Technische Universität Braunschweig

Formelsammlung zur Statistik
Wintersemester 2023/24

Institut für Mathematische Stochastik

Dr. Frank Palkowski
Prof. Dr. Jens-Peter Kreiß

26. Oktober 2023

Inhaltsverzeichnis

1 Deskriptive Statistik univariater Daten

- X heißt diskrete Variable, wenn X nur endlich (oder abzählbar unendlich viele) Werte annehmen kann.
- X heißt stetige Variable, wenn X alle möglichen Werte eines Intervalls annehmen kann.
- Nominal skalierte Variablen: Die Werte der Variablen sind lediglich Bezeichnungen. Es ist keine Reihung möglich. Z.B.: Farben, Antworten, Geschlecht.
- Ordinal skalierte Variablen: Die Werte der Variablen lassen sich sinnvoll anordnen. Z.B.: Zensuren.
- Metrisch skalierte Variablen: Die Reihenfolge und die Abstände der Werte der Variablen sind aussagekräftig. Z.B. Gehälter, Zeitdauern.

Für eine diskrete Variable X mit den möglichen Werten $a_1, \ldots, a_K$ liege eine Stichprobe $x_1, \ldots, x_n$ vom Umfang n vor.

Absolute Häufigkeiten: $h_k = |\{j \in \{1, \ldots, n\} : x_j = a_k\}|, k = 1, \ldots, K.$
h_k bezeichnet die Anzahl der Elemente der Stichprobe, die den Wert a_k annehmen.

Relative Häufigkeiten: $r_k = \frac{h_k}{n}, k = 1, \ldots, K.$
r_k bezeichnet die Anzahl der Elemente der Stichprobe, die den Wert a_k annehmen dividiert durch den Stichprobenumfang n.

Tabelle 1: Klassifizierte oder gruppierte Daten

Klassen − Nr.	Klasse	Klassenrepräsentant	Häufigkeit	relative Häufigkeit
1	$[a_0, a_1)$	x_1^*	h_1	r_1
2	$[a_1, a_2)$	x_2^*	h_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$[a_{k-1}, a_k)$	x_k^*	h_k	r_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	$[a_{K-1}, a_K)$	x_K^*	h_K	r_K

Als graphische Darstellung für gruppierte Stichproben bietet sich ein **flächentreues Histogramm** an. Die Erstellung eines flächentreuen Histogramms wird ausführlich in der Vorlesung besprochen. Wichtig ist, dass die **Flächen** des Histogramms den relativen Häufigkeiten entsprechen!

empirische Verteilungsfunktion F_n:

$$F_n(x) \quad := \quad \frac{1}{n}|\{j \in \{1,\ldots,n\} : x_j \leq x\}| = \frac{1}{n}\sum_{x_j \leq x} h_j = \sum_{x_j \leq x} r_j$$

: Anzahl der Beobachtungen x_j mit $x_j \leq x$ dividiert durch n
bzw. relative H. der Beobachtungen, die kleiner als oder gleich x sind.

1.1 Lageparameter

Mittelwert

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n} \tag{1.1}$$

Median:
Er berechnet sich wie folgt aus $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ der sog. **geordneten Stichprobe**:

$$\tilde{x} = Median(x_1, \ldots x_n) = \begin{cases} x_{(\frac{n+1}{2})} & , \ n \text{ ungerade} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & , \ n \text{ gerade.} \end{cases} \tag{1.2}$$

Der Median ist im Gegensatz zum Mittelwert robust gegen **Ausreißer**.

Über einen sog. **getrimmten Mittelwert** erreicht man ebenfalls eine Robustheit. Beim getrimmten Mittelwert werden jeweils die $\alpha \cdot 100\%$ kleinsten und größten Werte bei der Mittelwertbildung nicht berücksichtigt.

Lineare Transformation der Daten
$x_1, ..., x_n$ mit Mittelwert $\overline{x}$ und Median $\tilde{x}$.
Nun werden die Daten gemäß einer **linearen** Abbildung transformiert, d.h. wir erhalten neue Daten $y_1, \ldots, y_n$ wie folgt:

$$y_1 = a\,x_1 + b, y_2 = a\,x_2 + b, ..., y_n = a\,x_n + b, a \neq 0\,.$$

Dann gilt:

$$\overline{y} = a \cdot \overline{x} + b \text{ und } \tilde{y} = a \cdot \tilde{x} + b\,.$$

Empirische Quantile
Für jedes p mit $0 < p < 1$ ist das **empirische p*100%-Quantil** x_p der

Stichprobe $x_1, ..., x_n$ der $n \cdot p$ kleinste Wert der Stichprobe.
D.h.: Finde zu gegebenem p die natürliche Zahl k mit

$$n \cdot p \leq k < n \cdot p + 1 \tag{1.3}$$

und setzte: $x_p = x_{(k)}$ für $p \neq 0,5$ und $x_{0,5} = \tilde{x} = Median(x_1, \dots x_n)$.

$x_{0,25}$ heißt auch **unteres Quartil** und $x_{0,75}$ heißt **oberes Quartil**.

Geometrisches Mittel:

$$\overline{x}_{geom} = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}\,. \tag{1.4}$$

Es wird bei relativen Änderungen wie der Berechnung durchschnittlicher Wachstumsraten verwendet.

Harmonisches Mittel:

$$\overline{x}_{harm} = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i}} \tag{1.5}$$

Mittelwert bei gruppierten Daten
Beachte Tabelle 1. Dann setzt man.

$$\overline{x}_{gruppiert} = \sum_{i=1}^{K} x_i^* \cdot r_i\,. \tag{1.6}$$

Median bei gruppierten Daten
Beachte Tabelle 1. Der Median der klassifizierten Daten ist der Klassenrepräsentant derjenigen Klasse, in der der Median liegt.

Modus ist derjenige Wert in der Stichprobe, der am häufigsten auftritt. Natürlich braucht dieser Wert nicht eindeutig bestimmt zu sein. Diesen Wert kann man im Gegensatz zu den anderen Lageparametern auch bei nominalen Daten bestimmen.

1.2 5-Punkte Zusammenfassung

Es liege Stichprobe $x_1, \dots, x_n$ einer metrisch skalierten Variablen vor.
5-Punkte Zusammenfassung:

$$x_{(1)}\ ,\ x_{0,25}\ ,\ Median\ ,\ x_{0,75}\ ,\ x_{(n)}\,. \tag{1.7}$$

Die graphische Darstellung der 5-Punkte Zusammenfassung nennt man **Boxplot**.

1.3 Streuungsmaße

Empirische Varianz

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2}{n-1} \tag{1.8}$$

Empirische Streuung oder **Standardabweichung**

$$s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2}{n-1}} \tag{1.9}$$

Variationskoeffizient

$$V_x = \frac{s_x}{\overline{x}} \tag{1.10}$$

Spannweite

$$x_{(n)} - x_{(1)}$$

bzw. besser (unter Robustheitsaspekten) der **Interquartilsabstand**

$$x_{0,75} - x_{0,25}\,. \tag{1.11}$$

(Relative) mittlere absolute Differenz:

$$MAD = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \tag{1.12}$$

$$RMAD = \frac{1}{n^2 \overline{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \tag{1.13}$$

Es gilt:

$$RMAD = 2 \cdot \left(2 \cdot \frac{1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \ldots + n \cdot x_{(n)}}{n \cdot (x_1 + \ldots + x_n)} - \frac{n+1}{n} \right) = 2 \left(\frac{2}{n^2 \overline{x}} \cdot \sum_{i=1}^{n} i x_{(i)} - \frac{n+1}{n} \right)$$

Lineare Transformation der Daten:
Gegeben seien Daten $x_1, ..., x_n$ mit Mittelwert $\overline{x}$. Nun werden die Daten gemäß einer linearen Abbildung transformiert, d.h. wir erhalten neue Daten

$$y_1 = a \cdot x_1 + b, y_2 = a \cdot x_2 + b, ..., y_n = a \cdot x_n + b, a \neq 0\,.$$

Dann gilt:

$$s_y^2 = a^2 \cdot s_x^2 \text{ und } s_y = |a| \cdot s_x \,. \text{ Falls } b = 0 \text{ gilt } V_y = V_x.$$

Berechnung der Streuung bei gruppierten Daten
Betrachte Tabelle 2.1. Dann berechnet sich die empirische Varianz wie folgt.

$$s_{x,gruppiert}^2 = \sum_{i=1}^{K} (x_i^* - \overline{x}_{gruppiert})^2 \cdot r_i \tag{1.14}$$

Die empirische Streuung aus gruppierten Daten ergibt sich zu:

$$s_{x,gruppiert} = \sqrt{\sum_{i=1}^{K} (x_i^* - \overline{x}_{gruppiert})^2 \cdot r_i} \tag{1.15}$$

Dispersion(sindex) nach Hammond und Householder:

$$DI = \frac{K}{K-1} \sum_{i=1}^{K} r_i(1 - r_i) = \frac{K}{K-1}(1 - \sum_{i=1}^{K} r_i^2) \tag{1.16}$$

Faustregel: $DI < 0,8$ bedeutet geringe Dispersion, $DI > 0,9$ starke Dispersion.

Evenness bzw. normierte Shannon-Wiener-Entropie:

$$E = \frac{1}{\log K} S = -\frac{1}{\log K} \sum_{i=1}^{K} r_i \log r_i \tag{1.17}$$

Diese letzten beiden Kennzahlen kann man im Gegensatz zu den anderen Streuungsparametern auch bei nominalen Daten bestimmen. Es gilt, dass diese beiden Streuungsmaße wegen der Normierungskonstanten $\frac{K}{K-1}$ bzw. $\frac{1}{\log K}$ nur Werte zwischen 0 und 1, jeweils einschließlich, annehmen können. Lässt man diese Normierungskonstanten weg, können bei DI Werte zwischen 0 und $\frac{K-1}{K}$, also kleiner als 1, und bei S Werte zwischen 0 und $\log K$, also auch größer als 1 angenommen werden.

1.4 Schiefe und Kurtosis

Quartilskoeffizient der Schiefe:

$$\frac{(x_{0,75} - Median) - (Median - x_{0,25})}{x_{0,75} - x_{0,25}}\,. \tag{1.18}$$

Momentenkoeffizient der Schiefe:

$$\frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x})^3}{s_x^3}\,. \tag{1.19}$$

Positive Werte dieser Maßzahlen für Schiefe zeigen eine **rechtsschiefe** Verteilung an und **linksschiefe** Verteilungen besitzen negative Maßzahlen. Maßzahlen in der Nähe von Null deuten auf symmetrische Verteilungen hin.

Lageregel

symmetrische Verteilung	Modus $\approx$ Median $\approx \overline{x}$
rechtsschiefe Verteilung	Modus <Median $< \overline{x}$
linksschiefe Verteilung	Modus $>$ Median $> \overline{x}$

Lineare Transformation der Daten:
Gegeben seien Daten $x_1, ..., x_n$ mit Mittelwert $\overline{x}$. Nun werden die Daten gem ß einer linearen Abbildung transformiert, d.h. wir erhalten neue Daten

$$y_1 = a \cdot x_1 + b, y_2 = a \cdot x_2 + b, ..., y_n = a \cdot x_n + b, a \neq 0\,.$$

Dann gilt für den Quartilskoeffizienten der Schiefe QKS:

$$QKS_y = QKS_x.$$

Exzess bzw. **Überkurtosis**:

$$\frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x})^4}{s_x^4} - 3\,. \tag{1.20}$$

Die Größe

$$\frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x})^4}{s_x^4} \tag{1.21}$$

heißt **Kurtosis**.

