierenden Störvariablen U in der Beobachtungsperiode i. Die die Lage der Geraden $y = \alpha + \beta x$ determinierenden Parameter α und β heißen **Regressionskoeffizienten**. Für das Modell werden folgenden Annahmen getroffen: Modellannahmen

A1: Außer X werden keine weiteren exogenen Variablen zur Erklärung von Y benötigt.

A2: Die Parameter α und β sind konstant.

A3a: Die Störterme u_i sind Ausprägungen von Zufallsvariablen mit Erwartungswert 0 und Varianz σ^2.

A3b: Störvariablen aus unterschiedlichen Beobachtungsperioden sind unkorreliert.

A3c: Die Störvariablen sind normalverteilt.

A4: Die Werte der unabhängigen Variable X sind determiniert.

A5: Die Variable X ist nicht konstant für $i = 1, \ldots, n$ (Ausschluss eines trivialen Falls).

Ohne den Störterm u_i würden die Beobachtungsdaten (x_i, y_i) alle auf einer Geraden $y = \alpha + \beta x$ liegen (Regressionsgerade). Diese „wahre“ Gerade ist unbekannt, d. h. die sie determinierenden Regressionskoeffizienten α und β müssen anhand der Daten geschätzt werden. Die Regressionskoeffizienten der geschätzten Geraden Kleinst-Quadrat-Schätzung

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta} x$$

bestimmt man meist nach der **Methode der kleinsten Quadrate** (**KQ-Schätzung**). Bei dieser greift man auf die als **Residuen** bezeichneten Abweichungen

$$\widehat{u_i} = y_i - \widehat{y_i} = y_i - \widehat{\alpha} - \widehat{\beta} x_i \qquad i = 1, \ldots, n$$

zwischen dem Beobachtungswert y_i und dem Wert $\widehat{y_i}$ der Regressionsgeraden in der Beobachtungsperiode i zurück. Man wählt bei der KQ-Methode aus der Menge aller denkbaren Anpassungsgeraden diejenige Regressionsgerade $\widehat{R}$ aus, bei der die Summe der *quadrierten* Residuen $\widehat{u_i}^2$ bezüglich der beiden Geradenparameter minimal ist:

$$\sum_{i=1}^{n} \widehat{u_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 \to Min.$$

Die KQ-Schätzungen der Regressionskoeffizienten β und α errechnen sich nach KQ-Schätzungen

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}; \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta} \cdot \overline{x}.$$

Die Varianz der Störvariablen lässt sich anhand der Summe der quadrierten Residuen $\widehat{u}_i^2$ schätzen, die man noch durch $n-2$ dividiert:

$$\widehat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \widehat{u}_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2.$$

Für die KQ-Schätzfunktionen $\widehat{\beta}$, $\widehat{\alpha}$ und $\widehat{\sigma}^2$ lässt sich mit den getroffenen Modellannahmen ableiten, dass sie erwartungstreu sind:

$$E(\widehat{\beta}) = \beta; \quad E(\widehat{\alpha}) = \alpha; \quad E(\widehat{\sigma}^2) = \sigma^2.$$

Als Maß für die Anpassungsgüte eines bivariaten Datensatzes an eine Regressionsgerade wird das **Bestimmtheitsmaß** R^2 verwendet. Dieses Gütemaß setzt den durch die lineare Regression erklärten Varianzanteil $s_{\widehat{y}}^2$ ins Verhältnis zur Gesamtvariation s_y^2 der endogenen Variablen. Ausgangspunkt für die Herleitung von R^2 ist eine Zerlegung der Gesamtvarianz s_y^2 der abhängigen Variablen in zwei Komponenten:

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2}_{s_{\widehat{y}}^2} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{u}_i - \overline{\widehat{u}})^2}_{s_{\widehat{u}}^2}.$$

Dabei beinhaltet $s_{\widehat{y}}^2$ die durch den Regressionsansatz erklärte Varianz und $s_{\widehat{u}}^2$ die durch den Ansatz nicht erklärte Restvarianz. Bei Beachtung von $\overline{\widehat{u}} = 0$ und $\overline{\widehat{y}} = \overline{y}$ sowie $\widehat{u}_i = y_i - \widehat{y}_i$ kann man die beiden Komponenten auch wie folgt schreiben:

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{s_{\widehat{y}}^2} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{s_{\widehat{u}}^2}.$$

Das Anpassungsgütemaß R^2, für das $0 \leq R^2 \leq 1$ gilt, ist somit gegeben durch

$$R^2 = \frac{s_{\widehat{y}}^2}{s_y^2} = 1 - \frac{s_{\widehat{u}}^2}{s_y^2}.$$

Formeln für R^2

Wenn man die letzte der beiden obigen Varianzzerlegungen mit n erweitert und die resultierenden Summen SQ von quadrierten Abweichungen jeweils gemäß

