Formelsammlung Statistik

Version: 6. September 2013

Definition 1. *„**statistische Variable**“: Sei eine Grundmenge G gegeben: Jede Zerlegung (= Partition) Z der Grundmenge G nennen wir „(statistische) Variable“. Die (nichtleeren) Elemente (= Äquivalenzklassen) von Z nennen wir „Ausprägungen der Variable“ oder „Werte der Variablen“ (die Elemente von Z sind Mengen).*

Skalen:

Eine Abbildung $f : G \to \mathbb{R}$ ist eine ..... Skala auf die Grundgesamtheit G bezüglich der Variable V genau dann, wenn es Relationen T und $\circ$ auf G gibt, so dass:	
(1) $(x, y) \in R_V$ genau dann, wenn $f(x) = f(y)$ (R_V ist die durch die Variabale V festgelegte Äquivalenzrelation auf G).	**nominale**
(2) T ist transitiv und asymmetrisch. (3)$(x, y) \in T$ genau dann, wenn $f(x) > f(y)$. (4) $(x, y) \in T$ oder $(y, x) \in T$ genau dann, wenn $f(x) \neq f(y)$.	**ordinale**
(5) $f(x \circ y) = f(x) + f(y)$ (Additionsprinzip oder Linearität).	**metrische**

Tabelle 1: Zusammenstellung über den Zusammenhang zwischen den drei behandelten Skalen

Häufigkeitesverteilungen und Verteilungsfunktion

Definition 2. *: k ist eine absolute Häufigkeitsverteilung einer Grundmenge G bezüglich der f-skalierten Variable V genau dann, wenn*
$k : Bild(f) \to \mathbb{N}$
$k(x) := |\{z \mid f(z) = x \text{ und } z \in G\}|$
($Bild(f)$ = Bild der Abbildung f):

Definition 3. *h ist eine relative Häufigkeitsverteilung der f-skalierten Variable V genau dann, wenn*
$h : Bild(f) \to [0, 1]$
$h(z) := \frac{k(z)}{n}$

Definition 4. *k_k ist eine absolute, kumulative Häufigkeitsverteilung einer Grundmenge G bezüglich der f-skalierten Variable V genau dann, wenn*
$k_k : Bild(f) \to$
$k_k(x_i) := \sum_{x_j \leq x_i} k(x_j)$ *(für $x_j, x_i \in Bild(f)$)*

Definition 5. *h_k ist eine relative, kumulative Häufigkeitsverteilung einer Grundmenge G bezüglich der f-skalierten Variable V genau dann, wenn*

$h_k : Bild(f) \to \mathbb{N}$ $(n = |G|)$
$h_k(x_i) := \frac{\sum\limits_{x_j \le x_i} k(x_j)}{n}$ $\quad$ $(für\ x_j, x_i \in Bild(f))$

Definition 6. *F ist eine (empirische) Verteilungsfunktion (Punktform) genau dann, wenn*
$F : Bild(f) \to [0,1]$
$F(x_i) := \frac{|\{x_j:\ x_j \le x_i\ und\ x_j \in Bild(f)\}|}{|G|}$

Kennwerte der mittleren Lage

Definition 7. *(arithmetisches Mittel):* $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ *für den Datensatz* $x = (x_1, ..., x_n)$

Definition 8. *(Median):* $\tilde{x} = \begin{cases} x_k & für\ k = (n+1)\cdot 0.5 \\ x_k + (x_{k+1} - x_k)\cdot 0.5 & sonst \end{cases}$
$x_k = k-tes$ *Datum des geordneten Datensatzes* x.
$k =$ *Zahl vor dem Komma der Zahl* $(n+1)\cdot 0.5$.

Definition 9. *(Median für klassierte Daten):* $\tilde{x} = u + \frac{\left(\frac{n+1}{2} - m\right)}{l} k$
$k =$ *die Intervalllänge des Intervalls* K*, in das der Median zu liegen kommt.*
m *die Zahl der Daten, die in den tiefer als* K *liegenden Klassen zu liegen kommen*
l *die Anzahl der Daten des Intervalls, in das der Median zu liegen kommt und*
u *die untere Grenze des Intervalls* K.

Definition 10. *(Modus):* $modalwert(x) = x_i$ *genau dann, wenn* $k(x_i) = \max\{k(x_j) : x_j \in Bild(f)\}$ *(*max *ist nur bestimmt, wenn es genau einen Wert gibt, der in der Menge maximal ist)*

Streuungskennwerte

Definition 11. *Spannweite der Messwerte* $x = (x_1, ..., x_n)$ *eines Datensatzes* $:= \max(x) - \min(x)$

Definition 12. *Durchschnittlicher Abweichungsbetrag* $= \frac{1}{n}\sum\limits_{i=1}^{n} |x_i - \bar{x}|$

Definition 13. *Varianz* $\sigma^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n} (x_i - \bar{x})^2$

Definition 14. *Standardabweichung* $\quad \sigma = \sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$

Definition 15. *Variationskoeffizient:* $V = \frac{\sigma}{\bar{x}}$.

Definition 16. *standardisiertes Datum:* $z_i = \frac{x_i - \bar{x}}{\sigma_x}$

Definition 17. *Lorenzkurve: Die Punkte* $\left(\frac{i}{n}, a(\frac{i}{n})\right)$ *und* $\left(\frac{i+1}{n}, a(\frac{i+1}{n})\right)$ *(für* $0 \le i \le n-1$*) werden jeweils mit einer Strecke verbunden.*

$\frac{i}{n}$ *= Anzahl Objekte mit weniger oder gleichviel von der betrachteten Grösse;* $a(\frac{i}{n}) := \frac{\sum\limits_{x_j \le x_i} x_j}{\sum\limits_{i=1}^{n} x_i}$.