1.5 Standardisierung

Standardisierung einer Stichprobe $x_1, \ldots, x_n$ (**z-Transformation**):

$$y_i = \frac{x_i - \overline{x}}{s_x} \ , \ i = 1, \ldots, n \, . \tag{1.22}$$

Nach dieser Transformation gilt: $\overline{y} = 0$ und $s_y = 1$.

1.6 Konzentrationsmessung

Gegeben sei nach wie vor eine Stichprobe $x_1, ..., x_n$ eines metrisch skalierten Merkmals.

Ordnungsstatistik: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Konzentrationsrate K_m gibt den Anteil der m größten Werte an der Gesamtsumme an.

$$K_m = \frac{x_{(n)} + x_{(n-1)} + \ldots + x_{(n-m+1)}}{x_1 + \ldots + x_n}, m = 1, \ldots, n \, . \tag{1.23}$$

Der Graph der Abbildung $m \to K_m$, wobei zwischen den Punkten (m, K_m) linear interpoliert wird, heißt **Konzentrationskurve**. Die Konzentrationskurve ist konkav und liegt oberhalb der Geraden durch die Punkte $(0, 0)$ und $(n, 1)$.

Lorenzkurve $\frac{m}{n} \to L_m$ gibt jeweils den Anteil der m kleinsten Werte an der Gesamtsumme:

$$L_m = \frac{x_{(1)} + x_{(2)} + \ldots + x_{(m)}}{x_1 + \ldots + x_n}, m = 1, \ldots, n \, . \tag{1.24}$$

Vielfach trägt man die folgende Abbildung $\frac{m}{n} \to L_m, m = 1, \ldots, n$ in einem Diagramm auf. Diese Kurve ist konvex und liegt unterhalb der Geraden durch die Punkte $(0, 0)$ und $(1, 1)$.

Je dichter die Konzentrationskurve bzw. die Lorenzkurve an der jeweils genannten Geraden liegen, desto gleichmäßiger ist die Verteilung der Werte (z.B. Einkommen, Umsätze, Vermögen) in der Stichprobe.

Man spricht von einer **Nullkonzentration**, wenn $L_m = \frac{m}{n} = K_m$ für alle m und von einer **Maximalkonzentration**, wenn $L_m = 0$ für alle $m < n$ und $L_n = 1$ bzw. wenn $K_m = 1$ für alle $m = 1, \ldots, n$.

Gini-Koeffizient: Maßzahl für den Grad der relativen Konzentration

$$G = 2 \cdot \frac{1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \ldots + n \cdot x_{(n)}}{n \cdot (x_1 + \ldots + x_n)} - \frac{n+1}{n} = 1 - \frac{1}{n} \sum_{i=1}^{n} (L_i + L_{i-1}) \,.$$

G gibt gerade den doppelten Flächeninhalt zwischen der Lorenzkurve und der Diagonalgeraden an.

Im Fall der Nullkonzentration gilt $G = 0$ und im Fall der Maximalkonzentration gilt $G = \frac{n-1}{n}$.

Gini-Koeffizient bei gruppierten Daten

$$G = 1 - \sum_{i=1}^{K} r_i (L_i^* + L_{i-1}^*) \,. \tag{1.25}$$

Dabei ist r_i die relative Häufigkeit des Anteils $\frac{x_i^*}{\sum_{j=1}^{n} x_j}$ an der Gesamtsumme $\sum_{j=1}^{n} x_j$.

Der Gini-Koeffizient ist ein Maß zur Messung der relativen Konzentration, und hängt nicht von der Stichprobengröße n ab.

Ein Maß zur Messung der absoluten Konzentration verwendet den Flächeninhalt oberhalb der **Konzentrationskurve** und unterhalb der Parallelen zur Abszisse (waagerechten Achse) durch den Punkt $(0, 1)$. Dieses Maß nennt man Rosenbluth-Index K_R. Es ist definiert als der Kehrwert des doppelten Flächeninhaltes oberhalb der Konzentrationskurve, und kann mit Hilfe des Gini-Koeffizienten G geschrieben werden als $K_R = \frac{1}{n(1-G)}$.

Ein weiteres häufiger verwendetes Maß zur Messung der absoluten Konzentration ist der Herfindahl-Index H, der definiert ist als die Summe der quadrierten Anteile an der Gesamtsumme $\sum_{i=1}^{n} x_i$ und mit Hilfe des Variationskoeffizienten V_x geschrieben werden kann als

$$H = \frac{1}{n} \cdot \left(\frac{n-1}{n} V_x^2 + 1 \right) \,. \tag{1.26}$$

Für diese Maße der absoluten Konzentration gilt: $\frac{1}{n} \leq K_R \leq 1$ bzw. $\frac{1}{n} \leq H \leq 1$. Im Falle der Maximalkonzentration, d.h. Konzentrationsrate $K_1 = 1$ nehmen diese Maße den Wert 1 an.

Der Herfindahl-Index und die Konzentrationsrate K_1 sind Spezialfälle einer Klasse von Maßen zur Messung der absoluten Konzentration: $D(\alpha) := \left(\sum_{i=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_j}\right)^{\alpha}\right)^{\frac{1}{\alpha-1}}$, und zwar für die Fälle $\alpha = 2$ bzw. $\alpha = \infty$ (genauer gesagt den Grenzwert von $D(\alpha)$ für $\alpha \to \infty$). Ein weiterer Spezialfall, der Grenzwert von $D(\alpha)$ für $\alpha \to 1$, führt auf den sogenannten *Exponentialindex*, dessen negativer natürlicher Logarithmus gerade die Shannon-Entropie ist. Der Exponentialindex ist ein gewichtetes Geometrisches Mittel der Anteile an der Gesamtsumme.

Ein anderes Maß zur Messung der relativen Konzentration, das durch den *Ungleichheitsaversionsparameter* $\epsilon > 0$ Werturteile einfließen lässt, ist das Atkinson-Maß $A(\epsilon)$

$$A(\epsilon) = \begin{cases} 1 - \frac{(\overline{x^{\epsilon-1}}_{harm})^{\frac{1}{\epsilon-1}}}{\overline{x}} & , \text{ falls } \epsilon > 1 \\ 1 - \frac{\overline{x}_{geom}}{\overline{x}} & , \text{ falls } \epsilon = 1 \\ 1 - \frac{(\overline{x^{1-\epsilon}})^{\frac{1}{1-\epsilon}}}{\overline{x}} & , \text{ falls } \epsilon \neq 1. \end{cases}$$

Dabei bedeutet die Notation $\overline{x^c} = \frac{1}{n}\sum_{i=1}^{n} x_i^c$, und entsprechend für das harmonische Mittel. Das Atkinson-Maß gehört für den Fall $0 < \epsilon \leq 1$ zu einer Klasse von Maßen zur Messung der relativen Konzentration, die durch den verallgemeinerten Entropie-Index $GE(\alpha)$ dargestellt werden können:

$$GE(\alpha) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left(\frac{\overline{x^\alpha}}{\overline{x}^\alpha} - 1\right) & , \text{ falls } \alpha \neq 0, \alpha \neq 1 \\ -\frac{1}{n}\sum_{i=1}^{n} \ln \frac{x_i}{\overline{x}} & , \text{ falls } \alpha = 0 \\ \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{\overline{x}} \ln \frac{x_i}{\overline{x}} & , \text{ falls } \alpha = 1. \end{cases}$$

$GE(0) =: T_L$ und $GE(1) =: T_T$ heißen auch *Theil*-Indices T_L und T_T. $GE(2) = \frac{n-1}{2n} V_x^2$.

Das Atkinson-Maß lässt sich für den Fall $0 < \epsilon < 1$ mit dem verallgemeinerten Entropie-Index darstellen als: $A(\epsilon) = 1 - [\epsilon(\epsilon-1)GE(1-\epsilon) + 1]^{\frac{1}{1-\epsilon}}$.

Ein Nachteil des verallgemeinerten Entropie-Indexes ist, dass er, im Gegensatz zum Atkinson-Index, nicht normiert ist, also auch Werte größer als 1 annehmen kann. Für die beiden Spezialfälle der Theil-Indices gibt es die Möglichkeit der Normierung $T_L^* := 1 - \exp(-T_L) = A(1)$ und $T_T^* := 1 - \exp(-T_T) = 1 - \frac{1}{n}\exp S$, wobei S die Shannon-Entropie ist, und damit $\exp S$ der Kehrwert des Exponentialindexes.

Theil-Indices bei gruppierten Daten

$$T_T = \sum_{i=1}^{K} (L_i^* - L_{i-1}^*) \ln \frac{L_i^* - L_{i-1}^*}{r_i} \qquad T_L = \sum_{i=1}^{K} r_i \ln \frac{r_i}{L_i^* - L_{i-1}^*}$$

Atkinson-Maß bei gruppierten Daten

$$A(\epsilon) = \begin{cases} 1 - \left(\sum_{i=1}^{K} r_i^\epsilon (L_i^* - L_{i-1}^*)^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}} & , \text{ falls } \epsilon \neq 1 \\ 1 - \prod_{i=1}^{K} \left(\frac{L_i^* - L_{i-1}^*}{r_i}\right)^{r_i} & , \text{ falls } \epsilon = 1. \end{cases}$$

$A(\epsilon)$ lässt sich für $\epsilon \neq 1$ auch in der Form

$$1 - \left(\sum_{i=1}^{K} r_i \left(\frac{L_i^* - L_{i-1}^*}{r_i} \right)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}$$

schreiben. Wie der Gini-Koeffizient und der als nächstes noch vorgestellte Hoover-Index lässt sich auch das Atkinson Maß anhand der Lorenzkurve interpretieren, und zwar als die Differenz zwischen der Steigung der Winkelhalbierenden, also 1, (was Nullkonzentration oder vollkommen gleichmäßige Verteilung bedeutet) und dem Hölder-Mittel der Steigungen der Lorenzkurve. Hierdurch werden kleine Steigungen, und das bedeutet, da die Lorenzkurve konvex (linksgekrümmt) ist, kleine Einkommen oder kleine Vermögen desto stärker gewichtet, je größer ϵ ist, in Übereinstimmung mit der obigen Interpretation von ϵ als Ungleichheitsaversionsparameter. Für $\epsilon = 1$ ergibt sich $A(1)$ als eins minus geometrisches Mittel der Steigungen der Lorenzkurve, für $\epsilon = 2$ als eins minus harmonisches Mittel der Steigungen der Lorenzkurve.

Ein weiteres wie der Gini-Koeffzient auch als relatives Streuungsmaß verwendbares (siehe relative mittlere absolute Differenz RMAD) und auch auf geometrischen Eigenschaften der Lorenzkurve beruhendes Maß der relativen Konzentration ist der

Hoover- bzw. Robin-Hood- bzw. Ricci-Schutz- bzw. Pietra-Index

$$HI = \frac{1}{2n\bar{x}} \sum_{i=1}^{n} |x_i - \bar{x}| = \frac{1}{2n} \sum_{i=1}^{n} \left| \frac{x_i}{\bar{x}} - 1 \right| = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{x_i}{\sum_{j=1}^{n} x_j} - \frac{1}{n} \right|$$

Hoover-Index bei gruppierten Daten

$$HI = \frac{1}{2} \sum_{i=1}^{K} |L_i^* - L_{i-1}^* - r_i| = \max_{i \in \{1,\ldots,K\}} \left| L_i^* - L_{i-1}^* - r_i \right|$$

Der Hoover-Index lässt sich auch schreiben als

$$HI = \frac{1}{2} \sum_{i=1}^{K} r_i \left| 1 - \frac{L_i^* - L_{i-1}^*}{r_i} \right| = \max_{i \in \{1,\ldots,K\}} r_i \left| 1 - \frac{L_i^* - L_{i-1}^*}{r_i} \right| ,$$

also als der mittlere Abstand bzw. das Maximum der Abstände zwischen der Steigung der Winkelhalbierenden (eins) und der Steigung der Lorenzkurve in dem jeweils betrachten Abschnitt der Lorenzkurve bzw. der maximale vertikale Abstand zwischen der Lorenzkurve und der Winkelhalbierenden.