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{SQ_{Total}} = \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{SQ_{Regression}} + \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{SQ_{Residual}}.$$

mit einem aussagekräftigen Index versieht, erhält man eine weitere Darstellung für R^2:

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}.$$

Erwähnt sei noch die für die praktische Berechnung von R^2 nützliche Formel

$$R^2 = \frac{\widehat{\beta} s_{xy}}{s_y^2} = \frac{(s_{xy})^2}{s_x^2 s_y^2} = r^2.$$

Regressionsanalyse (multiples Regressionsmodell)

Eine Verallgemeinerung des Modellansatzes mit nur *einer* erklärenden Variablen ist das **multiple lineare Regressionsmodell**

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i \qquad i = 1, \ldots, n$$

mit k erklärenden Variablen. Für das Modell gelten folgende Annahmen:

Modellannahmen

MA1: Alle k erklärenden Variablen liefern einen relevanten Erklärungsbeitrag.

MA2: Die $k+1$ Parameter β_0, $\beta_1, \beta_2, \ldots, \beta_k$ sind konstant.

MA3a: Die Störterme u_i des Modells sind Realisationen von Zufallsvariablen mit Erwartungswert 0 und fester Varianz σ^2.

MA3b: Störvariablen aus unterschiedlichen Beobachtungsperioden sind unkorreliert.

MA3c: Die Störvariablen sind normalverteilt.

MA4: Die Werte der k unabhängigen Variablen $X_1, X_2, \ldots, X_k$ sind determiniert.

MA5: Zwischen den k Regressoren existieren keine linearen Abhängigkeiten.

Die n Gleichungen des multiplen Regressionsmodells lassen sich auch mit Vektoren und Matrizen darstellen:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Wenn man die drei obigen Vektoren mit $\mathbf{y}$, $\boldsymbol{\beta}$ und $\mathbf{u}$ bezeichnet und die Matrix mit $\mathbf{X}$, kann man kürzer schreiben

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Zur Schätzung der Regressionskoeffizienten kann erneut die **Methode der kleinsten Quadrate** eingesetzt werden, bei der hier aus der Menge aller denkbaren Anpassungshyperebenen ($k > 2$) – im Falle k = 2 ist dies eine Ebene – diejenige ausgewählt wird, bei der die Summe der *quadrierten* Residuen $\widehat{u_i}^2$ bezüglich der Regressionskoeffizienten minimal ist. Die Minimierungsaufgabe hat hier die Gestalt

$$\sum_{i=1}^{n} \widehat{u_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik})^2 \rightarrow Min.$$

Bei Verwendung von Vektoren und Matrizen kann man äquivalent schreiben

$$\sum_{i=1}^{n} \widehat{u}_i^{\,2} = \widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \rightarrow Min.$$

Die im Sinne der KQ-Methode optimale Regressionshyperebene ist durch einen Vektor

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)'$$

definiert, der die KQ-Schätzungen $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ für die Regressionskoeffizienten zusammenfasst. Er errechnet sich aus der Datenmatrix $\mathbf{X}$ und dem Datenvektor $\mathbf{y}$ gemäß

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Grundzüge der Varianzanalyse

Mit der **Varianzanalyse** lassen sich Niveauunterschiede in mehr als zwei Teilpopulationen untersuchen. Man geht hier wieder von einem linearen Zusammenhang zwischen einer Einflussgröße X oder mehreren Einflussgrößen $X_1, X_2, ..., X_k$ (**Faktoren**) und einer zu erklärenden Variablen Y aus (Responsevariable). Letztere wird als stetig modelliert, während die Einflussgrößen *diskret* sind. Die Ausprägungen der Einflussgrößen heißen auch **Faktorstufen**. Wenn die Faktorstufen von vorneherein festgelegt sind, spricht man von einem **Modell der Varianzanalyse mit festen Effekten**, bei einer zufälligen Auswahl von einem **Modell der Varianzanalyse mit zufälligen Effekten**. Es wird unterschieden zwischen **einfaktorieller Varianzanalyse** (*eine* Einflussgröße) und **mehrfaktorieller Varianzanalyse** (*mehrere* Einflussgrößen).

Einfaktorielle Varianzanalyse

Beim *einfaktoriellen* Modell der Varianzanalyse geht man von einer Grundgesamtheit aus, für die eine Zufallsstichprobe des Umfangs n vorliegt. Die Stichprobe wird in s Teilmengen (Gruppen) des Umfangs n_i zerlegt ($i = 1, 2, \ldots, s; n_1 + n_2 + \ldots n_s = n$), die jeweils einer anderen Intensität (Faktorstufe) eines einzigen Einflussfaktors X ausgesetzt sind. Die Responsevariable Y ist in allen Teilstichproben unabhängig $N(\mu_i, \sigma^2)$-verteilt mit einem gruppenspezifischen Erwartungwert μ_i. Die Schwankungen der Responsevariablen innerhalb der Gruppen werden wie beim Regressionsmodell durch eine Störvariable U mit $E(U) = 0$ repräsentiert. Das Modell lautet also, wenn man die Stichprobenwerte als Ausprägungen von Zufallsvariablen interpretiert,

$$Y_{ik} = \mu_i + U_{ik} \qquad i = 1, \ldots, s; \quad k = 1, \ldots, n_i.$$

Zerlegt man den Erwartungswert μ_i der i-ten Gruppe noch in eine Basiskomponente μ und eine gruppenspezifische Komponente α_i, erhält man das **Modell der einfaktoriellen Varianzanalyse in Effektdarstellung**:

$$Y_{ik} = \mu + \alpha_i + U_{ik} \qquad i = 1, \ldots, s; \quad k = 1, \ldots, n_i.$$