Definition 18. *Sei* x *ein geordneter Datenvektor. Dann definieren wir:*
$i - Quantil(x) := \begin{cases} x_j\ für\ j = (n-1)i + 1 \\ x_j + r(x_{j+1} - x_j)\ für\ j < (n-1)i + 1 < j + 1 \end{cases}$
mit:
$x_j = j-tes$ *Datum des geordneten Datensatzes* x.
$r = (n-1)i + 1 - j$ *(= Zahl hinter dem Komma der Zahl* $(n-1)i + 1$*).*

Definition 19. *Der Quartilsabstand = 0.75-Quantil − 0.25-Quantil (= 3. Quartil - 1. Quartil). (für $x = (x_1, ..., x_n)$)*

Definition 20. *Simpsonindex: :* $D = \frac{m}{m-1}\left(1 - \sum_{i=1}^{m} (h(x_i))^2\right)$
$h(x_i)$ = relativen Häufigkeit des Skalenwertes x_i;
m = Anzahl Ausprägungen.

Definition 21. *Schiefekennwert:* $g_m = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{\sigma})^3$

Indizes

Definition 22. *Wertindex:* $W = \frac{\sum_{i=1}^{n} p_{t_i} m_{t_i}}{\sum_{i=1}^{n} p_{0_i} m_{0_i}}$

Definition 23. *Preisindex:* $P_p = \frac{\sum_{i=1}^{n} p_{t_i} m_{t_i}}{\sum_{i=1}^{n} p_{0_i} m_{t_i}}; P_L = \frac{\sum_{i=1}^{n} p_{t_i} m_{0_i}}{\sum_{i=1}^{n} p_{0_i} m_{0_i}}$

Definition 24. *Mengenindex:* $M_p = \frac{\sum_{i=1}^{n} p_{t_i} m_{t_i}}{\sum_{i=1}^{n} p_{t_i} m_{0_i}}; M_L = \frac{\sum_{i=1}^{n} p_{0_i} m_{t_i}}{\sum_{i=1}^{n} p_{0_i} m_{0_i}}$

Bivariate beschreibende Statistik

Definition 25. *Sei f die Skala der Variable X auf die Grundmenge G und g die Skala der Variable Y auf die Grundmenge G, und $Bild(f)$ das Bild von f und $Bild(g)$ das Bild von g. Die gemeinsame Häufigkeitsverteilung k der Variablen X und Y ist die Abbildung*

$$k_{X,Y} : Bild(f) \times Bild(g) \to \mathbb{N}$$
$$k_{ij} := k_{X,Y}((x_i, y_j)) := |\{u : f(u) = x_i \text{ und } g(u) = y_j \text{ und } u \in G\}|$$

Definition 26. *Chancen-Verhältnis (Odds-Ratio; OR)*

$$OR = \frac{k_{11}k_{22}}{k_{12}k_{21}}$$

oder

$$OR = \frac{(k_{11} + 0.5)(k_{22} + 0.5)}{(k_{12} + 0.5)(k_{21} + 0.5)}$$

Definition 27. *Kontingenz-Koeffizient*

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Korrigierter Kontingenz-Koeffizient

$$K^* = \frac{1}{\sqrt{\frac{\min\{I,J\}-1}{\min\{I,J\}}}} K$$

mit $\chi^2 := \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(k_{ij} - k_{ij}^e)^2}{k_{ij}^e}$; *$k_{ij}$ absolute Häufigkeiten der Zelle i,j der Kreuztabelle;* $k_{ij}^e = \frac{k_{i+}k_{+j}}{n}$; *$k_{i+}$ Summe der $i-$ten Zeile; k_{+j} Summe der $j-$ten Spalten; n Anzahl Daten; I Anzahl Zeilen der Kreuztabelle; J Anzahl Spalten der Kreutztabellen.*

Definition 28. *Gamma:*

$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}$
$\Pi_c = \sum_{i=1}^{I} \sum_{j=1}^{J} H_{ij} \left(\sum_{I \geq h > i} \sum_{J \geq k > j} H_{hk} \right)$
$\Pi_d = \sum_{i=1}^{I} \sum_{j=1}^{J} H_{ij} \left(\sum_{I \geq h > i} \sum_{1 \leq k < j} H_{hk} \right)$

Definition 29. *Pearson-Korrelationskoeffizient:* $r_{xy} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$

Definition 30. *Kovarianz:* $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Definition 31. *Spearmankorrelation* $r_S = \frac{1}{n-1} \sum_{i=1}^{n} (r_i - \bar{x})(k_i - \bar{y})$

r_i Rang des $i-$ten Objektes bezüglich der Variable X. k_i Rang des i-ten Objektes bezüglich der Variable Y. $\bar{x} = \bar{y} = \frac{n(n+1)}{2}$ (= Rangmittel).
(bei gleichen Ausprägungen wird das jeweilige arithmetische Mittel der Durchnummerierungen zugeordnet)

Satz 32. *Lineare Regression: Für $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n)$ mit mindestes einen Datenpaar $x_i \neq x_j$, sowie $f(x) = ax + b$ ist*

$$S : \mathbb{R}^2 \to \mathbb{R}$$
$$S(a, b) := \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

minimal genau dann, wenn

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$
$$b = \bar{y} - a\bar{x}$$

Satz 33. *Für den Regressionssteigungskoeffizienten a, die Pearson-Korrelation r, die Standardabweichung s_x der x-Daten und die Standardabweichung s_y der y-Daten gilt:*

$$a = r\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

Satz 34. *$SQT = SQR + SQE$ mit*

$$SQR := \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

$$SQE := \sum_{i=1}^{n} (f(x_i) - \bar{y})^2.$$

$$SQT := \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Satz 35.

$$R^2 := \frac{SQE}{SQT}$$

(die Spearmankorrelation $r = \sqrt{R^2}$*)*

Definition 36. *Partielle Korrelation* $r_{xy/z} = r_{r_x r_y}$ *mit den Residuen* $r_{x_i} = x_i - f(z_i)$ *und* $r_{y_i} = y_i - g(z_i)$*, wobei* f *die Regressionsgerade des Datenvektors* x *in Abhängigkeit vom Datenvektor* z *ist und* g *die Regressionsgerade des Datenvektors* y *in Abhängigkeit vom Datenvektor* z *ist (d.h.* $r_{r_x r_y}$ *ist die Korrelation zwischen den Residuenvektoren* r_x *und* r_y*).*