Wegen der Dreiecksungleichung gilt: $G \leq 2HI$.

Da der Rang R_j eines Stichprobenwertes x_j, $j = 1, \ldots, n$ so definiert ist: $R_j = i \Leftrightarrow x_j = x_{(i)}$, $i = 1, \ldots, n$, gilt für den Gini-Koeffizienten G:

$$G = \frac{2}{n^2 \bar{x}} \sum_{j=1}^{n} R_j (x_j - \bar{x}) = \frac{2}{n^2} \sum_{j=1}^{n} R_j \left(\frac{x_j}{\bar{x}} - 1 \right)$$

Das bedeutet, während beim Hoover-Index alle Stichprobenwerte x_j gleich gewichtet werden, werden beim Gini-Koeffizienten, die Werte x_j gemäß ihrem Rang R_j gewichtet, und damit große Werte (deutlich) stärker gewichtet als kleine Werte. Dadurch wirkt sich Umverteilung (z.B. durch Steuerentlastungen) im oberen (z.B. Einkommens- oder Vermögens-)Bereich stärker auf den Gini-Koeffizienten aus als Umverteilung im unteren Bereich (bzw. zugunsten des unteren Bereiches).

Im Gegensatz dazu wichtet z.B. das Atkinson-Maß kleine Werte stärker als große Werte, sodass sich hier Umverteilung im unteren Bereich (bzw. zugunsten des unteren Bereiches) stärker auswirkt.

Der Variationskoeffizient V_x ist ein weiteres relatives Streuungsmaß, das sich auch zur Messung der relativen Konzentration verwenden lässt, allerdings ist er nicht normiert und kann Werte größer als 1 annehmen.

Für $\bar{x} > 0$ gilt:

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

Für alle vorgestellten Maße der relativen Konzentration gilt, dass sie sich darstellen lassen als

$$\sum_{i=1}^{n} g\left(\frac{x_i}{\bar{x}} \right)$$

mit einer *Gewichtsfunktion* $g : \mathbb{R} \to \mathbb{R}$.

2 Kontingenztafeln

In diesem Abschnitt befassen wir uns mit **nominal** skalierten mehrdimensionalen diskreten Variablen. Wir nehmen dazu an, dass die erste Variable die Ausprägungen (den Wertebereich) $A_1, ..., A_I$ und die zweite Variable den Wertebereich $B_1, ..., B_J$ besitzt. Für n Beobachtungseinheiten (Probanden) werden beide Merkmale erfasst.

2.1 Darstellung in Häufigkeitstabellen

Kontingenztafel, auch **Kreuztabelle**:

A \ B	B_1	$\cdots$	B_J	Summe
A_1	$h_{1,1}$	$\cdots$	$h_{1,J}$	$h_{1,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	$h_{I,1}$	$\cdots$	$h_{I,J}$	$h_{I,\bullet}$
Summe	$h_{\bullet,1}$	$\cdots$	$h_{\bullet,J}$	$h_{\bullet,\bullet} = n$

Kontingenztafel der relativen Häufigkeiten:

A \ B	B_1	$\cdots$	B_J	Summe
A_1	$r_{1,1}$	$\cdots$	$r_{1,J}$	$r_{1,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	$r_{I,1}$	$\cdots$	$r_{I,J}$	$r_{I,\bullet}$
Summe	$r_{\bullet,1}$	$\cdots$	$r_{\bullet,J}$	$r_{\bullet,\bullet} = 1$

Die Zeilen- bzw. Spaltensummen beschreiben die absoluten bzw. relativen Häufigkeiten bzgl. nur eines Merkmals. Sie heißen **Randverteilungen**.

2.2 Bedingte relative Häufigkeiten

Die bedingte relative Häufigkeit der Variablen B bei gegebener Information, dass die Variable A den Wert A_i annimmt, lautet

	B_1	$\cdots$	B_J	Summe
gegeben A_i	$h_{i,1}/h_{i,\bullet}$	$\cdots$	$h_{i,J}/h_{i,\bullet}$	$h_{i,\bullet}/h_{i,\bullet} = 1$

Eine große Heterogenität der bedingten Verteilungen (für die verschiedenen Bedingungen $A_1, \ldots, A_I$) deutet auf eine Abhängigkeit der beiden Merkmale A und B hin.

Völlig analog kann man auch bedingte relative Häufigkeiten für die Variable A gegeben, dass die Variable B den Wert B_j annimmt berechnen. Man erhält:

	gegeben B_j
A_1	$h_{1,j}/h_{\bullet,j}$
A_2	$h_{2,j}/h_{\bullet,j}$
$\vdots$	$\vdots$
A_I	$h_{I,j}/h_{\bullet,j}$
Summe	$h_{\bullet,j}/h_{\bullet,j} = 1$

2.3 Erwartete relative Häufigkeiten und Chi-Quadrat Koeffizient

Man nennt die aus den Rändern gemäß

$$r_{i,j}^{erwartet} = r_{i,\bullet} \cdot r_{\bullet,j} \text{ für alle } i, j\,.$$

rekonstruierte Kontigenztafel, die **Kontingenztafel der unter Unabhängigkeit erwarteten relativen Häufigkeiten**.

Chi-Quadrat-Koeffizient:

$$\chi_n^2 = n \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(r_{i,j} - r_{i,\bullet} \cdot r_{\bullet,j})^2}{r_{i,\bullet} \cdot r_{\bullet,j}}\,. \tag{2.1}$$

Je größer dieser Wert ist, desto stärker ist die Abweichung von der Unabhängigkeit der beiden Merkmale A und B.

Besonders einfach berechnet sich der Chi-Quadrat-Koeffizient für sog. 2×2-Kontingenztafeln der Form:

A \ B	B_1	B_2	Summe
A_1	h_{11}	h_{12}	$h_{1\bullet}$
A_2	h_{21}	h_{22}	$h_{2\bullet}$
Summe	$h_{\bullet 1}$	$h_{\bullet 2}$	n

$$\chi_n^2 = n \cdot \frac{(h_{11} \cdot h_{22} - h_{12} \cdot h_{21})^2}{h_{1\bullet} h_{2\bullet} h_{\bullet 1} h_{\bullet 2}} \,. \tag{2.2}$$

Korrigierter Kontingenzkoeffizient:

$$K^* = \sqrt{\frac{\chi^2}{n + \chi^2} \cdot \frac{M}{M-1}} \,,$$

wobei M das Minimum der Zeilen- und der Spaltenzahl in der Kontingenztafel bezeichnet: $M := \min(I, J)$. Stets gilt $0 \leq K^* \leq 1$.

Cramers V wird definiert durch

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

und nimmt wie K^* Werte im Intervall $[0, 1]$ an.

Phi-Koeffizient berechnet sich in **2×2**-Tafeln wie folgt:

$$\phi = \sqrt{\chi_n^2} \,.$$

Es gilt stets: $-1 \leq \phi \leq 1$.

3 Wahrscheinlichkeitstheoretische Grundlagen der Statistik

3.1 Zufälliges Ereignis und Wahrscheinlichkeit

Ein *Zufallsexperiment* ist ein zufälliger Vorgang, der unter **gleichen** Bedingungen **beliebig oft** nach bestimmter Vorschrift **wiederholbar** ist und der verschiedene vorher allesamt **bekannte mögliche Ausgänge** hat, wobei **unbekannt** ist, welcher dieser Ausgänge eintreten wird.

Eine *Wahrscheinlichkeit* ist eine Maßzahl mit möglichen Werten zwischen 0 und 1, die jedem Ausgang eines Zufallsexperiments zugeordnet werden kann. Sie gibt an, wie wahrscheinlich der jeweilige Ausgang des Zufallsexperiments ist. (Genauere Definition der Wahrscheinlichkeit folgt später.)

Ein *Ergebnis* ω ist ein möglicher Ausgang des Experiments.

Ein *Ereignis* ist eine Menge von Ergebnissen.

Ein *Elementarereignis* ist ein Ereignis, das nur ein Ergebnis enthält: $\{\omega\}$.

Der *Stichprobenraum* oder *Ergebnisraum* oder *Grundraum* Ω ist die Menge aller möglichen Ergebnisse ω.

DEFINITION (Wahrscheinlichkeits-)*Axiome von Kolmogorov*: Eine Abbildung P, die jedem zufälligen Ereignis A (aus einem bestimmten System von Teilmengen aus Ω) eine Wahrscheinlichkeit zuordnet, heißt *Wahrscheinlichkeitsmaß* bzw. *Wahrscheinlichkeitsverteilung* auf Ω, falls für P gilt:

1. $0 \leq \mathsf{P}(A) \leq 1$
2. $\mathsf{P}(\Omega) = 1$, Ω heißt deswegen auch *sicheres* Ereignis.
   (Bemerkung: $\mathsf{P}(\emptyset) = 0$, $\emptyset$ heißt deswegen auch *unmögliches* Ereignis.)
3. Wenn A und B unvereinbar sind (d.h. $A \cap B = \emptyset$), dann gilt $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$.

SATZ: Für ein Wahrscheinlichkeitsmaß P gilt

1. $\mathsf{P}(\overline{A}) = 1 - \mathsf{P}(A)$
2. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$ für beliebige Ereignisse A und B.

3. $A \subseteq B$, dann $\mathsf{P}(A) \leq \mathsf{P}(B)$

3.2 Laplace-Modell

Allgemein wird die Laplace-Wahrscheinlichkeit $\mathsf{P}(A)$ für ein Ereignis A definiert als Quotient aus Anzahl der *günstigen* durch Anzahl der möglichen Ergebnisse, in mathematischer Notation

$$\mathsf{P}(A) = \frac{|A|}{|\Omega|} \quad \text{mit} \quad A \subset \Omega \ ,$$

wobei $|A|$ die *Mächtigkeit* der Menge A bezeichnet, d.h. die Anzahl der Elemente der Menge A.

3.3 Zähldichte

Insbesondere gilt für *diskrete*, d.h. Ergebnisräume Ω mit endlich oder abzählbar vielen Elementen, d.h. $\Omega = \{\omega_1, \omega_2, \ldots\}$, also z.B. $\Omega = \{1, \ldots, n\}$ oder $\Omega = \mathbb{N}$, $\Omega = \mathbb{Z}$, $\Omega = \mathbb{Q}$:

Jede Abbildung $p : \Omega \to [0, 1]$ mit $\sum_{\omega \in \Omega} p(\omega) = 1$ heißt *Zähldichte*.

Für jedes Ereignis $A \subset \Omega$ lässt sich gemäß $\mathsf{P}(A) = \sum_{\omega \in A} p(\omega)$ eine Wahrscheinlichkeitsverteilung auf der Potenzmenge von Ω erklären, d.h. die Wahrscheinlichkeit für das Eintreten von A ist gerade die Summe der Einzelwahrscheinlichkeiten für die ω, die in A liegen. Die Wahrscheinlichkeit für das Eintreten von ω ist $p(\omega)$.