Dabei ist $n_1 \cdot \alpha_1 + n_2 \cdot \alpha_2 + \ldots n_s \cdot \alpha_s = 0$. Die einfaktorielle Varianzanalyse ermöglicht anhand eines F-Tests auch eine Entscheidung darüber, ob die Veränderung von Faktorstufen einen signifikanten Einfluss auf den Erwartungswert der Responsevariablen hat. Man testet

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_s \quad \text{gegen} \quad H_1 : \mu_i \neq \mu_j \text{ für mind. ein } (i, j)$$

bzw. bei Zugrundelegung des Modells in Effektdarstellung

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_s = 0 \quad \text{gegen} \quad H_1 : \alpha_i \neq 0 \text{ und } \alpha_j \neq 0 \text{ für mind. ein } (i, j).$$

Bei der Konstruktion einer Teststatistik wird ausgenutzt, dass sich die Streuung der n Beobachtungen aus allen s Stichproben (Gesamtstreuung) zerlegen lässt in eine die Variabilität *zwischen* den Gruppen widerspiegelnde Komponente SQ_{zwischen} (Behandlungseffekt) und eine die Variation *innerhalb* der Stichproben repräsentierende Restkomponente SQ_{Residual} (Reststreuung). Es gilt also die Streuungszerlegungsformel

$$SQ_{\text{Total}} = SQ_{\text{zwischen}} + SQ_{\text{Residual}}.$$

Die Gesamtstreuung ist gegeben durch

$$SQ_{\text{Total}} = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_{..})^2$$

und für die beiden Komponenten gilt

$$SQ_{\text{zwischen}} := \sum_{i=1}^{s} n_i \cdot (\overline{y}_{i.} - \overline{y}_{..})^2; \qquad SQ_{\text{Residual}} := \sum_{i=1}^{s} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_{i.})^2.$$

Um zu testen, ob die Variation von Faktorstufen einen signifikanten Einfluss auf den Erwartungswert der Responsevariablen hat, verwendet man die Teststatistik

$$F := \frac{\frac{1}{s-1} \cdot SQ_{\text{zwischen}}}{\frac{1}{n-s} \cdot SQ_{\text{Residual}}} = \frac{n-s}{s-1} \cdot \frac{SQ_{\text{zwischen}}}{SQ_{\text{Residual}}}.$$

Dieser Quotient folgt unter H_0 einer **F-Verteilung** mit $s-1$ und $n-s$ Freiheitsgraden, weil unter der hier getroffenen Normalverteilungsannahme $SQ_{\text{zwischen}} \sim \chi^2_{s-1}$ und $SQ_{\text{Residual}} \sim \chi^2_{n-s}$ gilt. Die Alternativhypothese H_1 wird dann als statistisch gesichert angesehen mit Irrtumswahrscheinlichkeit α, wenn der genannte Quotient das $(1-\alpha)$-Quantil $F_{s-1;n-s;1-\alpha}$ der F-Verteilung mit $s-1$ und $n-s$ Freiheitsgraden überschreitet.

Wenn man den Einfluss von *zwei* Einflussgrößen X_1 und X_2 mit s resp. r Faktorstufen auf eine Responsevariable Y betrachtet, erhält man eine Modelldarstellung, die sich auf $s \cdot r$ Faktorstufenkombinationen bezieht: Zweifaktorielle Varianzanalyse

$$Y_{ijk} = \mu_{ij} + U_{ijk} \quad i = 1, \ldots, s; \quad j = 1, \ldots, r; \quad k = 1, \ldots, n_{ij},$$

wobei die Störvariablen als unabhängig identisch $N(0; \sigma^2)$-verteilt spezifiziert sind. Zerlegt man die Erwartungswerte μ_{ij} der Responsevariablen in den $s \cdot r$ Gruppen wieder additiv in einen für alle Gruppen identischen Basisanteil μ und in faktorstufenspezifische Komponenten α_i (Effekt der i-ten Stufe des Faktors X_1) sowie β_j (Effekt der j-ten Stufe des Faktors X_2) und berücksichtigt bei der Modellformulierung noch einen mit $(\alpha\beta)_{ij}$ bezeichneten möglichen Wechselwirkungseffekt zwischen der i-ten Stufe von X_1 und der j-ten Stufe von X_2, erhält man das **Modell der zweifaktoriellen Varianzanalyse in Effektdarstellung**:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + U_{ijk} \quad i = 1, \ldots, s; \;\; j = 1, \ldots, r; \;\; k = 1, \ldots, n_{ij}.$$

Effekte auf die Responsevariable Y, die durch die Veränderung von Stufen von Faktor X_1 oder von Faktor X_2 hervorgerufen werden, heißen **Haupteffekte**. Wirkungen auf Y, die durch Interaktion der beiden Faktoren induziert werden, nennt man **Wechselwirkungseffekte** oder **Interaktionseffekte**.