Satz 37. $r_{xy/z} = \frac{r_{xy} \quad - \quad (r_{xz} \; r_{yz})}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$

Kombinatorik

1. Ziehen mit Zurücklegen in Reihenfolge (n Anzahl Objekte Urne; m Anzahl gezogene Objekte):

   $$\overline{V}(n,m) := n^m$$

2. Ziehen ohne Zurücklegen in Reihenfolge (n Anzahl Objekte Urne; m Anzahl gezogene Objekte): mit

   $$n! := \begin{cases} \prod_{i=0}^{n-1} (n-i) & \text{für } n \geq 1 \\ 1 & \text{für } n = 0 \end{cases}$$

   und $n \in \mathbb{N}^0$

   $$V(n,m) = \frac{n!}{(n-m)!}$$

3. Ziehen ohne Zurücklegen ohne Reihenfolge (n Anzahl Objekte Urne; m Anzahl gezogene Objekte):

   $$K(n,m) = \frac{n!}{m!(n-m)!} =: \binom{n}{m}$$

   (für $m, n \in \mathbb{N}$)

4. Ziehen mit Zürücklegen mit Reihenfolge (n Anzahl Objekte Urne; m Anzahl gezogene Objekte): (für $m, n \in \mathbb{N}$)

   $$\overline{K}(n,m) = \binom{n+m-1}{m}$$

Wahrscheinlichkeitstheorie

Definition 38. *Ergebnis und Ergebnismenge: Die Ausgänge, die beim Ausführen eines Zufallsexperimentes auftreten können, heissen Ergebnisse. Die Menge aller Ergebnisse heisst Ergebnismenge* Ω*. Ein abzählbare (endlich oder nicht) Ergebnismenge heisst „diskret".*

Definition 39. ***Ereignis****: E ist ein Ereignis genau dann, wenn* $E \in \mathcal{P}(\Omega)$ *(*$\mathcal{P}(\Omega)$ *ist die Potenzmenge der diskreten Ergebnismenge* Ω*, d.h. die Menge aller Teilmengen von* Ω*).*

Definition 40. ***"Ereignis trifft ein":*** *Ein Ereignis triff genau dann ein, wenn eines seiner Elemente (= Ergebnisse) eintrifft. So trifft das Ereignis, eine gerade Zahl zu würfeln ein, wenn man eine 2, eine 4 oder eine 6 würfelt.*

Definition 41. ***Wahrscheinlichkeitsfunktion:*** *Sei Ω eine diskrete Ergebnismenge.*
f ist eine Wahrscheinlichkeitsfunktion auf Ω, genau dann, wenn gilt:
$f : \Omega \to [0,1]$ *und* $\sum_{\omega \in \Omega} f(\omega) = 1$.

Definition 42. ***Diskrete Wahrscheinlichkeitsverteilung*** P ***auf*** Ω*:*
$P : \mathcal{P}(\Omega) \to [0,1]$
$P(E) := \sum_{\omega \in E} f(\omega)$
(für $E \subset \Omega$; Ω ist ein diskreter Stichprobenraum,
$\mathcal{P}(\Omega)$ ist der Ereignisraum von Ω = Potenzmenge von Ω)

Definition 43. ***diskreter Wahrscheinlichkeitsraum:*** *Das Tripel $(\Omega, \mathcal{P}(\Omega), P)$ wird diskreter Wahrscheinlichkeitsraum genannt (Ω = diskrete Ergebnismenge, $\mathcal{P}(\Omega)$ = Potenzmenge von Ω, P Wahrscheinlichkeitsverteilung auf Ω).*

Satz 44. ***Additionssatz:*** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Definition 45. *Seien n Stichprobenräume Ω_i gegeben. Dann nennen wir ein beliebiges Wahrscheinlichkeitsmass auf $\mathcal{P}(\Omega_1 \times ... \times \Omega_n)$ eine gemeinsame Verteilung von $\Omega_1, ..., \Omega_n$.*

Definition 46. *Seien zwei diskrete Ergebnismengen Ω_1 und Ω_2 mit zwei Wahrscheinlichkeitsfunktionen f_1 und f_2 gegeben. Wir nennen die gemeinsame Wahrscheinlichkeitsfunktion $f_{\Omega_1 \times \Omega_2}$ mit*

$$f_{\Omega_1 \times \Omega_2} : \Omega_1 \times \Omega_2 \to [0,1]$$
$$f_{\Omega_1 \times \Omega_2}((\omega_i, \omega_j)) = f_1(\omega_i) \cdot f_2(\omega_j), \omega_i \in \Omega_1, \omega_j \in \Omega_2$$

die Produktewahrscheinlichkeitsfunktion auf $\Omega_1 \times \Omega_2$.
(Diese Namengebung gilt auch für die durch Multiplizieren der Wahrscheinlichkeiten erzeugte Wahrscheinlichkeitsfunktion auf das Produkt von n Ergebnismengen)

Definition 47. *Seien zwei Stichprobenräume Ω_1 und Ω_2 mit den zwei Wahrscheinlichkeitsfunktionen f_1 und f_2 gegeben. Die durch*

$$f : \Omega_1 \times \Omega_2 \to [0,1]$$
$$f((\omega_i, \omega_j)) := f_1(\omega_i) \cdot f_2(\omega_j)$$

definierte Wahrscheinlichkeitsverteilung nennen wir das Produktesmass von P_1 und P_2. (Definition für $n \in \mathbb{N}_2^*$;Analog für $n > 2$).