Dieses P erfüllt die Kolmogorov-Axiome (1-3) und die Eigenschaften 1-3 aus dem oben angegebenen Satz.

p heißt dann *Zähldichte* der Wahrscheinlichkeitsverteilung P.

Beispiel: Laplace-Verteilung aus vorigem Abschnitt mit Zähldichte $p(\omega) = \frac{1}{|\Omega|}$.

3.4 Stochastische Unabhängigkeit zufälliger Ereignisse, bedingte Wahrscheinlichkeit, Formel von Bayes

- Zwei Ereignisse A und B heißen **stochastisch unabhängig**, falls $\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$ gilt.

- Die Wahrscheinlichkeit des Ereignisses A unter der Bedingung, dass das Ereignis B eintritt (mit $\mathsf{P}(B) > 0$), ist definiert durch

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}$$

  und heißt **bedingte Wahrscheinlichkeit** von A unter B (*conditional probability*).

- Wenn $\mathsf{P}(A|B) = \mathsf{P}(A)$, d. h. Eintreten des Ereignisses B ändert Wahrscheinlichkeit von A nicht, dann heißen A und B **stochastisch unabhängig** (*independent*).

  Sei $\mathsf{P}(B) > 0$. $\mathsf{P}(A|B) = \mathsf{P}(A)$ gilt genau dann, wenn $\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$.

- **Formel der totalen Wahrscheinlichkeit:** Die Grundmenge Ω wird in unvereinbare Ereignisse B_j, $j = 1, \ldots, k$ zerlegt. ($\Omega = B_1 \cup \cdots \cup B_k$, $B_j \cap B_i = \emptyset$ für $j \neq i$). Dann gilt für ein beliebiges Ereignis A

$$\mathsf{P}(A) = \mathsf{P}(A|B_1)\mathsf{P}(B_1) + \cdots + \mathsf{P}(A|B_k)\mathsf{P}(B_k) \tag{3.1}$$

- **Bayessche Formel** Für beliebige Ereignisse A und B mit positiver Wahrscheinlichkeit gilt

$$\mathsf{P}(B|A) = \frac{\mathsf{P}(A|B)\mathsf{P}(B)}{\mathsf{P}(A)} \tag{3.2}$$

  Formel (1) und (2) liefern zusammen

$$\mathsf{P}(B_j|A) = \frac{\mathsf{P}(A|B_j)\mathsf{P}(B_j)}{\mathsf{P}(A|B_1)\mathsf{P}(B_1) + \cdots + \mathsf{P}(A|B_k)\mathsf{P}(B_k)}$$

3.5 Zufallsvariable, Wahrscheinlichkeitsverteilung

DEFINITION: Gegeben sei ein Ergebnisraum Ω, ein auf Ω definiertes Wahrscheinlichkeitsmaß P und ein Bildraum $\mathcal{X}$, wobei wir uns im folgenden im wesentlichen auf $\mathcal{X} = \mathbb{R}$ beschränken werden. Dann heißt eine Abbildung $X : \Omega \to \mathcal{X}$ *Zufallsvariable*.

P^X mit $\mathsf{P}^X(B) = P\{\omega \in \Omega : X(\omega) \in B\} = \mathsf{P}(X^{-1}(B))$, $B \subset \mathcal{X}$, wobei $X^{-1}(B)$ das Urbild von B unter der Abbildung X bezeichnet, heißt *Wahrscheinlichkeitsverteilung* der Zufallsvariable X.

Sei $\mathcal{X} = \mathbb{R}$ und $B = (-\infty, x]$. Dann heißt $F(x) := \mathsf{P}(X \leq x)$ ***Verteilungsfunktion*** der Zufallsvariablen X. Für $B = (-\infty, x]$ gilt

$F(x) := \mathsf{P}(X \leq x) = \mathsf{P}(X \in (-\infty, x]) = \mathsf{P}^X(-\infty, x] = \mathsf{P}^X(B)$

3.6 Spezielle *diskrete* Wahrscheinlichkeitsverteilungen

In diesem Abschnitt betrachten wir also nur endliche oder abzählbare Ergebnisräume. Die Wahrscheinlichkeitsverteilung heißt in diesem Fall deswegen auch *Zähl*dichte.

Die Verteilung

$$\mathsf{P}(X = k) := \frac{1}{n}\,, \quad k \in \{1, 2, 3, \cdots, n\},$$

Laplace-Verteilung bzw. **diskrete Gleichverteilung**. (Die Zufallsvariable X, die Werte aus $\mathcal{X}$ annehmen kann, heißt dann *Laplace*-verteilt oder (*diskret*) *gleichverteilt* auf $\{1, 2, 3, \ldots, n\}$.)

$$\mathsf{P}(X = k) := \binom{n}{k} p^k (1-p)^{n-k}\,, \quad \text{wobei } \binom{n}{k} = \frac{n!}{k!\,(n-k)!}\,, \quad k \in \{0, 1, 2, \cdots, n\},$$

heißt **Binomialverteilung** mit den Parametern n und p (Erfolgswahrscheinlichkeit).

$$\mathsf{P}(X = k) := e^{-\lambda}\frac{\lambda^k}{k!}\,, \quad , \quad k \in \mathbb{N} \cup \{0\},$$

Poisson-Verteilung mit dem Parameter λ, wobei $\mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \cdots\}$.

3.7 Stetige Wahrscheinlichkeitsverteilungen

Definition: Eine Funktion $f : \mathbb{R} \to [0, \infty)$ heißt **Wahrscheinlichkeitsdichte** oder kurz **Dichte**, falls gilt:

i) $\forall x \in \mathbb{R} : f(x) \geq 0$
ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Für jede Dichte f gilt: $\int_{-\infty}^{x} f(t)dt = F(x)$, wobei F die Verteilungsfunktion zur Dichte f ist, mit $F(x) = \mathsf{P}(X \leq x)$.

Spezielle stetige Verteilungen

$$f(x) := \begin{cases} \frac{1}{b-a} & \text{falls} \quad a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

heißt Dichte der *Rechteckverteilung* oder *stetige Gleichverteilung* auf dem Intervall $[a, b]$.

Eine Zufallsvariable X mit dieser Dichte, die Werte aus $\mathbb{R}$ annehmen kann, heißt dann *rechteck*verteilt oder (*stetig*) *gleichverteilt* auf $[a, b]$.

Die Verteilungsfunktion F mit

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

heißt Rechteckverteilung oder stetige Gleichverteilung auf $[a, b]$.

Ein wichtiger Spezialfall ist die Rechteckverteilung auf dem Intervall $[0, 1]$, die man für $a = 0$ und $b = 1$ erhält.

$$f(x) := \begin{cases} \lambda e^{-\lambda x} & \text{falls} \quad x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

heißt Dichte der **Exponentialverteilung** mit Parameter λ.

$$\varphi_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

heißt Dichte der **Normalverteilung** mit Erwartungswert μ und Varianz σ^2.

Stichprobe		**Population**	
Daten $x_1, \ldots, x_n$		Zufallsvariable X	
Häufigkeitsverteilung		**Wahrscheinlichkeitsverteilung**	
Definition und Darstellung			
diskretes Merkmal mit möglichen Werten $x_1^*, \ldots, x_m^*$		diskrete ZV mit möglichen Werten $x_1^*, \ldots, x_m^*$	
relative Häufigkeit von x_j^*: $h_n(x_j^*)$		Einzelwahrscheinlichkeit von x_j^*: $p(x_j^*) = \mathsf{P}(X = x_j^*)$	
Häufigkeitstabelle, Balkendiagramm		Wahrscheinlichkeitstabelle, Balkendiagramm	
Beobachtung eines stetigen Merkmals		stetige ZV	
Intervalleinteilung rel. Häufigk. des Intervalls I_j: $h_n(I_j)$		Wahrscheinlichk. eines Intervalls: $\mathsf{P}(X \in I_j) = \int_{I_j} f(x)\,\mathrm{d}x$, f Dichte	
Histogramm		Wahrscheinlichk.: Fläche unter f	
Eigenschaften			
diskret	stetig	diskret	stetig
$h_n(x_j^*) \geq 0$	$h_n(I_k) \geq 0$	$\mathsf{P}(X = x_j^*) \geq 0$	$f(x) \geq 0$
$\sum_{j=1}^{m} h_n(x_j^*) = 1$	$\sum_{k=1}^{l} h_n(I_k) = 1$	$\sum \mathsf{P}(X = x_j^*) = 1$	$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$

3.8 Populationskenngrößen

Falls die folgenden Reihen und (uneigentliche) Integrale existieren, definiert man:

$$\mathsf{E}X := \sum_{i=1}^{\infty} x_i \mathsf{P}(X = x_i)$$

heißt **Erwartungswert** der **diskreten** Zufallsvariable X,

$$\mathsf{Var}X := \sum_{i=1}^{\infty}(x_i - \mathsf{E}X)^2\mathsf{P}(X = x_i)$$

heißt **Varianz** der **diskreten** Zufallsvariable X, und $\sqrt{\mathsf{Var}X}$ die Streuung von X.

$$\mathsf{E}X := \int_{-\infty}^{\infty} xf(x)dx$$

heißt Erwartungswert der **stetigen** Zufallsvariable X.

$$\mathsf{Var}X := \int_{-\infty}^{\infty}(x - \mathsf{E}X)^2 f(x)dx$$

heißt **Varianz** der **stetigen** Zufallsvariable X, und $\sqrt{\mathsf{Var}X}$ die Streuung von X.

x_q mit $0 < q < 1$ heißt q**-Quantil** einer Zufallsvariable X, falls

$$\mathsf{P}(X \leq x_q) \geq q \quad \text{und} \quad \mathsf{P}(X \geq x_q) \geq 1 - q$$

.