Anmerkungen und Ergänzungen für Studierende im BSc „Psychologie“

Ergänzungen zur multiplen Regressionsanalyse

Regressoren mit kategorialem Skalenniveau erfordern eine spezifische Behandlung. Kodierte Merkmalsausprägungen – z. B. ‘ledig’=1, ‘verheiratet’=2, ‘geschieden’=3, ‘verwitwet’=4 – können nicht wie reelle Zahlen in die Berechnung von Parameterschätzungen einbezogen werden, da den Kodierungen nicht notwendigerweise eine Ordnung zugrunde liegt und Abstände bei ordinalen Merkmalen nicht definiert sind. Um diesem Problem zu begegnen, werden kategoriale Regressoren umkodiert. Hierfür gibt es die Möglichkeit der Dummy- oder Effektkodierung. Bei beiden Varianten wird ein kategorialer Regressor mit k möglichen Merkmalsausprägungen in $k-1$ neue Regressoren (Dummys) umgewandelt. Eine der Originalkategorien (Merkmalsausprägungen) wird als sogenannte **Referenzkategorie** ausgewählt. Nach deren Wahl ergeben sich die Dummys X_i $(i = 1, \ldots, k)$ bei der **Dummykodierung** gemäß

$$x_i = \begin{cases} 1 & \text{falls Kategorie } i \text{ vorliegt,} \\ 0 & \text{sonst} \end{cases}$$

und bei der **Effektkodierung** nach

$$x_i = \begin{cases} 1 & \text{falls Kategorie } i \text{ vorliegt,} \\ -1 & \text{falls Kategorie } j \text{ vorliegt} \quad (i \neq j), \\ 0 & \text{sonst.} \end{cases}$$

Ergänzungen zur Varianzanalyse

Im Kurs Statistik II werden die Ausprägungen der Störvariablen bei der Regressions- und Varianzanalyse mit ϵ und nicht mit u bezeichnet. Bei der Varianzanalyse wird für die Restkomponente SQ_{Residual}, welche die Variation *innerhalb* der Stichproben widerspiegelt, im Kurs Statistik II auch die Bezeichnung $SQ_{innerhalb}$ verwendet. Ferner wird für die durch die Anzahl der Freiheitsgrade (kurz: df; *d*egrees of *f*reedom) dividierten Streuungskomponenten SQ die Abkürzung MQ herangezogen. Die Prüfstatistik F hat also mit den vorstehend genannten Notationen im Kurs Statistik II im Falle der einfaktoriellen Varianzanalyse die Gestalt

$$F = \frac{\frac{1}{s-1} \cdot SQ_{\text{zwischen}}}{\frac{1}{n-s} \cdot SQ_{\text{Residual}}} = \frac{MQ_{\text{zwischen}}}{MQ_{\text{Residual}}}.$$

Der Behandlungseffekt SQ_{zwischen} wird im Kurs Statistik II bei mehrfaktoriellen Designs i. a. nach dem jeweiligen Faktor benannt, also z.B. SQ_A und SQ_B bei einem zweifaktoriellen Modell mit den Faktoren A und B, und die Anzahl der Faktorstufen mit a resp. mit b. Es wird ferner im Kurs Statistik II bei der Behandlung der zweifaktoriellen Varianzanalyse angenommen, dass für jede Faktorstufenkombination $(i;j)$ genau r Beobachtungen vorliegen. Für die Durchführung und Ergebnisdarstellung einer zweifaktoriellen Varianzanalyse wird dann das folgende Schema verwendet, bei der $N = a \cdot b \cdot r$ die Gesamtzahl der Beobachtungen bezeichnet:

Ursache	SQ	df	MQ	F
Faktor A	SQ_A	$a-1$	MQ_A	F_A
Faktor B	SQ_B	$b-1$	MQ_B	F_B
Wechselwirkung $A \times B$	$SQ_{A\times B}$	$(a-1)(b-1)$	$MQ_{A\times B}$	$F_{A\times B}$
Fehler	$SQ_{Residual}$	$N-ab$ $= ab(r-1)$	$MQ_{Residual}$	
Total	SQ_{Total}	$N-1$		