Satz 48. *Sei das Produktemass $P_{\Omega_1 \times \Omega_2}$ von P_1 und P_2 auf die diskrete Ergebnismenge $\Omega_1 \times \Omega_2$ gegeben und $A \subset \Omega_1$ sowie $B \subset \Omega_2$. Dann gilt: $P_{\Omega_1 \times \Omega_2}(A \times B) = P_{\Omega_1}(A) \cdot P_{\Omega_2}(B)$*

Definition 49. *Ereignisse A und B beeinflussen sich gegenseitig nicht genau dann, wenn $P_{\Omega_1 \times \Omega_2}(A \times B) = P_{\Omega_1}(A) \cdot P_{\Omega_2}(B)$.*

Definition 50. $P(E_2 \mid E_1) := \frac{P(E_2 \cap E_1)}{P(E_1)}$ *(bedingte Wahrscheinlichkeit)*

Satz 51. *(Satz von der totalen Wahrscheinlichkeit): Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $(A_1, ..., A_n)$ eine Zerlegung von Ω und $B \subset \Omega$, dann gilt:*

$$P(B) = \sum_{i=1}^{n} (P(B \mid A_i) \cdot P(A_i))$$

Satz 52. *(Satz von Bayes) Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum, $(A_1, ..., A_n)$ eine Zerlegung von Ω und $B \subset \Omega$, $P(A_i) > 0$ und $P(B \mid A_i) > 0$ für mindestens ein i, dann gilt:*

$$P(A_j \mid B) = \frac{P(B \mid A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)} = \frac{P(B \mid A_j) \cdot P(A_j)}{P(B)}$$

Satz 53. *Ist* $P(A_2 \mid A_1) = P(A_2)$*, dann gilt* $P(A_2)P(A_1) = P(A_2 \cap A_1)$.

Definition 54. *Sei der diskrete Wahrscheinlichkeitsraum* $(\Omega, \mathcal{P}(\Omega), P)$ *gegeben. Für Ereignisse* $C, D \subset \Omega$ *gilt: Ereignisse* C *und* D *sind unabhängig bezüglich des Verteilung* P *genau dann, wenn* $P(C) \cdot P(D) = P(C \cap D)$

Definition 55. X *ist eine Zufallsvariable, wenn* X *eine Abbildung ist mit:* $X : \Omega \to \mathbb{R}$
($(\Omega, \mathcal{P}(\Omega), P)$ *ist ein diskreter Wahrscheinlichkeitsraum)*

Definition 56. *- Der Wert* $X(\omega)$ *einer Zufallsvariable* X *realisiert sich (= tritt ein), wenn* ω *eintritt.*
- Sei $\mathfrak{X} := B(X)$ *das Bild der Zufallsvariable* X *und seien Ereignisse aus* $\mathfrak{X}$ *als Teilmengen von* $\mathfrak{X}$ *definiert: Ein Ereignis* E *aus* $\mathfrak{X}$ *tritt ein, wenn es ein eintretendes* $\omega \in \Omega$ *gibt, so dass* $X(\omega) \in E$.

Definition 57. *Sei* $(\Omega, \mathcal{P}(\Omega), P)$ *ein diskreter Wahrscheinlichkeitsraum.*
Sei X *eine Zufallsvariable* $X : \Omega \to \mathbb{R}$ *und* $\mathfrak{X}$ *das Bild von* X. $f : \Omega \to [0, 1]$ *sei die durch* P *festgelegte Wahrscheinlichkeitsfunktion auf* Ω. *Dann definieren wir:*
$f_X : \mathfrak{X} \to [0, 1]$
$f_X(x) := \sum\limits_{X(\omega)=x} f(\omega)$. *(für* $\omega \in \Omega$*).*

Definition 58. *Sei* $(\Omega, \mathcal{P}(\Omega), P)$ *ein diskreter Wahrscheinlichkeitsraum.*
Sei X *eine Zufallsvariable* $X : \Omega \to \mathbb{R}$ *und* $\mathfrak{X}$ *das Bild von* X. $f_X : \mathfrak{X} \to [0, 1]$ *die von* P *bestimmte Wahrscheinlichkeitsfunktion auf das Bild von* X.
Dann definieren wir:
$P_X : \mathcal{P}(\mathfrak{X}) \to [0, 1]$ *(*$\mathcal{P}(\mathfrak{X})$ *ist die Potenzmenge des Bildes von* X*)*
$P_X(E) := \sum\limits_{x \in E} f_X(x)$ *(für* $E \in \mathcal{P}(\mathfrak{X})$*)*

Definition 59. *Sei* $(\Omega, \mathcal{S}, P)$ *ein beliebiger Wahrscheinlichkeitsraum.*
Sei X *eine Zufallsvariable* $X : \Omega \to \mathbb{R}$ *und* $E \subset \mathfrak{X}$ *(= Bild von* X*) und* $a, b \in \mathbb{R}$.
$P(X \in E) := P_X(E)$ *(= Die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert aus* E *annimmt).*
$P(a \leq X \leq b) := P(X \in [a, b] \cap \mathfrak{X})$
$P(X \leq b) := P(X \in \ ]-\infty, b] \cap \mathfrak{X})$
$P(X \geq b) := P(X \in [b, \infty[\ \cap \mathfrak{X})$
$P(X = a) := P(X \in \{a\})$ *(etc. d.h. analog z.B. für* $<, >$*)*

Definition 60. *Verteilungsfunktion: Eine Funktion*
$F : \mathfrak{X} \to [0, 1]$
$F(x) := P(X \leq x)$
wird „Verteilungsfunktion der Zufallsvariable X *unter* P*“ genannt.*

Definition 61. *Sei* $(\Omega, \mathcal{P}(\Omega), P)$ *ein diskreter Wahrscheinlichkeitsraum und* X *eine Zufallsvariable auf* Ω *und* $\mathfrak{X}$ *(= Bild von* Ω *unter* X*). Der Erwartungswert* $E(X)$ *der Zufallsvariable* X *unter der diskreten Wahrscheinlichkeitsverteilung* P *ist definiert als:*
$E(X) = \sum\limits_{x \in \mathfrak{X}} [P(X = x) \cdot x]$.
(sofern $\sum\limits_{x \in \mathfrak{X}} [P(X = x) \cdot x]$ *definiert ist. (Weitere Symbolisierung:* $E_P(X)$*).*

Satz 62. $E(a) = a$ *(*a *ist eine beliebige reelle Zahl).*

Satz 63. $E(E(X)) = E(X)$

Satz 64. $E(bX) = bE(X)$ *(*$b \in \mathbb{R}$*)*

Satz 65. $E(XE(Y)) = E(Y)E(X)$

Definition 66. *Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein diskreter Wahrscheinlichkeitsraum und X eine Zufallsvariable auf Ω und $\mathfrak{X}$ (= Bild von Ω unter X). Die Varianz der Zufallsvariable X unter der diskreten Wahrscheinlichkeitsverteilung P ist definiert als: $V(X) = \sum_{x \in \mathfrak{X}} P(X = x)(x - E(X))^2$.*
(Sofern $\sum_{x \in \mathfrak{X}} P(X = x)(x - E(X))^2$ definiert ist).