Stichprobenparameter	**Populationsparameter**	
Mittelwert $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Erwartungswert $\mu = \sum_{j=1}^{\infty} x_j^* \, p(x_j^*)$	$\mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$
Streuung, empirische Varianz $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	Varianz $\sigma^2 = \sum_{j=1}^{\infty}(x_j^* - \mu)^2 \, p(x_j^*)$	$\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,\mathrm{d}x$
empirische Standardabweichung $s = \sqrt{s^2}$	Standardabweichung $\sigma = \sqrt{\sigma^2}$	
Stichprobenmedian für mindestens 50% der Daten $x_i \geq M$ und für mindestens 50% der Daten $x_i \leq M$ ("Halbierung der Stichprobe")	Median $\mathsf{P}(X \leq x_{med}) \geq 0,5$ und $\mathsf{P}(X \geq x_{med}) \geq 0,5$ ("Halbierung der Wk-masse")	
Stichprobenschiefe $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^3}}$	Schiefe $\frac{\sum_j (x_j^* - \mu)^3 \, p(x_j^*)}{\sigma^3}$	$\frac{\int_{-\infty}^{\infty}(x-\mu)^3 f(x)\,\mathrm{d}x}{\sigma^3}$
Daten: $(x_1, y_1), \ldots, (x_n, y_n)$	Zufallsvariable X und Y	
emp. Korrelationskoeffizient $\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X \, s_Y}$	Korrelationskoeffizient $\rho = \frac{\mathsf{E}((X - \mu_X)(Y - \mu_Y))}{\sigma_X \, \sigma_Y}$ Kovarianz $\mathsf{Cov}(X,Y) = \mathsf{E}((X - \mu_X)(Y - \mu_Y))$	

Für Erwartungswert und Varianz gelten die folgenden Rechenregeln für (diskrete und stetige) Zufallsvariablen:

$\mathsf{E}(aX + b) = a\mathsf{E}X + b$, a und b sind Konstanten
$\mathsf{E}(X + Y) = \mathsf{E}X + \mathsf{E}Y$ (Linearität)
$\mathsf{Cov}(X, Y) = \mathsf{E}(X \cdot Y) - \mathsf{E}X \cdot \mathsf{E}Y$
$\mathsf{Var}X = \mathsf{E}X^2 - (\mathsf{E}X)^2$
$\mathsf{E}(X \cdot Y) = \mathsf{E}X \cdot \mathsf{E}Y$, falls X und Y *unkorreliert*)
$\mathsf{Var}(X + Y) = \mathsf{Var}X + \mathsf{Var}Y + \mathsf{Cov}(X, Y)$, (allgemein)
$\mathsf{Var}(X + Y) = \mathsf{Var}X + \mathsf{Var}Y$, falls X und Y unkorreliert, d.h. $\mathsf{Cov}(X.Y) = 0$
$\mathsf{Var}(aX + b) = a^2\mathsf{Var}X$, a und b sind Konstanten

DEFINITION: Eine Funktion $f : \mathbb{R}^2 \to [0, \infty)$ heißt (bivariate) *Wahrscheinlichkeitsdichte* oder kurz *Dichte*, falls gilt:

i) $\forall (x, y) \in \mathbb{R}^2 : f(x, y) \geq 0$
ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

DEFINITION: Seien X und Y Zufallsvariablen mit endlichem Erwartungswert und positiver endlicher Varianz und seien X und Y entweder *stetig* und habe der Zufallsvektor (X, Y) die Dichte f oder seien X und Y *diskrete* Zufallsvariablen und habe (X, Y) die gemeinsame Verteilung $\mathsf{P}(X = x_i, Y = y_j)$ für alle i und j. Dann heißt

$$\begin{aligned} \mathsf{Cov}(X, Y) &= \mathsf{E}[(X - \mathsf{E}X)(Y - \mathsf{E}Y)] \\ &= \begin{cases} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i - \mathsf{E}X)(y_j - \mathsf{E}Y)\mathsf{P}(X = x_i, Y = y_j) & \text{falls} \quad X, Y \text{ diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathsf{E}X)(y - \mathsf{E}Y) f(x, y) dx dy & \text{falls} \quad X, Y \text{ stetig} \end{cases} \end{aligned}$$

die *Kovarianz* von X und Y,

$$\rho(X, Y) := \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}X}\sqrt{\mathsf{Var}Y}}$$

die *Korrelation* von X und Y.

Es gilt: $-1 \leq \rho(X, Y) \leq 1$.

DEFINITION:

i) Zwei Zufallsvariablen X und Y heißen *unkorreliert*, falls $\rho(X, Y) = 0$.

ii) Zwei Zufallsvariablen X und Y heißen *stochastisch unabhängig*, falls $\mathsf{P}(X \leq x, Y \leq y) = \mathsf{P}(X \leq x) \cdot \mathsf{P}(Y \leq y)$ für alle $x, y \in \mathbb{R}$.

Zwei *diskrete* Zufallvariablen X mit Werten x_i und Y mit Werten y_i, $i \in \mathbb{N}$ heißen *stochastisch unabhängig*, falls $\mathsf{P}(X = x_i, Y = y_i) = \mathsf{P}(X = x_i) \cdot \mathsf{P}(Y = y_i)$ für alle $i \in \mathbb{N}$.

Insbesondere gilt: Sind zwei Zufallsvariablen X und Y stochastisch unabhängig, dann sind sie auch unkorreliert.

Umgekehrt sind zwei unkorrelierte Zufallsvariablen im allgemeinen **nicht** stochastisch unabhängig.

DEFINITION: Seien X und Y *stetige* Zufallsvariablen mit Verteilungsfunktionen F und G und habe der Zufallsvektor (X, Y) die Dichte f. Dann heißt

$$\begin{aligned} \rho_S(X,Y) &:= \rho(F(X), G(Y)) = 12\mathsf{E}[F(X)G(Y)] - 3 \\ &= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x)G(y)f(x,y)dxdy - 3 \end{aligned}$$

die Rangkorrelation von X und Y.

SATZ (CEBYSEV-UNGLEICHUNG): Sei X eine (beliebige diskrete oder stetige) Zufallsvariable mit $\mathsf{E}X = \mu$ und $\mathsf{Var}X = \sigma^2 > 0$. Dann gilt für alle $\epsilon > 0$:

$$\mathsf{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Setzt man nun speziell $\epsilon = k\sigma$, dann gilt für jede Zufallsvariable deshalb insbesondere die $k\sigma$-Regel für alle $k \in \mathbb{N}$:

$$\mathsf{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Gleichwertig hierzu ist:

$$\mathsf{P}\left(\left|\frac{X - \mu}{\sigma}\right| \geq k\right) \leq \frac{1}{k^2}.$$

Die Variable $Z := \frac{X-\mu}{\sigma}$ nennt man auch standardisiert (bzw. z-transformiert), denn mit den Rechenregeln für Erwartungswert und Varianz von Zufallsvariablen erhält man: $\mathsf{E}Z = 0$ und $\mathsf{Var}Z = 1$. (siehe auch Kapitel 1.5 mit $z := \frac{x-\bar{x}}{s_x}$, $\bar{z} = 0$ und $s_z = 1$)

3.9 Gesetz der Großen Zahlen und Zentraler Grenzwertsatz

Schwaches Gesetz der Großen Zahlen (Weak Law of Large Numbers)

SATZ: Seien $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilte Zufallsvariablen mit Erwartungswert $\mathsf{E}X_1 = \mu$ und $\mathsf{Var}X_1 = \sigma^2 \in (0, \infty)$ und sei $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Dann gilt für alle $\epsilon > 0$:

$$\mathsf{P}(|\bar{X} - \mu| > \epsilon) \to 0 \quad \text{für} \quad n \to \infty.$$

Zentraler Grenzwertsatz (Central Limit Theorem)

SATZ: Seien $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilte Zufallsvariablen mit Erwartungswert $\mathsf{E}X_1 = \mu$ und $\mathsf{Var}X_1 = \sigma^2 \in (0, \infty)$ und sei $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ und $Z_n := \sigma^{-1}\sqrt{n}(\bar{X} - \mu)$. Dann gilt für alle $a < b$:

$$\mathsf{P}(a < Z_n < b) \to \int_a^b \varphi(x)dx \quad \text{für} \quad n \to \infty$$

Sei F_{Z_n} die Verteilungsfunktion von Z_n und Φ die Verteilungsfunktion der Standardnormalverteilung, dann lässt sich die Folgerung des Zentralen Grenzwertes auch formulieren als:

Dann gilt für alle Stetigkeitsstellen x von F_{Z_n}:

$$F_{Z_n}(x) \to \Phi(x) \quad \text{für} \quad n \to \infty.$$

- Der zentrale Grenzwertsatz besagt, dass sich die Verteilung der (standardisierten) Größe

  $$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

  für große Stichprobenumfänge n approximativ (näherungsweise) durch die Standardnormalverteilung beschreiben läßt.

- Sprechweisen: Die Verteilung von Z_n konvergiert für $n \to \infty$ gegen die Standardnormalverteilung, oder

  Z_n ist asymptotisch normal (verteilt)

  $$\boldsymbol{Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \to \mathbf{N(0,1)}}$$

4 Schätzen

4.1 Momentenschätzer

Das k-te Moment $\mathrm{E}X^k$, $k = 1, 2, 3, \ldots$ der Zufallsvariable X wird hier durch das k-te empirische Moment $\frac{1}{n}\sum_{i=1}^{n} x_i^k$ geschätzt.

4.2 Kleinste-Quadrate-Schätzer

Siehe Kapitel *Lineare Regression*.

Ein weiteres häufig verwendetes Schätzverfahren ist die Maximum-Likelihood-Methode. Diese wird in dieser Vorlesung aber nicht besprochen.

4.3 Statistische Eigenschaften von Schätzern

$\{P_\vartheta : \vartheta\}$ sei eine Familie von Wahrscheinlichkeitsmaßen mit unbekanntem Verteilungsparameter ϑ, z.B. $\vartheta = \mu$, Erwartungswert der W-Verteilung. E_ϑ bzw. Var_ϑ seien die unter den W-Maßen P_ϑ berechneten Erwartungsgwerte bzw. Varianzen $\hat{\vartheta}$ sei ein Schätzer für ϑ.

Gilt für alle ϑ

$\mathrm{E}_\vartheta\hat{\vartheta} \to \vartheta$ für $n \to \infty$ (d.h. $\hat{\vartheta}$ *asymptotisch* erwartungstreuer Schätzer für ϑ) und

$\mathrm{Var}_\vartheta\hat{\vartheta} \to 0$ für $n \to \infty$,

dann heißt $\hat{\vartheta}$ *konsistenter* Schätzer für ϑ.

Man sagt auch der Schätzer $\hat{\vartheta}$ konvergiert dann *nach Wahrscheinlichkeit* gegen ϑ:

Für alle $\epsilon > 0$ gilt $\mathrm{P}(|\hat{\vartheta} - \vartheta| > \epsilon) \to 0$ für $n \to \infty$.

Siehe auch **Schwaches Gesetz der Großen Zahlen** (Weak Law Of Large Numbers):

Seien $X_1, \ldots, X_n$ stochastisch *unabhängig* und *identisch* verteilte Zufallsvariablen mit Erwartungswert $\mathrm{E}X_i = \mu$ für $i = 1, \ldots, n$, wobei $\mu \in \mathbb{R}$, und Varianz $\mathrm{Var}X_i = \sigma^2$, wobei $\sigma^2 \in (0, \infty)$, und sei $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Dann gilt

$$\mathrm{P}(|\bar{X} - \mu| > \epsilon) \to 0 \text{ für } n \to \infty.$$

D.h. das *Stichproben*mittel $\bar{X}$ konvergiert nach Wahrscheinlichkeit gegen das *Populations*mittel (den Erwartungswert) μ der Zufallsvariablen X_i, $i = 1, \ldots, n$.

kurz: $\bar{X}$ ist konsistenter Schätzer für μ.

4.4 Bereichsschätzer: Konfidenzintervalle

- aus den Daten ermitteltes **Intervall plausibler Werte für ϑ**
- Schreibweisen: $[\hat{\vartheta}_u, \hat{\vartheta}_o]$, I_n,
- Das Intervall $\boldsymbol{[\hat{\vartheta}_u, \hat{\vartheta}_o]}$ heißt Konfidenzintervall zum Niveau $\boldsymbol{1-\alpha}$ für den Parameter $\boldsymbol{\vartheta}$, wenn

$$\mathbf{P}(\hat{\vartheta}_u \leq \vartheta \leq \hat{\vartheta}_o) \geq 1 - \alpha.$$

- übliche Werte für $\boldsymbol{\alpha}$: 0,10; 0,05; 0,01

Konfidenzintervalle, wenn Normalverteilungsannahme gerechtfertigt
Die Daten $x_1, \ldots, x_n$ seien Realisierungen unabhängiger $\mathsf{N}(\mu, \sigma^2)$-verteilter ZV.