3 Matrizen, statistische Tabellen und Konzeptpapier

Grundzüge der Matrizenrechnung

Spalten- und Zeilenvektoren

Wenn man ein n-Tupel von reellen Zahlen vertikal anordnet, erhält man einen **Spaltenvektor**, der i. a. mit einem fett gesetzten lateinischen oder griechischen Kleinbuchstaben abgekürzt wird. Ordnet man das n-Tupel horizontal an, resultiert ein **Zeilenvektor**. Die Überführung eines Spaltenvektors in einen Zeilenvektor wird als *Transponieren* des Vektors bezeichnet und durch einen hochgestellten Strich gekennzeichnet:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)' = \mathbf{x}'.$$

Spezielle Vektoren sind der nur aus Nullen bestehende **Nullvektor 0** und der nur aus Einsen bestehende **Einsvektor 1**. Reelle Zahlen, die ja die Elemente eines Vektors konstituieren, heißen **Skalare**.

Bildung von Matrizen

Hat man nicht nur einen, sondern k Datensätze $\mathbf{x}_j = (x_{1j}, ..., x_{nj})'$ $(j = 1, 2, ..., k)$ des Umfangs n und stellt man die Elemente der k Spaltenvektoren nebeneinander, erhält man ein als **Matrix** bezeichnetes rechteckiges Schema mit Tabellenstruktur. Matrizen werden i. a. mit fetten lateinischen oder griechischen Großbuchstaben abgekürzt:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{nk} \end{pmatrix} = (x_{ij})_{i=1,..,n;j=1,...,k}.$$

Eine Matrix mit n Zeilen und k Spalten heißt (n x k)-Matrix. Man verwendet auch die kürzere Schreibweise $\mathbf{X} = (x_{ij})$, wenn sich der Laufbereich der Indizes i (Anzahl der Zeilen) und j (Anzahl der Spalten) aus dem Kontext erschließt.

Spezialfälle

Vektoren lassen sich als spezielle Matrizen interpretieren – ein Zeilenvektor lässt sich als Matrix mit nur einer Zeile und ein Spaltenvektor als Matrix mit nur einer Spalte interpretieren. Eine Matrix, deren Elemente alle Nullen sind, heißt **Nullmatrix**. Ein weiterer Spezialfall ist eine **quadratische Matrix** (gleiche Zeilen- und Spaltenzahl).

Sind bei einer quadratischen Matrix alle Elemente x_{ij} mit $i \neq j$ Null, spricht man von einer **Diagonalmatrix**. Deren Elemente $x_{11}, x_{22}, ..., x_{nn}$ konstituieren die **Hauptdiagonale**. Ein Sonderfall einer Diagonalmatrix ist die i. a. mit $\mathbf{I}$ oder – bei Ausweis der Dimension – mit $\mathbf{I}_n$ abgekürzte **Einheitsmatrix**. Für diese ist kennzeichnend, dass die Elemente auf der Hauptdiagonalen alle den Wert 1 haben.

Auch Matrizen lassen sich transponieren. Die zu einer Matrix $\mathbf{X}$ gehörende **transponierte Matrix $\mathbf{X}'$** entsteht durch Vertauschen der Zeilen und Spalten von $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \overset{Transponieren}{\longrightarrow} \quad \mathbf{X}' = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{i1} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{i2} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{ik} & \dots & x_{nk} \end{pmatrix}.$$

Eine Matrix $\mathbf{X}$ mit der Eigenschaft $\mathbf{X} = \mathbf{X}'$ heißt *symmetrisch*.

Die Multiplikation einer Matrix mit einer reellen Zahl λ (lies: *lambda*) erfolgt, indem man jedes Element einer Matrix $\mathbf{X} = (x_{ij})$ einzeln mit dem Skalar λ multipliziert:

$$\lambda \cdot \mathbf{X} = \lambda \cdot (x_{ij}) = (\lambda \cdot x_{ij}).$$

Addition von Matrizen

Bei der Addition von Matrizen $\mathbf{A} = (a_{ij})$ und $\mathbf{B} = (b_{ij})$ gleicher Dimension werden die an gleicher Position stehenden Elemente addiert, d. h. es ist

$$\mathbf{A} + \mathbf{B} = \mathbf{C} = (c_{ij}) \qquad \text{mit} \qquad c_{ij} = a_{ij} + b_{ij}.$$

Für Matrizen ungleicher Dimension ist die Addition nicht erklärt. Auch die Multiplikation von Matrizen ist nur unter bestimmten Voraussetzungen möglich. Das Produkt zweier Matrizen $\mathbf{A}$ und $\mathbf{B}$ ist erklärt, wenn die Anzahl der Spalten von $\mathbf{A}$ mit der Anzahl der Zeilen von $\mathbf{B}$ übereinstimmt. Hat etwa die Matrix $\mathbf{A}$ die Dimension $(n \times k)$ und $\mathbf{B}$ die Dimension $(k \times m)$, so ist die Matrix $\mathbf{C} := \mathbf{A} \cdot \mathbf{B}$ von der Dimension $(n \times m)$:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & \dots & b_{1l} & \dots & b_{1m} \\ b_{21} & \dots & b_{2l} & \dots & b_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{k1} & \dots & b_{kl} & \dots & b_{km} \end{pmatrix} = \begin{pmatrix} c_{11} & \dots & c_{1l} & \dots & c_{1m} \\ \vdots & \vdots & \vdots & \vdots & \\ c_{i1} & \dots & c_{il} & \dots & c_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{n1} & \dots & c_{nl} & \dots & c_{nm} \end{pmatrix}$$