Satz 67. $V(X) = E\left[(X - E(X))^2\right]$

Definition 68. *$P_{(X_1,X_2)}$ bezeichnet ein beliebiges Wahrscheinlichkeitsmass auf $\mathcal{P}(\mathfrak{X} \times \mathfrak{X})$.*
$P_{(X_1,X_2)}$ wird „gemeinsame Verteilung von X_1 und X_2"genannt.

Definition 69. *Seien $(\Omega, \mathcal{P}(\Omega), P)$ ein diskreter Wahrscheinlichkeitsraum und seien X_i Zufallsvariablen auf Ω und $\mathfrak{X}$ die Bilder von X_i $(i \in \mathbb{N}_2)$.*
Sei $f_{(X_1,X_2)}$ die durch P auf $\mathcal{P}(\Omega \times \Omega)$ bestimmte Wahrscheinlichkeitsfunktion auf $\mathfrak{X} \times \mathfrak{X}$. Dann halten wir fest:
$P_{(X_1,X_2)}$ ist mit dem Produktemass auf $\mathcal{P}(\mathfrak{X} \times \mathfrak{X})$ identisch, genau dann wenn $f_{(X_1,X_2)}((x_i, x_j)) = f_{X_1}(x_i) \cdot f_{X_2}(x_j)$ (für alle $(x_i, x_j) \in \mathfrak{X} \times \mathfrak{X}$))
$P_{(X_1,X_2)}$ wird in diesem Fall „Produktemass von X_1 und X_2"genannt.

Definition 70. *Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein diskreter Wahrscheinlichkeitsraum und X_i Zufallsvariablen auf Ω und $\mathfrak{X}$ die Bilder von X_i $(i \in \mathbb{N}_2)$. Dann definieren wir:*
Die Zufallsvariablen X_1 und X_2 sind unabhängig bezüglich P genau dann, wenn für alle $A \subset \mathfrak{X}$ und $B \subset \mathfrak{X}$ gilt: $P_{(X_1,X_2)}(A \times B) = P_{X_1}(A) \cdot P_{X_2}(B)$.

Definition 71. *Eine Zufallsvariable T, die auf das kartesiche Produkt von Bildern $\mathfrak{X}_i$ von Zufallsvariabalen $X_1, ..., X_n$ definiert ist, nennt man „Statistik".*

Übersicht **Statistiken:**

Ω^n	$\xrightarrow{(X_1,...,X_n)}$ $\overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i$	$\xrightarrow{T}$ $\mathbb{R}$
P	$P_{(X_1,...X_n)}$	P_T
Ω = Stichprobenraum der einzelnen n Zufallsexperimente		
$(X_1, ..., X_n) = n-$Tupel von Zufallsvariablen $X_i : \Omega \to \mathbb{R}$		
$\overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i$ = Bild von X		
T = Zufallsvariable auf $\overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i$ (= Statistik auf $\overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i$)		
P = Wahrscheinlichkeitsmass auf $\mathcal{P}(\Omega)$.		
$P_{(X_1,...X_n)}$ = gemeinsame Verteilung auf $\mathcal{P}\left(\overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i\right)$		
(= Produktmass bei unabhängigen X_i, „$\mathcal{P}$" für Potenzmenge)		
P_T Bildmass von $P_{(X_1,...X_n)}$ unter T.		

Definition 72. *Seien X_i Zufallsvariablen. $\mathfrak{X}$ die Bilder dieser Zufallsvariablen. Dann schreiben wir künftig für*
$T : \overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i \to \mathbb{R}$
$T((x_1, ..., x_n) = \sum_{i=1}^{n} x_i :$
$\sum_{i=1}^{n} X_i := T.$
Mit $x := (x_1, ..., x_n)$ schreiben wir auch: $\sum_{i=1}^{n} X_i(x) := \sum_{i=1}^{n} X_i((x_1, ..., x_n)) = \sum_{i=1}^{n} x_i$.

Satz 73. *Seien X_i unabhängige Zufallsvariablen auf Ω. Dann gilt $E_{P_{(X_1,...,X_n)}}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E_P(X_i)$.*

Satz 74. $E(a+bX) = a + bE(X) \qquad (a, b \in \mathbb{R}, b \neq 0)$

Definition 75. *Der Mittelwert $\bar{X}_n$ von n Zufallsvariablen X_i wird definiert als:*
$\bar{X}_n : \times_{i=1}^{n} \mathfrak{X}_i \to \mathbb{R}$
$\bar{X}_n(x) := \frac{1}{n} \sum_{i=1}^{n} x_i \quad$ *(für $x = (x_1, ..., x_n)$)*

Satz 76. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

Definition 77. *Statt „n unabhängig und identisch verteilte Zufallsvariablen" verwenden wir künftig auch „n i.i.d.-Zufallsvariablen" (i.i.d. = independant and identically distributed).*

Definition 78. *Seien X_i i.i.d. Dann gilt:* $E_{P_{(X_1,...,X_n)}}(\sum_{i=1}^{n} X_i) = nE_P(X_i)$

Satz 79. *Für n i.i.d. X_i gilt:* $E_{P_{(X_1,...,X_n)}}(\bar{X}_n) = E_P(X_i)$

Satz 80. $V(X) = E(X^2) - [E(X)]^2$

Satz 81. $V(a) = 0 \qquad (a \in \mathbb{R})$

Satz 82. $V(bX) = b^2 V(X) \qquad (b \in \mathbb{R})$

Satz 83. $V(a + bX) = b^2 V(X) \qquad (b, a \in \mathbb{R}, b \neq 0)$

Definition 84. $\prod_{i=1}^{n} X_i : \overset{n}{\underset{i=1}{\times}} \mathfrak{X}_i \to \mathbb{R}$
$\prod_{i=1}^{n} X_i(x) := \prod_{i=1}^{n} x_i \quad (mit\ x = (x_1, ..., x_n))$