1. Intervall für den Erwartungswert μ bei unbekannter Varianz

$$\left[\overline{x} - t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \; , \; \overline{x} + t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right].$$

   Hierbei ist $\overline{x}$ der Mittelwert, s die Stichproben-Standardabweichung und $t_{n-1;1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$ - Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

2. Intervall für den Erwartungswert μ bei bekannter Varianz σ^2

$$\left[\overline{x} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; , \; \overline{x} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

   Hierbei ist $u_{1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$ - Quantil der Standardnormalverteilung.

3. Intervall für die Varianz σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}\ ,\ \frac{(n-1)s^2}{\chi^2_{n-1;\frac{\alpha}{2}}}\right].$$

Hierbei ist $\chi^2_{n-1;1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$ - Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden.

4. (2 unverbundene Stichproben) Die Daten $x_1,\ldots,x_{n_1}$ und $y_1,\ldots,y_{n_2}$ seien Realisierungen zweier unabhängiger Stichproben von unabhängigen $\mathsf{N}(\mu_x,\sigma_x^2)$- bzw. $\mathsf{N}(\mu_y,\sigma_y^2)$verteilten ZV.

Intervall für die Differenz $\mu_x-\mu_y$

Es gelte $\sigma_x^2=\sigma_y^2=\sigma^2$.

$$\left[\overline{x}-\overline{y}\ -\ t_{n_1+n_2-2;1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{n_1+n_2}{n_1n_2}}\ ,\ \overline{x}-\overline{y}\ +\ t_{n_1+n_2-2;1-\frac{\alpha}{2}}\ \hat{\sigma}\sqrt{\frac{n_1+n_2}{n_1n_2}}\right],$$

wobei $\hat{\sigma}^2$ der gepoolte Varianzschätzer ist: **'gepoolte Varianz'**

$$\hat{\sigma}^2\ =\ \frac{(n_1-1)s_x^2\ +\ (n_2-1)s_y^2}{n_1+n_2-2}.$$

Approximative Konfidenzintervalle bei großem Stichprobenumfang

Die Daten $x_1,\ldots,x_n$ seien Realisierungen unabhängiger identisch verteilter ZV (d.h. die Voraussetzungen für den **Zentralen Grenzwertsatz**, s. Kapitel 4.11, sind erfüllt), die Varianz von X_1 σ^2 sei **unbekannt** und der **Stichprobenumfang n** sei **groß**.

1. Intervall für den Erwartungswert μ

$$\left[\overline{x}\ -\ u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\ ,\ \overline{x}\ +\ u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right].$$

Hierbei ist $u_{1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$ - Quantil der Standardnormalverteilung.

2. Intervall für die Wahrscheinlichkeit eines Ereignisses, d.h. für den Parameter p der Bernoulli-Verteilung

$$\left[\hat{p}\ -\ u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p}\ +\ u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Hierbei ist $\hat{p}$ die relative Häufigkeit des Ereignisses, d.h. $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$, wobei $x_i = 1$, wenn das betreffende Ereignis beobachtet wird und $x_i = 0$ sonst.

5 Statistische Tests

5.1 Allgemeines zum Testen

Ziel:
mit Hilfe von Beobachtungen (in Stichproben) sollen Vermutungen über Grundgesamtheiten (Populationen) überprüft werden

- Vermutungen betreffen bestimmte Parameter der Verteilung, den Verteilungstyp oder Unabhängigkeit bzw. Abhängigkeit von Merkmalen
- Vermutungen werden als **Hypothesen** formuliert.
- Die Hypothesen (Nullhypothese H und Gegenhypothese/Alternative K) über den Parameter ϑ einer Verteilung haben im Allgemeinen die folgende Form:

$$H : \vartheta \leq \vartheta_0 \text{ gegen } K : \ \vartheta > \vartheta_0 \text{ bzw. } H : \ \vartheta \geq \vartheta_0 \text{ gegen } K : \ \vartheta < \vartheta_0 \tag{5.1}$$

oder

$$H : \ \vartheta = \vartheta_0 \text{ gegen } K : \ \vartheta \neq \vartheta_0, \tag{5.2}$$

wobei ϑ_0 ein vorgegebener bekannter Wert ist.

Die Probleme (5.1) heißen **einseitiges** Testproblem, das Problem (5.2) heißt **zweiseitig**.

- **Fehlermöglichkeiten** Ergebnis eines Tests ist die Entscheidung für eine der beiden Hypothesen. Da diese Entscheidung auf der Basis einer Stichprobe getroffen wird, sind Fehlentscheidungen möglich. Wir unterscheiden dabei:

	H richtig	K richtig
H angenommen	Entscheidung richtig	Fehler 2. Art
K angenommen	Fehler 1. Art	Entscheidung richtig

Der Fehler, der entsteht, wenn H abgelehnt wird, obwohl H gilt, heißt **Fehler 1. Art**. Dieser kann z.B. dadurch auftreten, dass die Stichprobe, die ja zufällig ausgewählt wird, typische Merkmale der hypothetischen Grundgesamtheit, ausgedrückt durch H, nicht repräsentiert,

obwohl sie aus ihr stammt. Der Fehler, der dadurch entsteht, dass H angenommen wird, obwohl K gilt, heißt **Fehler 2. Art**.

- Die Wahrscheinlichkeiten für beide Fehlerarten können nicht gleichzeitig minimiert werden, deshalb betrachtet man **Signifikanztests zum Niveau α**:

  Man gibt sich eine (kleine) Wahrscheinlichkeit α für den Fehler 1. Art vor und konstruiert **Testverfahren, bei denen die Wahrscheinlichkeit für den Fehler 1. Art dieses sogenannte Signifikanzniveau (significance level) α nicht überschreitet.**

  In einem weiteren Schritt, versucht man dann unter all den Signifikanztests zum Niveau α einen solchen zu finden, für den die Fehlerwahrscheinlichkeit 2. Art minimal wird.

Konsequenzen für Formulierung der Hypothesen:
Die Aussage, deren Richtigkeit nachgewiesen werden soll bzw. deren Konsequenzen schwerwiegender sind, wird man, wenn möglich, nicht als Nullhypothese formulieren, sondern als Alternative.

- **Ablehnung der Hypothese *H*** bedeutet dann, dass die Alternative statistisch gesichert ist.

  Redeweise: **Alternative statistisch gesichert**,
  (Abweichung von der Nullhypothese ist signifikant)

  hierbei möglicherweise große Wahrscheinlichkeit, die Nullhypothese anzunehmen, obwohl sie falsch ist. (Fehler 2. Art)

  Annahme von der Hypothese *H* bedeutet **nicht**, dass diese Hypothese statistisch gesichert ist.

  Redeweise: **Nullhypothese durch die Beobachtungen nicht widerlegt**

- Übliche α-Werte sind: 0.01, 0.05, 0.001
  Man beachte, dass ein kleineres α zu einer größeren Wahrscheinlichkeit des Fehlers 2. Art führt.

5.2 Allgemeines Prozedere zur Konstruktion eines Tests

1. Aufstellen des statistischen Modells (Verteilungsannahme)[1]

[1] Dieser Abschnitt dient zum Verständnis des Prinzips eines Testverfahrens. Im Rahmen der Statistikvorlesung wird nicht erwartet, dass Sie selbständig Testverfahren aufstellen.

2. Formulierung der Null-und Alternativhypothese H und K
3. Vorgabe des Signifikanzniveaus α
4. Wahl der **Test-(Prüf-) Statistik t**
   Erwünscht:
   - t soll K gut widerspiegeln
   - die **Verteilung von t unter der Annahme, dass H gilt**, soll berechenbar sein

   Typische Form von t:

   $\boldsymbol{t} =$ Abstand zwischen beobachtetem Wert und hypothetischem Wert (geeignet normiert)
5. Bestimmung der Verteilung der Teststatistik unter der Hypothese H
6. Angabe des kritischen Bereichs S, d.h. **Berechnung des kritischen Werts,** d.h. c_α so, dass

   $$\mathsf{P}_H(t > c_\alpha) \leq \alpha \text{ bzw. } \mathsf{P}_H(t < c_\alpha) \leq \alpha \text{ bzw. } \mathsf{P}_H(|t| < c_\alpha) \leq \alpha$$

   wobei P_H für die Wahrscheinlichkeit unter der Hypothese H steht.
7. Berechnung der Teststatistik $t = t(x_1, \ldots, x_n)$ für die gegebenen Beobachtungswerte
8. **Entscheidungsregel: Gilt für die Beobachtungswerte $(x_1, \ldots, x_n)$**

   $$\boldsymbol{t > c_\alpha} \text{ (bzw. } \boldsymbol{t < c_\alpha} \text{ bzw. } \boldsymbol{|t| \geq c_\alpha})$$

   so wird H abgelehnt; andernfalls ist gegen H nichts einzuwenden.

   (Äquivalent dazu: H ablehnen, wenn Signifikanzwert kleiner als α.)

5.3 Güteaussagen

Unter der Güte (Trennschärfe, power) eines Tests versteht man die Wahrscheinlichkeit, dass die Hypothese H abgelehnt wird, wenn ein Parameter aus der Alternativmenge richtig ist (das heißt 1-Wahrscheinlichkeit des Fehlers 2. Art).
Allgemein kann man über die Güte eines Tests folgende Aussagen treffen:

- Optimale Tests sind so definiert, dass sie bei vorgegebenem α maximale Güte haben.
- Die Güte steigt mit wachsendem n.
- Die Güte sinkt, wenn α kleiner wird.
- Die Güte steigt, wenn die Variabilität kleiner wird.
- Die Güte bei einseitigen Tests ist größer.

5.4 Konfidenzintervalle und Tests

Wenn bei einem zweiseitigen Testproblem der hypothetische Wert ϑ_0 vom Konfidenzintervall überdeckt wird, dann ist H nicht zu verwerfen; **ist der hypothetische Wert nicht Element des Konfidenzintervalls, so wird H abgelehnt**.

5.5 Testen des Erwartungswerts einer Population

Testprobleme der folgenden Art werden betrachtet:

$$\text{a) } H: \ \vartheta = \vartheta_0 \text{ gegen } K: \ \vartheta \neq \vartheta_0$$

$$\text{b) } H: \ \vartheta \leq \vartheta_0 \text{ gegen } K: \ \vartheta > \vartheta_0$$

$$\text{c) } H: \ \vartheta \geq \vartheta_0 \text{ gegen } K: \ \vartheta < \vartheta_0$$

Ferner seien u_q, $t_{n;q}$ und $\chi^2_{n;q}$ das q-Quantil der Standardnormalverteilung, der t-Verteilung mit n Freiheitsgraden bzw. der Chi-Quadrat-Verteilung mit n Freiheitsgraden.

Der interessierende Parameter ϑ ist der Erwartungswert der Population; d.h. $\vartheta = \mu$. Es soll untersucht werden, ob dieser Parameter gleich (kleiner/gleich, größer/gleich) einem vorgegebenen Wert $\mu_0 = \vartheta_0$ ist.

Lösung:

A Wenn angenommen werden kann, dass die Daten aus einer **normalverteilten Population** stammen, also $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, und die Varianz σ^2 unbekannt ist, und mit Daten aus einer Stichprobe z.B.

durch die Stichprobenvarianz s^2 geschätzt werden muss, dann verwendet man den (exakten) ***t*-Test**.

B Wenn die Normalverteilungsannahme nicht gerechtfertigt ist, kann man den **Gauß-Test** oder den ***t*-Test** als **approximatives** Verfahren verwenden, wenn der Stichprobenumfang $\boldsymbol{n}$ **groß** ist.