Produkt zweier Matrizen

Das vorstehend durch Rasterung betonte Element c_{il} der $(n \times m)$-Produktmatrix $\mathbf{C}$ ergibt sich, indem man die ebenfalls in der obigen Gleichung gerastert dargestellten k Elemente der i-ten Zeile von $\mathbf{A}$ $(i = 1, ..., n)$ und die k Elemente der l-ten Spalte von $\mathbf{B}$ $(l = 1, \dots, m)$ gliedweise miteinander multipliziert und aufsummiert:

$$\underbrace{\mathbf{A}}_{n \times k} = (a_{ij}), \quad \underbrace{\mathbf{B}}_{k \times m} = (b_{jl}) \quad \Rightarrow \quad \mathbf{A} \cdot \mathbf{B} = \underbrace{\mathbf{C}}_{n \times m} = (c_{il}) \quad \text{mit} \quad c_{il} = \sum_{j=1}^{k} a_{ij} \cdot b_{jl}.$$

Inversion von Matrizen

Wenn eine quadratische Matrix $\mathbf{B}$ die Eigenschaft hat, dass das Produkt $\mathbf{A} \cdot \mathbf{B}$ die Einheitsmatrix ist, nennt man sie die **Inverse** zur Matrix $\mathbf{A}$ und schreibt $\mathbf{A}^{-1}$ der Matrix $\mathbf{A}$). Für die Inverse $\mathbf{A}^{-1}$ einer quadratischen Matrix $\mathbf{A}$ ist neben $\mathbf{A} \cdot \mathbf{A}^{-1}$ stets auch $\mathbf{A}^{-1} \cdot \mathbf{A}$ erklärt und es gilt $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}$.

Verteilungsfunktion der Binomialverteilung

Es sei $X \sim B(n,p)$ eine mit Parametern n und p binomialverteilte Zufallsvariable. Deren Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ ist durch

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

und die **Verteilungsfunktion** $F(x) = P(X \leq x)$ durch

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \qquad x = 0, 1, \ldots, n.$$

gegeben. Um das Verhalten von X vollständig zu charakterisieren, benötigt man nur eine der beiden obigen Funktionen.

In der nachstehenden Tabelle sind Werte $F(x)$ der Verteilungsfunktion einer $B(n,p)$-verteilten Zufallsvariablen X für $n = 1, 2, \ldots, 8$ und $p = 0,05, 0,10, \ldots, 0,50$ zusammengestellt. Man entnimmt der Tabelle z. B., dass $F(x)$ im Falle $n = 7$ und $p = 0,40$ für $x = 3$ den Wert $F(3) = 0,7102$ annimmt. Dieser Wert entspricht der Summe $f(0)$, $f(1)$, $f(2)$, $f(3)$ aller Werte der Wahrscheinlichkeitsfunktion bis zur Stelle $x = 3$.