Definition 85. $KOV(X_1, X_2) := E_{P_{(X_1,X_2)}}((X_1 - E_P(X_1) \cdot (X_2 - E_P(X_2))$

Satz 86. $KOV(X_1, X_2) = E_{P_{(X_1,X_2)}}(X_1 \cdot X_2) - E_P(X_1)E_P(X_2)$

Satz 87. $KOV(X_1, X_1) = V(X_1)$

Satz 88. $KOV(X_1, X_2) = 0$ *genau dann, wenn* $E_{P_{(X_1,X_2)}}(X_1X_2) = E_P(X_1)E_P(X_2)$

Satz 89. $V_{P_{(X_1,X_2)}}(X_1 + X_2) = V_P(X_1) + V_P(X_2) + KOV(X_1, X_2)$

Definition 90. X_1 *und* X_2 *sind unkorreliert genau dann, wenn* $KOV(X_1, X_2) = 0$.

Satz 91. *Wenn X_1 und X_2 unkorreliert sind, dann* $V_{P_{(X_1,X_2)}}(X_1 + X_2) = V_P(X_1) + V_P(X_2)$

Satz 92. *Wenn X_1 und X_2 unabhängig sind, dann gilt:* $E_{P_{(X_1,X_2)}}(X_1X_2) = E_P(X_1)E_P(X_2)$

Satz 93. *Wenn X_1 und X_2 unabhängig sind, dann sind X_1 und X_2 unkorreliert.*

Satz 94. *Für die Summe von n i.i.d.Zufallsvariablen X_i gilt:* $V_{P_{(X_1,...,X_n)}}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V_P(X_i) = nV_P(X_i)$

Satz 95. *Seien X_i i.i.d., dann gilt:* $V_{P_{(X_1,...,X_n)}}(\bar{X}_n) = \frac{V_P(X_i)}{n}$

Definition 96. *Sei $(X_1, ..., X_n)$ eine Folge von Zufallsvariablen und $\bar{X}_n$ die Statistik Mittelwert auf das kartesische Produkt der Bilder von X_i.*
Wenn für beliebige $k > 0$ gilt: $P(\{x : \left|\bar{X}_n(x) - E(X_i)\right| > k\}) = 0$ *für $n \to \infty$, dann gilt für die Folge von Zufallsvariablen X_i das schwache Gesetz der grossen Zahlen.*

Satz 97. *Für eine Folge $(X_1, ..., X_n)$ von i.i.d Zufallsvarablen X_i mit $V(X_i) < \infty$ gilt das schwache Gesetz der grossen Zahlen.*

Definition 98. *Die Zufallsvariable X ist bernoulliverteilt genau dann, wenn*

$$\mathfrak{X} = \{0, 1\}$$
$$P_X(\{1\}) = p$$
$$P_X(\{0\}) = 1 - p$$

$(0 \leq p \leq 1)$
$\mathfrak{X}$ = *Bild von* X; p *ist der Parameter der Verteilungfamilie)*
$X \sim B(1, p)$ *ist eine Abkürzung für „die Zufallsvariable* X *ist bernoulliverteilt mit dem Parameter* p*“.*

Satz 99. *Für* $X \sim B(1, p)$ *gilt:* $E(X) = p$ *und* $V(X) = p(1 - p)$

Definition 100. X *ist eine binomialverteilte Zufallsvariable genau dann, wenn*

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$(x \in \mathbb{N}_n)$.
(Abkürzung: $X \sim B(n, p)$ *für „die Zufallsvariable* X *ist binomialverteilt“)*
n *und* p *werden die Parameter der Binomialverteilung genannt. Durch sie ist eine Binomialverteilung bestimmt.*

Satz 101. *Für* $X \sim B(n, p)$ *gilt* $E(X) = np$ *und* $V(x) = np(1 - p)$

Definition 102. *Eine Zufallsvariable* X *ist poisson-verteilt genau dann, wenn*

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$(x \in \mathbb{N},\ \lambda > 0)$.
(Abkürzung: $X \sim P(\lambda)$*)*
(„λ*“ wird als „Lamda“ gelesen.* λ *ist der Parameter der Verteilung).*

Satz 103. *Für* $X \sim P(\lambda)$ *gilt:* $E(X) = \lambda$ *und* $V(X) = \lambda$.

Definition 104. X *ist eine hypergeometrisch verteilte Zufallsvariable genau dann, wenn*

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$
$$0 \leq x \leq n; x \leq M; n - x \leq N - M; x \in \mathbb{N}$$

(N *= Anzahl der Objekte der Grundgesamtheit;* M *= Anzahl der ausgezeichneten Objekte in der Grundgesamtheit;* n *= Stichprobengrösse,* x *= Anzahl der ausgezeichneten Objekte in der Stichprobe)*
Abkürzung: $X \sim H(N, n, M)$
N,M *und* n *sind die Parameter der Verteilung)*

Satz 105. *Für* $X \sim H(N, n, M)$ *gilt* $E(X) = n\frac{M}{N}$ *und* $V(X) = \frac{N-n}{N-1} n \frac{M}{N}(1 - \frac{M}{N})$

Definition 106. *Ist* $F_X : \mathbb{R} \to [0, 1]$ *eine Verteilungsfunktion und ist* F *stetig, nennen wir die durch* F_X *definierte Wahrscheinlichkeitsverteilung* $P_X(]a, b]) := F_X(b) - F_X(a)$ *stetig.*
Die Zufallsvariable X *wird in diesem Fall „stetig verteilt“ genannt.*

Definition 107. *Sei* $f : \mathbb{R} \to \mathbb{R}$. f *ist eine Dichtefunktion einer Zufallsvariable* X *genau dann, wenn*
(1) f *ist eine Abbildung und*
(2) $f \geq 0$ *und*
(3) die Fläche zwischen f *und der* $x-$*Achse* $= 1$ *(d.h.* $\int_{-\infty}^{\infty} f dx = 1$*).*