C In der eher seltenen Situation, in der die Normalverteilungsannahme gerechtfertigt und die Varianz σ^2 der Population bekannt ist, kann man den Gauß Test als **exaktes** Verfahren verwenden.

D Wenn die Normalverteilungsannahme nicht gerechtfertigt ist, und der Stichprobenumfang nicht groß ist, muss man passen! (Wenn andere Verteilungsannahmen gerechtfertigt sind, gibt es in manchen Fällen spezielle Verfahren in der Literatur; z.B. bei Exponentialverteilung in der Lebensdauer-Theorie.)

Verfahren:

A ***t* - Test.** Die Normalverteilungsannahme sei gerechtfertigt, d.h. $x_1, \ldots, x_n$ sind Realisierungen unabhängiger $\mathsf{N}(\mu, \sigma^2)$-verteilter ZV; der interessierende Parameter ist $\vartheta = \mu$. Wenn $\mu = \mu_0$ ist die Teststatistik

$$\boldsymbol{t = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}}$$

t-verteilt mit $(n-1)$ Freiheitsgraden; H in a), b) bzw. c) wird abgelehnt, wenn

a)	$\lvert t\rvert \geq t_{n-1;1-\frac{\alpha}{2}}$	$P_{2-seitig} = \mathsf{P}(\lvert T\rvert \geq \lvert t\rvert) < \alpha$	μ_0 nicht vom $(1-\alpha)$-KI überdeckt
b)	$t \geq t_{n-1;1-\alpha}$	$P_{1-seitig} = \mathsf{P}(T \geq t) < \alpha$	μ_0 liegt unterhalb des $(1-2\alpha)$-KI
c)	$t \leq -t_{n-1;1-\alpha}$	$P_{1-seitig} = \mathsf{P}(T \leq t) < \alpha$	μ_0 liegt oberhalb des $(1-2\alpha)$-KI

2

[2] KI Abkürzung für Konfidenzintervall

B **approximativer Gauß-Test.** Die Normalverteilungsannahme sei nicht gerechtfertigt, aber der Stichprobenumfang n ist groß; der interessierende Parameter sei $\vartheta = \mu$, dann ist die Teststatistik

$$\boldsymbol{z = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}},$$

für $\mu = \mu_0$ nach dem ZGWS **annähernd** standardnormalverteilt; H wird abgelehnt, wenn

a)	$\lvert z\rvert \geq u_{1-\frac{\alpha}{2}}$	asympt.$P_{2-seitig} < \alpha$	μ_0 nicht vom asympt. $(1-\alpha)$-KI überdeckt
b)	$z \geq u_{1-\alpha}$	asympt.$P_{1-seitig} < \alpha$	μ_0 unterhalb des asympt.$(1-2\alpha)$-KI liegt
c)	$z \leq -u_{1-\alpha}$	asympt.$P_{1-seitig} < \alpha$	μ_0 liegt oberhalb des asympt.$(1-2\alpha)$-KI liegt

Semi- bzw. Nichtparametrische Tests

5.6 Chi-Quadrat-Anpassungstest

Die Daten $x_1, \ldots, x_n$ seien Realisierungen einer ZV X, die die Werte $x_1^*, \ldots, x_k^*$ mit den Wahrscheinlichkeiten $p_j = \mathsf{P}(X = x_j^*)$ annimmt. Die hypothetischen Wahrscheinlichkeiten seien p_{0j}, $j = 1, \ldots, k$. Getestet werden soll

$$H: \; p_j = p_{0j} \text{ für alle } j \text{ gegen } K: \; p_{j'} \neq p_{0j'} \text{ für mindestens ein } j'.$$

Sei h_j die absolute Häufigkeit des Werts x_j^* und $r_j = h_j/n$. Wenn die Nullhypothese gilt, dann kann die Verteilung der Teststatistik

$$\boldsymbol{t = \sum_{j=1}^{k} \frac{(h_j - np_{0j})^2}{np_{0j}} = n \sum_{j=1}^{k} \frac{(r_j - p_{0j})^2}{p_{0j}},}$$

durch eine χ^2-Verteilung mit $k-1$ Freiheitsgraden angenähert werden. Vorausgesetzt wird, dass n so groß ist , dass $np_{0j} \geq 5$ für alle j. Man lehnt H ab, wenn

$$\boldsymbol{t \geq \chi^2_{k-1;1-\alpha}} \text{ bzw. } \boldsymbol{P = \mathsf{P}(T \geq t) < \alpha.}$$

Realisierung in SPSS: Analysieren - Nichtparametrische Tests - Chi-Quadrat

(hypothetische Wahrscheinlichkeiten eingeben)

Realisierung in EXCEL: beobachtete Häufigkeiten eingeben; erwartete Häufigkeiten berechnen;
Prozedur CHITEST liefert Signifikanzwert $\mathsf{P}(T \geq t)$.

Bemerkung 1: Hängen die hypothetischen Wahrscheinlichkeiten noch von einem unbekannten Parameter ϑ ab, d.h. $p_{0j}(\vartheta)$, so kann man das Verfahren modifizieren. Die Teststatistik hat dann die Form

$$t = \sum_{j=1}^{k} \frac{(h_j - np_{0j}(\hat{\vartheta}))^2}{np_{0j}(\hat{\vartheta})},$$

wobei $\hat{\vartheta}$ ein geeigneter Schätzwert für den unbekannten Parameter ist. Unter bestimmten Voraussetzungen (siehe z.B. in Läuter/Pincus: Statistische Datenanalyse) kann gezeigt werden, dass man die Verteilung dieser Statistik

durch eine χ^2-Verteilung mit $k-1-m$ Freiheitsgraden annähern kann, wobei m die Dimension des unbekannten ϑ ist.

Bemerkung 2: Dieses Testverfahren ist auch anwendbar für das Testen der Verteilung einer stetigen ZV. Dann geht man über zu einer Klasseneinteilung. An die Stelle der Werte x_j^* treten die Klassen I_j; die Wahrscheinlichkeiten sind nun die Klassenwahrscheinlichkeiten $p_j = \mathsf{P}(X \in I_j)$ und die Häufigkeiten h_j sind nun Klassen-Häufigkeiten.

Der Übergang zur Klasseneinteilung ist auch sinnvoll bei diskreten ZV, die sehr viele verschiedene Werte x_j^* annehmen können.

Man beachte aber: Durch die Klasseneinteilung kommt es zu einem Informationsverlust, und die Testentscheidung ist abhängig von der gewählten Klasseneinteilung.

Kontingenztafeln

		Wert y						
		y_1^*	y_2^*	$\cdots$	y_m^*	$\cdots$	y_l^*	Summe
Wert x	x_1^*	h_{11}	h_{12}	$\cdots$	h_{1m}	$\cdots$	h_{1l}	$h_{1\cdot}$
	x_2^*	h_{21}	h_{22}	$\cdots$	h_{2m}	$\cdots$	h_{2l}	$h_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_j^*	h_{j1}	h_{j2}	$\cdots$	h_{jm}	$\cdots$	h_{jl}	$h_{j\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_k^*	h_{k1}	h_{k2}	$\cdots$	h_{km}	$\cdots$	h_{kl}	$h_{k\cdot}$
	Summe	$h_{\cdot 1}$	$h_{\cdot 2}$	$\cdots$	$h_{\cdot m}$	$\cdots$	$h_{\cdot l}$	$h_{\cdot\cdot} = n$

h_{jm} ist die Anzahl (oder absolute Häufigkeit) aller Paare in der Stichprobe, bei denen die x-Komponente den Wert x_j^* **und** die y-Komponente den Wert y_m^* annimmt

Eine (empirische, rein deskriptive) Maßzahl die die (stochastische) Abhängigkeit zwischen zwei Merkmalen beschreibt ist der *Chi-Quadrat-Koeffizient* oder der *quadratische Kontingenzkoeffizient*.

Dieser ist definiert durch

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \frac{1}{n} \cdot h_{i\cdot} \cdot h_{\cdot j})^2}{\frac{1}{n} \cdot h_{i\cdot} \cdot h_{\cdot j}}$$

oder, indem man statt der absoluten die relativen Häufigkeiten verwendet durch

$$\chi^2 = n \cdot \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(r_{ij} - r_{i\cdot} \cdot r_{\cdot j})^2}{r_{i\cdot} \cdot r_{\cdot j}}$$

und nimmt Werte im Intervall $[0, \infty)$ (Ist der Wert des Koeffizienten null, so kann man **nicht** von (stochastischer) Abhängigkeit der Merkmale ausgehen.) an, ist also nicht negativ, kann aber beliebig groß werden. Deswegen gibt es Normierungen dieses Maßes wie z.B. die folgenden:

Cramers V wird mit $M := \min(k, l)$, also bezeichnet M das Minimum der Zeilen- und der Spaltenzahl in der Kontingenztafel bezeichnet, definiert durch

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

und nimmt Werte im Intervall $[0, 1]$ an.

Der *korrigierte* Kontingenzkoeffizient wird mit $M := \min(k, l)$ definiert durch

$$K^* = \sqrt{\frac{\chi^2 \cdot M}{(\chi^2 + n)(M-1)}}$$

und nimmt Werte im Intervall $[0, 1]$ an.

5.7 Testen von Unabhängigkeit:

Chi-Quadrat-Unabhängigkeitstest. Beobachtet werden Datenpaare von **kategorischen**, d.h. nominalen oder ordinalen oder (endlich) diskreten Daten. Es soll getestet werden, ob zwischen den Größen eine Abhängigkeit besteht. Die Paare $(x_1, y_1) \ldots, (x_n, y_n)$ seien Realisierungen von ZV (X,Y), die die Werte $x_1^*, \ldots, x_k^*$ bzw. $y_1^*, \ldots, y_m^*$ mit den Wahrscheinlichkeiten $p_{jl} = \mathsf{P}(X = x_j^*, Y = y_l^*)$ annehmen;
getestet werden soll, ob X und Y unabhängig sind, d.h. ob

$$H: \ \mathsf{P}(X = x_j^*, Y = y_l^*) \ = \ \mathsf{P}(X = x_j^*) \cdot \mathsf{P}(Y = y_l^*) \text{ für alle } j \text{ und } l$$

gegen

$$K \ : \ \mathsf{P}(X = x_{j'}^*, Y = y_{l'}^*) \neq \mathsf{P}(X = x_{j'}^*) \cdot \mathsf{P}(Y = y_{l'}^*)$$

für mindestens ein (j', l'), wobei $\mathsf{P}(X = j)$ und $\mathsf{P}(Y = l)$ die entsprechenden Randwahrscheinlichkeiten sind.

Unter H kann die Verteilung der Teststatistik

$$\boldsymbol{t \ = \ n \sum_{j=1}^{k} \sum_{l=1}^{m} \frac{(h_{jl} - \frac{h_{j.} h_{.l}}{n})^2}{h_{j.} h_{.l}}},$$

durch eine χ^2-Verteilung mit $(k-1)(m-1)$ Freiheitsgraden **approximiert** (**angenähert**) werden. Hierbei:

h_{jl} Anzahl aller Paare (x_i, y_i), die den Wert (x_j^*, y_l^*) annehmen;
$h_{j\cdot} = \sum_l h_{jl}$ Anzahl aller Paare (x_i, y_i), bei denen $x_i = x_j^*$;
$h_{\cdot l} = \sum_j h_{jl}$ Anzahl aller Paare (x_i, y_i), bei denen $y_i = y_l^*$.