n	x	p=0,05	p=0,10	p=0,15	p=0,20	p=0,25	p=0,30	p=0,35	p=0,40	p=0,45	p=0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
1	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
2	1	0,9975	0,9900	0,9775	0,9600	0,9375	0,9100	0,8775	0,8400	0,7975	0,7500
2	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
3	1	0,9928	0,9720	0,9393	0,8960	0,8438	0,7840	0,7183	0,6480	0,5748	0,5000
3	2	0,9999	0,9990	0,9966	0,9920	0,9844	0,9730	0,9571	0,9360	0,9089	0,8750
3	3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
4	1	0,9860	0,9477	0,8905	0,8192	0,7383	0,6517	0,5630	0,4752	0,3910	0,3125
4	2	0,9995	0,9963	0,9880	0,9728	0,9492	0,9163	0,8735	0,8208	0,7585	0,6875
4	3	1,0000	0,9999	0,9995	0,9984	0,9961	0,9919	0,9850	0,9744	0,9590	0,9375
4	4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
5	1	0,9774	0,9185	0,8352	0,7373	0,6328	0,5282	0,4284	0,3370	0,2562	0,1875
5	2	0,9988	0,9914	0,9734	0,9421	0,8965	0,8369	0,7648	0,6826	0,5931	0,5000
5	3	1,0000	0,9995	0,9978	0,9933	0,9844	0,9692	0,9460	0,9130	0,8688	0,8125
5	4	1,0000	1,0000	0,9999	0,9997	0,9990	0,9976	0,9947	0,9898	0,9815	0,9688
5	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
6	1	0,9672	0,8857	0,7765	0,6554	0,5339	0,4202	0,3191	0,2333	0,1636	0,1094
6	2	0,9978	0,9842	0,9527	0,9011	0,8306	0,7443	0,6471	0,5443	0,4415	0,3438
6	3	0,9999	0,9987	0,9941	0,9830	0,9624	0,9295	0,8826	0,8208	0,7447	0,6563
6	4	1,0000	0,9999	0,9996	0,9984	0,9954	0,9891	0,9777	0,9590	0,9308	0,8906
6	5	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9982	0,9959	0,9917	0,9844
6	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
7	1	0,9556	0,8503	0,7166	0,5767	0,4449	0,3294	0,2338	0,1586	0,1024	0,0625
7	2	0,9962	0,9743	0,9262	0,8520	0,7564	0,6471	0,5323	0,4199	0,3164	0,2266
7	3	0,9998	0,9973	0,9879	0,9667	0,9294	0,8740	0,8002	0,7102	0,6083	0,5000
7	4	1,0000	0,9998	0,9988	0,9953	0,9871	0,9712	0,9444	0,9037	0,8471	0,7734
7	5	1,0000	1,0000	0,9999	0,9996	0,9987	0,9962	0,9910	0,9812	0,9643	0,9375
7	6	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9994	0,9984	0,9963	0,9922
7	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
8	1	0,9428	0,8131	0,6572	0,5033	0,3671	0,2553	0,1691	0,1064	0,0632	0,0352
8	2	0,9942	0,9619	0,8948	0,7969	0,6785	0,5518	0,4278	0,3154	0,2201	0,1445
8	3	0,9996	0,9950	0,9786	0,9437	0,8862	0,8059	0,7064	0,5941	0,4770	0,3633
8	4	1,0000	0,9996	0,9971	0,9896	0,9727	0,9420	0,8939	0,8263	0,7396	0,6367
8	5	1,0000	1,0000	0,9998	0,9988	0,9958	0,9887	0,9747	0,9502	0,9115	0,8555
8	6	1,0000	1,0000	1,0000	0,9999	0,9996	0,9987	0,9964	0,9915	0,9819	0,9648
8	7	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9993	0,9983	0,9961
8	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Verteilungsfunktion $F(x)$ der Binomialverteilung ($n = 1$ bis $n = 8$)

Verteilungsfunktion und Quantile der Standardnormalverteilung

Ist X eine mit Erwartungswert μ und Varianz σ^2 normalverteilte Zufallsvariable, also $X \sim N(\mu, \sigma^2)$, so lässt sie sich anhand ihrer Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

oder anhand ihrer Verteilungsfunktion $F(x) = P(X \leq x)$ charakterisieren, wobei die erste Ableitung $F'(x)$ der Verteilungsfunktion und die Dichtefunktion $f(x)$ über die Beziehung $F'(x) = f(x)$ verknüpft sind.

Man kann jede normalverteilte Zufallsvariable X über die Transformation $Z := \frac{X-\mu}{\sigma}$ in die **Standardnormalverteilung** überführen (Normalverteilung mit Erwartungswert 0 und Varianz 1). Daher genügt es, Werte der Verteilungsfunktion der Standardnormalverteilung zu tabellieren. Für diese Funktion hat sich die Bezeichnung $\Phi(z)$ etabliert und für die Dichtefunktion $\Phi'(z)$ der Standardnormalverteilung die Bezeichnung $\phi(z)$. Zwischen der Verteilungsfunktion $F(x)$ einer $N(\mu, \sigma^2)$-verteilten Zufallsvariablen und der Verteilungsfunktion $\Phi(z)$ der standardisierten Variablen Z besteht die Beziehung

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z).$$

In der nebenstehenden Tabelle (obere Tabelle) sind für den Bereich von $z = 0,00$ bis $z = 3,99$ Werte der Verteilungsfunktion $\Phi(z)$ auf vier Dezimalstellen genau wiedergegeben. Dabei ist die letzte Dezimalstelle der Werte z im Tabellenkopf ausgewiesen. Aufgrund der Symmetriebeziehung

$$\Phi(z) = 1 - \Phi(-z)$$

reicht es Werte $\Phi(z)$ für nicht-negative z zu tabellieren. Für $z = -1,65$ gilt z. B. $\Phi(-1,65) = 1 - \Phi(1,65) = 0,0495$.