Definition 108. *Der Erwartungswert einer stetig verteilten Zufallsvariable ist definiert durch*

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

Definition 109. *Die Varianz einer stetig verteilten Zufallsvariable ist definiert als:*

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

Definition 110. *Eine Abbildung $f_X : \mathbb{R} \to \mathbb{R}$ ist die Dichtefunktion einer stetigen, uniformen Verteilung genau dann, wenn es ein a, $b \in \mathbb{R}$ gibt, so dass $b > a$ und*

$$f_X = \begin{cases} 0 & \text{für } x \leq b \\ \frac{1}{b-a} & \text{für } a < x \leq b \\ 0 & \text{für } x > c \end{cases}$$

Für uniform verteilte Zufallsvariablen schreiben wir:

$$X \sim U(a, b).$$

Satz 111. *Für $X \sim U(a, b)$ gilt $E(X) = \frac{b-a}{2}$ und $V(x) = \frac{(b-2)^2}{12}$*

Definition 112. *F_X ist die Verteilungsfunktion einer Exponentialverteilung genau dann, wenn* $F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } 0 \leq x; \lambda > 0 \\ 0 & \text{für } x < 0 \end{cases}$
Abkürzung: $X \sim Exp(\lambda)$
λ ist der Parameter der Verteilungsfamilie.
Die entsprechende Dichtefunktion ist $f(x) = e^{-\lambda x}$

Satz 113. *Für $X \sim Exp(\lambda)$ gilt $E(X) = \frac{1}{\lambda}$ und $V(x) = \frac{1}{\lambda}$*

Satz 114. *Hauptsatz der Statistik: Sei*

- *$(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum ($\mathcal{A}$ ist das Mengensystem, auf das P definiert ist) und $X_i : \to \mathbb{R}$ abzählbar unendlich viele i.i.d. Zufallsvariablen.*
- *x eine Folge von Werten der abzählbar unendliche vielen Zufallsvariablen $(X_1, ...X_n, ...)$.*
- *$F_n : \mathbb{R} \times \mathbb{R}^\infty \to [0, 1]; F_n(t, x) := \frac{1}{n} |\{x_j : x_j \leq t, j \leq n\}|$ die empirische Verteilungsfunktion der n ersten Daten x_i von x (Treppenform; t wird von F_n die relative Häufigkeit der ersten n Daten von x zugeordnet, die kleiner-gleich t sind), und*
- *F die theoretische Verteilungsfunktion von X_i .*

  Dann gilt für beliebig kleine $k > 0$:
  $P(\{x : |F_n(t, x) - F(t)|) > k\}) = 0$ für n gegen Unendlich.
  (formal: $\lim_{n \to \infty} P(\{x : |F_n(t, x) - F(t)| > k\}) = 0$)

Definition 115. *$f_X : \mathbb{R} \to \mathbb{R}^+$ ist die Dichtefunktion einer normalverteilten Zufallsvariable genau X dann, wenn*

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

mit $\sigma > 0; \mu \in \mathbb{R}$.
σ und μ sind die Parameter der Verteilung.
Berechnung mit Excel: $P(X \leq x) = normvert(x, a, s, 1)$ für $X \sim N(a, s)$

Satz 116. *Für $X \sim N(\mu, \sigma^2)$ gilt: $E(X) = \mu$, $V(X) = \sigma^2$*

Satz 117. *Mit Φ = Verteilungsfunktion der Standardnormalverteilung gilt (ohne Diskretheitskorrektur):*

Für $X \sim B(n,p)$ und $np(1-p) \geq 9$ gilt:

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$$

Definition 118. *P-P-Plots: Sei $h(x_i)$ die relative Häufigkeit der Daten x_j mit $x_j \leq x_i$ und $F(x_i) = P(X \leq x_i)$ die Verteilungsfunktion der Zufallsvariable X. Dann stellt die graphische Darstellung der folgenden Punkte in einem Koordinatennetz einen P-P-Plot dar:*

$$(h(x_i), F(x_i))$$

Satz 119. *Zentraler Grenzwertsatz: $\bar{X}_n$ sei der Mittelwert von n Zufallsvariablen X_i. Für beliebig verteilte i.i.d. Zufallsvariablen X_i mit $0 < V(X_i) < \infty$ gilt:*

$$\lim_{n\to\infty} \frac{\bar{X}_n - E(X_i)}{\sqrt{\frac{V(X_i)}{n}}} \sim N(0,1)$$

(Für steigendes n nähert sich die Verteilung der Zufallsvariable Z der Standardnormalverteilung)

Satz 120. *Aus dem zentralen Grenzwertsatz folgt für $X \sim P(\lambda)$:*

$$\frac{X-\lambda}{\sqrt{\lambda}} \sim N(0,1) \text{ (näherungsweise, für } \lambda \geq 9)$$

Satz 121. *Y ist χ^2_n-verteilt genau dann, wenn*
(i) $Y = \sum_{i=1}^{n} X_i^2$
(ii) $X_i \sim N(0,1)$ *und*
(ii) X_i *sind unabhängig.*

Satz 122. *Y ist t_n-verteilt genau dann, wenn*
(i) $Y = \frac{X_o}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$
(ii) $X_i \sim N(0,1)$ *für* $0 \leq i \leq n$
(iii) die X_i sind unabhängig (für $0 \leq i \leq n$).

Satz 123. *Y ist eine $F_{m,n}-$verteilte Zufallsvariable genau dann, wenn*
(i) $Y = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$
(ii) $X_i \sim N(0,1)$ *für* $1 \leq i \leq m$ *und* $1 \leq i \leq n$
(iii) die X_i sind unabhängig.

Testtheorie

Einstichproben Z-Test

Ziel: Getestet wird die Abweichung des Mittelwertes $\bar{x}$ (Wert der Zufallsvariable $\bar{X}$) einer Stichprobe vom Parameter $\bar{x}_o$.
Voraussetzungen: die Zufallsvariablen X_i, als deren Werte die Daten x_i der Stichprobe betrachtet werden, sind unabhängig und identisch verteilt. Die Varianz s_o^2 der Zufallsvariablen unter der

Nullhypothese ist bekannt (bei kleinen Stichproben - $n < 30$ - müssen die Daten normalverteilt sein) (für stetig, metrisch skalierte Variablen).