Die Größen h_{jl} und $\frac{h_{j\cdot} h_{\cdot l}}{n}$ heißen die *"beoabachtete"* und die *"unter der Hypothese erwartete Häufigkeit"*.

Man lehnt H ab, wenn

$$\boldsymbol{t \geq \chi^2_{(k-1)(m-1);1-\alpha}} \text{ bzw. p-Wert } \boldsymbol{P = \mathsf{P}(T \geq t) < \alpha.}$$

Bemerkung: Für den Fall $k = m = 2$ gibt es den sogenannten exakten **Fisher-Test** , der bei einem Stichprobenumfang von $n < 20$ anzuwenden ist, allerdings sehr rechenintensiv ist.

Bei Durchführung des Chi-Quadrat-Tests bei einer Stichprobengröße von $20 \leq n \leq 200$ wird noch eine Korrekturgröße eingeführt, die die Approximationsgüte verbessert.

6 Lineare Regression

Zweidimensionale stetige Variable (X, Y) mit Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ vom Umfang n.

6.1 Streuungsdiagramm

Streuungsdiagramm bzw. **Scatterplot**:
Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ werden in ein Koordinatensystem eingetragen.

6.2 Empirische Kovarianz und Korrelation

Empirische Kovarianz:

$$s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y} \right) . \quad (6.1)$$

Empirischer Korrelationskoeffizient (Korrelationskoeffizient nach Pearson):

$$\begin{aligned} r_{xy} &= \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2 \right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n \cdot (\overline{y})^2 \right)}} . \end{aligned} \quad (6.2)$$

Der Korrelationskoeffizient nach Pearson besitzt die schöne Eigenschaft

$$-1 \leq r_{xy} \leq 1 .$$

Außerdem ändert der Korrelationskoeffizient nach Pearson seinen Wert **nicht**, wenn die Daten $x_1, ..., x_n$ und $y_1, ..., y_n$ linear transformiert werden!

Im Falle $r_{xy} \lesssim +1$ spricht man von einer **hohen positiven Korrelation** und im Fall $r_{xy} \gtrsim -1$ spricht man von **hoher negativer Korrelation**.

Im Falle einer hohen positiven Korrelation liegen die Datenpunkte im Streuungsdiagramm (sehr) nahe an einer steigenden Geraden, während die Datenpunkte im Falle einer hohen negativen Korrelation erneut (sehr) dicht an einer Geraden, diesmal aber an einer fallenden Geraden, liegen.

Wie steil die Gerade steigt oder fällt hat keinerlei Einfluss auf die Höhe der Korrelation. Die Höhe der Korrelation hängt einzig von der Güte der Beschreibung der Datenpunkte durch die Gerade ab.

Im Falle $r_{xy} \approx 0$ spricht man von (nahezu) **unkorrelierten** Datenpunkten.

Der Korrelationskoeffizient nach Pearson beschreibt die Stärke eines **linearen Zusammenhangs** zwischen zwei Variablen.

6.3 Empirische Kovarianz und Korrelation bei klassifizierten Daten

Diesmal seien die Merkmale X und Y nur in Klassen erfasst. Und zwar mit den Repräsentanten $x_1^*, \ldots, x_I^*$ bzw. $y_1^*, \ldots, y_J^*$. Die Ergebnisse der Stichprobe hält man vorteilhaft in einer sog. **Kontingenztafel** fest:

X \ Y	y_1^*	$\cdots$	y_j^*	$\cdots$	y_J^*	
x_1^*						
$\vdots$						
x_i^*			$r_{i,j}$			$r_{i,\bullet}$
$\vdots$						
x_I^*						
			$r_{\bullet,j}$			$r_{\bullet,\bullet} = 1$

Dabei bezeichnen

$r_{i,j}$: Anzahl der Stichprobenpaare mit erster Komponente gleich x_i^* und zweiter Komponente gleich y_j^* geteilt durch n

$r_{i,\bullet}$: Anzahl der Stichprobenpaare mit erster Komponente gleich x_i^* geteilt durch n

$r_{\bullet,j}$: Anzahl der Stichprobenpaare mit zweiter Komponente gleich y_j^* geteilt durch n

Es gilt:

$$r_{i,\bullet} = \sum_{j=1}^{J} r_{i,j} \quad \text{und} \quad r_{\bullet,j} = \sum_{i=1}^{I} r_{i,j}\,.$$

Korrelationskoeffizient nach Pearson für gruppierte Daten:

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_i^* - \overline{x})\cdot(y_j^* - \overline{y})\cdot r_{i,j}}{\sqrt{\left(\sum_{i=1}^{I}(x_i^* - \overline{x})^2 \cdot r_{i,\bullet}\right)\cdot\left(\sum_{j=1}^{J}(y_j^* - \overline{y})^2 \cdot r_{\bullet,j}\right)}} \\ &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} x_i^*\cdot y_j^*\cdot r_{i,j} - \overline{x}\cdot\overline{y}}{\sqrt{\left(\sum_{i=1}^{I}(x_i^*)^2 \cdot r_{i,\bullet} - (\overline{x})^2\right)\cdot\left(\sum_{j=1}^{J}(y_j^*)^2 \cdot r_{\bullet,j} - (\overline{y})^2\right)}} \end{aligned} \tag{6.3}$$

6.4 Regressionsgerade

Gemäß

$$\widehat{m} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{6.4}$$

$$\widehat{b} = \overline{y} - \widehat{m}\cdot\overline{x}\,. \tag{6.5}$$

erhält man die **Regressionsgerade**

$$m_{\text{Regression}}(x) = \widehat{m}\cdot x + \widehat{b} \tag{6.6}$$

Man beachte, dass die Regressionsgerade immer durch den Punkt $(\overline{x}, \overline{y})$ verläuft!

Die Formel für $\widehat{m}$ sieht sehr ähnlich zur Formel des Korrelationskoeffizienten r_{xy} aus, aber die beiden Formel sind nicht identisch! Man kann die beiden Werte wie folgt ineinander umrechnen:

$$\widehat{m} = r_{xy}\cdot\sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = r_{xy}\cdot\frac{s_y}{s_x} \tag{6.7}$$

$$r_{xy} = \widehat{m}\cdot\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \widehat{m}\cdot\frac{s_x}{s_y}\,. \tag{6.8}$$

(i) Im Falle **positiver** Korrelation der Daten, also wenn $r_{xy} > 0$, **steigt** die Regressionsgerade an, also $\widehat{m} > 0$.

(ii) Im Falle **negativer** Korrelation der Daten, also wenn $r_{xy} < 0$, **fällt** die Regressionsgerade ab, also $\widehat{m} < 0$.

(iii) Im Falle **unkorrelierter** Daten, also wenn $r_{xy} = 0$, verläuft die Regressionsgerade **waagerecht**, also $\widehat{m} = 0$.

(iv) Eine Maßzahl für die Güte der Anpassung der Daten an die Regressionsgerade wird durch das **Bestimmtheitsmaß** $R^2 = r_{xy}^2$ gegeben. R^2 liegt immer zwischen 0 und 1.

(v) Im Falle $R^2 \approx 1$ bzw. $r_{xy} \approx 1$ oder ≈ -1 liegen die Datenpunkte $(x_1, y_1), \ldots, (x_n, y_n)$ sehr dicht an der Regressionsgeraden. Die Anpassung durch die Regressionsgerade ist sehr gut.

(vi) Im Falle $R^2 = 1$ bzw. $r_{xy} = 1$ oder $= -1$ liegen alle Datenpunkte $(x_1, y_1), \ldots, (x_n, y_n)$ genau auf der Regressionsgeraden. Die Anpassung durch die Regressionsgerade ist perfekt.

(vii) Im Falle $R^2 \approx 0$ bzw. $r_{xy} \approx 0$ ist die Anpassung durch die Regressionsgerade nicht gut oder schlecht.

6.5 Vorhersage mit Hilfe der Regressionsgerade

Vorhersagen:

$$\widehat{y} = \widehat{m} \cdot x + \widehat{b}\,. \tag{6.9}$$

Vorsicht bei *out-of-sample* Vorhersagen!

Mit $\widehat{y}_i = \widehat{m} \cdot x_i + \widehat{b}, i = 1, \ldots, n$ gilt:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{s_{\widehat{y}}^2}{s_y^2}\,. \tag{6.10}$$

Bestimmtheitsmaß R^2: Anteil an der Gesamtvarianz s_y^2 an, der durch die Regression erklärt werden kann.

6.6 Regressionsgerade bei gruppierten Daten

Regressionsgerade $m_{Regression}(x) = \widehat{m} \cdot x + \widehat{b}$ aus gruppierten Daten:

X \ Y	y_1^*	$\cdots$	y_J^*	Summe
x_1^*	$r_{1,1}$	$\cdots$	$r_{1,J}$	$r_{1,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_I^*	$r_{I,1}$	$\cdots$	$r_{I,J}$	$r_{I,\bullet}$
Summe	$r_{\bullet,1}$	$\cdots$	$r_{\bullet,J}$	$r_{\bullet,\bullet} = 1$

$$
\begin{aligned}
\overline{x}_{gruppiert} &= \sum_{i=1}^{I} x_i^* \cdot r_{i,\bullet} \\
\overline{y}_{gruppiert} &= \sum_{j=1}^{J} y_j^* \cdot r_{\bullet,j} \\
s^2_{x,gruppiert} &= \sum_{i=1}^{I} (x_i^*)^2 \cdot r_{i,\bullet} - (\overline{x}_{gruppiert})^2 \\
s^2_{y,gruppiert} &= \sum_{j=1}^{J} (y_j^*)^2 \cdot r_{\bullet,j} - (\overline{y}_{gruppiert})^2 \qquad (6.11) \\
s_{xy,gruppiert} &= \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^* \cdot y_j^* \cdot r_{i,j} - \overline{x}_{gruppiert} \cdot \overline{y}_{gruppiert} \\
r_{xy,gruppiert} &= \frac{s_{xy,gruppiert}}{s_{x,gruppiert} \cdot s_{y,gruppiert}} \\
\widehat{m} &= \frac{s_{xy,gruppiert}}{s^2_{x,gruppiert}} = r_{XY,gruppiert} \cdot \frac{s_{y,gruppiert}}{s_{x,gruppiert}} \\
\widehat{b} &= \overline{y}_{gruppiert} - \widehat{m} \cdot \overline{x}_{gruppiert} \, .
\end{aligned}
$$

6.7 Rangkorrelation

Rangkorrelationskoeffizient nach Spearman:

$$r_{xy}^{Rang} = \frac{\sum_{i=1}^{n}(R_i - (n+1)/2) \cdot (R_i' - (n+1)/2)}{\sqrt{\sum_{i=1}^{n}(R_i - (n+1)/2)^2 \cdot \sum_{j=1}^{n}(R_j' - (n+1)/2)^2}} \tag{6.12}$$

Stets gilt:

$$-1 \leq r_{xy}^{Rang} \leq 1\,.$$

Falls keine Bindungen vorliegen:

$$r_{xy}^{Rang} = 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_i - R_i')^2}{n \cdot (n^2 - 1)}\,. \tag{6.13}$$

Falls r_{xy}^{Rang} nahe bei $+1$, dann stimmen die Beurteilungen sehr gut überein gleich, d.h. $R_i = R_i'$.
Falls r_{xy}^{Rang} nahe bei -1, dann sind die Beurteilungenstark gegensätzlich, d.h. $R_i = n - R_i' + 1$.

Der Rangkorrelationskoeffizient beschreibt die Stärke eines **monotonen Zusammenhangs** zwischen zwei Variablen.