Ein **p-Quantil** z_p der Standardnormalverteilung ist durch $\Phi(z_p) = p$ $(0 < p < 1)$ definiert und markiert den Punkt auf der z-Achse, bis zu dem die Fläche unter der Dichte gerade p ist. Die nebenstehende Tabelle (unten) weist einige ausgewählte p-Quantile aus. Dabei ist $p \geq 0,5$. Quantile für $p < 0,5$ erhält man über die Beziehung $z_p = -z_{1-p}$, die sich aus der Symmetrie von Dichte- und Verteilungsfunktion bezüglich $z = 0$ ergibt. Mit $z_{0,95} = 1,6449$ gilt also z. B. $z_{0,05} = -1,6449$.

z	0	1	2	3	4	5	6	7	8	9
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	0,7704	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	0,8079	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	0,9956	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	0,9994	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996
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
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Werte der Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung

p	0,500	0,600	0,700	0,800	0,900	0,950	0,975	0,990	0,995	0,999
z_p	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0902

Quantile z_p der Standardnormalverteilung

Quantile der χ^2-Verteilung

In der folgenden Tabelle sind Quantile $\chi^2_{n;p}$ der χ^2-Verteilung mit n Freiheitsgraden für $n = 1$ bis $n = 40$ und ausgewählte Werte p zusammengestellt. Man entnimmt der Tabelle z. B., dass das $0,95$-Quantil der χ^2-Verteilung mit $n = 8$ Freiheitsgraden den Wert $\chi^2_{8;0,95} = 15,507$ besitzt.

n	p =0,01	p =0,025	p =0,05	p =0,1	p =0,9	p =0,95	p =0,975	p =0,99	p =0,995
1	-	-	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	16,919	19,023	21,666	23,589
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688	29,819
14	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,335
30	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
31	15,655	17,539	19,281	21,434	41,422	44,985	48,232	52,191	55,002
32	16,362	18,291	20,072	22,271	42,585	46,194	49,480	53,486	56,328
33	17,073	19,047	20,867	23,110	43,745	47,400	50,725	54,775	57,648
34	17,789	19,806	21,664	23,952	44,903	48,602	51,966	56,061	58,964
35	18,509	20,569	22,465	24,797	46,059	49,802	53,203	57,342	60,275
36	19,233	21,336	23,269	25,643	47,212	50,998	54,437	58,619	61,581
37	19,960	22,106	24,075	26,492	48,363	52,192	55,668	59,893	62,883
38	20,691	22,878	24,884	27,343	49,513	53,384	56,895	61,162	64,181
39	21,426	23,654	25,695	28,196	50,660	54,572	58,120	62,428	65,475
40	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766

Quantile der Chi-Quadrat-Verteilung

Quantile der t-Verteilung

Nachstehend sind **Quantile** $t_{n;p}$ der t-Verteilung mit n Freiheitsgraden für $n = 1$ bis $n = 40$ und ausgewählte Werte p zusammengestellt. Aus der Tabelle geht z. B. hervor, dass das $0,975$-Quantil der t-Verteilung mit $n = 8$ Freiheitsgraden den Wert $t_{8;0,975} = 2,306$ besitzt. Quantile einer t-Verteilung mit größerer Anzahl von Freiheitsgraden lassen sich gut durch die entsprechenden Quantile z_p der Standardnormalverteilung approximieren.

n	0,800	0,850	0,900	0,950	0,975	0,990	0,995
1	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,979	1,250	1,638	2,353	3,182	4,541	5,841
4	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,870	1,080	1,350	1,771	2,160	2,650	3,012
14	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,854	1,055	1,310	1,697	2,042	2,457	2,750
31	0,853	1,054	1,310	1,696	2,040	2,453	2,744
32	0,853	1,054	1,309	1,694	2,037	2,449	2,739
33	0,853	1,053	1,308	1,692	2,035	2,445	2,733
34	0,852	1,053	1,307	1,691	2,032	2,441	2,728
35	0,852	1,052	1,306	1,690	2,030	2,438	2,724
36	0,852	1,052	1,306	1,688	2,028	2,435	2,720
37	0,851	1,051	1,305	1,687	2,026	2,431	2,715
38	0,851	1,051	1,304	1,686	2,024	2,429	2,712
39	0,851	1,050	1,304	1,685	2,023	2,426	2,708
40	0,851	1,050	1,303	1,684	2,021	2,423	2,705

Quantile der t-Verteilung

Quantile der F-Verteilung

Die folgende Tabelle weist **Quantile** $F_{m;n;p}$ einer F-Verteilung mit m und n Freiheitsgraden für $p = 0,95$ aus. Die Freiheitsgrade für m liegen im Bereich von 1 bis 10, die von n im Bereich von 1 bis 100. Der Tabelle entnimmt man z. B., dass für das $0,95$-Quantil der F-Verteilung mit $m = 5$ und $n = 10$ Freiheitsgraden $F_{5;10;0,95} = 3,33$ gilt.

	m									
n	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,14	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

Quantile der F-Verteilung ($p = 0,95$, $m = 1$ bis $m = 10$)

Konzeptpapier (keine Bewertung), Blatt 1

Konzeptpapier (keine Bewertung), Blatt 2

Konzeptpapier (keine Bewertung), Blatt 3

Konzeptpapier (keine Bewertung), Blatt 4

Bei weiterem Bedarf an Konzeptpapier wenden Sie sich bitte an die Klausuraufsicht.