$Z = \frac{\bar{X} - \bar{x}_o}{\frac{s_o}{\sqrt{n}}} \sim N(0,1)$ (gilt asymptotisch)
Zufallsvariable X mit dem Wert $\bar{x}$ (Mittelwert der n Daten)
$\bar{x}_o$: Erwartungswert der ZV unter der Nullhypothese
s_o : Standardabweichung der ZV unter der Nullhypothese

Einstichprobenvarianztest

Ziel: Getestet wird die Abweichung der Varianz s^2 einer Stichprobe von einer vorgegebenen Varianz s_0^2 (=Varianz der der Zufallsvariablen unter der Nullhypothese)
Voraussetzungen: Die Zufallsvariablen X_i, als deren Werte wir die Daten x_i auffassen, sind identisch normalverteilt und unabhängig. Die Varianz s_0^2 ist bekannt (stetig, metrisch skalierte Variablen)

$\frac{(n-1)S^2}{s_o^2}$ ist $\chi^2_{n-1}-$verteilt.
Zufallsvariable S^2 mit dem Wert s^2 der Stichprobe
s_o^2 : Varianz der ZV unter der Nullhypothese
n : Anzahl Daten

Einstichproben t-Test

Ziel: Getestet wird die Abweichung des Mittelwertes einer Stichprobe von einem vorgegebenen Mittelwert (Mittelwert der Zufallsvariablen unter der Nullhypothese)
Voraussetzungen: die Zufallsvariablen X_i, als deren Werte die Daten x_i der Stichprobe betrachtet werden, sind unabhängig und identisch normalverteilt. Die Varianz der Zufallsvariablen unter der Nullhypothese ist unbekannt und wird aus den Daten geschätzt (stetig metrisch skalierte Variable).

$T = \frac{\bar{X} - \bar{x}_0}{\frac{S}{\sqrt{n}}}$ ist $t_{(n-1)}-$verteilt (bei normalverteilten Daten).
$\bar{x}_0$ Erwartungswert der ZV unter der Nullhypothese
X : Zufallsvariable mit $\bar{x}$ der Stichprobe als Wert
S : Zufallsvariable mit s der Stichprobe als Wert.
n: Anzahl der Daten

Einstichproben-Vorzeichentest

Ziel: Getestet wird, ob die Verteilung der Daten x_i einer Stichprobe zur Verteilung der Grundgesamtheit passt, indem man den (bekannten) Median med_o einer Grundmenge verwendet
Voraussetzungen: die Zufallsvariablen X_i, als deren Werte die Daten x_i der Stichprobe betrachtet werden, sind unabhängig und identisch verteilt. (ordinal oder metrisch skalierte Variable)

$X \sim B(n, 0.5)$
$n =$ Anzahl der Daten, die nicht auf med_0 fallen.
$med_0 =$ Median der Grundgesamtheit

Einstichproben Wilcoxon-Test

Ziel: Getestet wird, ob die Verteilung der Daten x_i einer Stichprobe zur Verteilung der Zufallsvariablen unter der Nullhypohtese passt, indem man den (vorgegebenen) Median med_o unter der Nullhypothese verwendet

Voraussetzungen: die Zufallsvariablen X_i, als deren Werte die Daten x_i der Stichprobe betrachtet werden, sind unabhängig und identisch verteilt. (metrisch skalierte Variable)
Vorgehen: es wird $|x_i - med_0|$ berechnet. Daten x_i mit $x_i - med_0 = 0$ werden weggelassen. Dann wird $|x_i - med_0|$ der Grösse nach geordnet und es werden Ränge zugeordnet (wie bei der Spearman-Korrelation) Es wird dann die Rangsumme der Zahlen x_i mit $x_i - med_0 < 0$ (oder die Rangsumme der Zahlen x_i mit $x_i - med_0 > 0$) gebildet. Kritische Werte in Tabelle oder

	Kritische Werte (einseitig) bei		
n	$\alpha = 0.025$	0.01	0.005
	Kritische Werte (zweiseitig) bei		
	$\alpha = 0.05$	0.02	0.01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Tabelle 2: Kritische Werte (= Grenzen des Verwerfungsbereichs) beim Wilcoxon-Test

Berechnung mit asymptotischer Verteilung.

$Z = \frac{R^i - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$ (asymptotisch) $\quad n > 25$ (für $n \leq 25$ Tabelle!!)
n: Anzahl der von 0 verschiedenen Differenzen
R^i : Rangsummen der negativen (oder positiven) Differenzen

Einstichproben Anteilstest bei zwei Ausprägungen

Ziel: Bei den Tests bezüglich Anteilen geht es darum, zu untersuchen, ob der Anteil einer Stichprobe (zwei Ausprägungen) zum Anteil einer Grundgesamtheit passt. (Voraussetzung: i.i.d X_i)

$X \sim B(n,p)$
$n =$ Anzahl der Daten.
$p =$ Anteil der ausgezeichneten Objekte an Grundgesamtheit

Anteilstest bei mehr als zwei Ausprägungen

Ziel: Es werden die Häufigkeiten einer Stichprobe mit den auf die Stichprobengrösse umgerechneten Häufigkeiten unter der Nullhypothese verglichen.

Voraussetzungen: die Daten sind Realisierungen unabhängiger, identisch verteilter Zufallsvariablen. Die erwarteten Häufigkeiten sind in maximal 20% der Zellen kleiner als 5.

$\sum_{i=1}^{m} \frac{(K_i - k_i^e)^2}{k_i^e}$ ist asymptotisch $\chi^2_{m-1}-$verteilt.
K_i = Zufallsvariable, welche die absoluten
Häufigkeiten k_i der Stichprobenklassen annimmt
k_i^e = absolute Häufigkeiten unter der Nullhypothese auf Stichprobengrösse
umgerechnet.
m = Anzahl Klassen

Verbundener Zweistichproben-t-Test

Ziel: Zwei verbundene Stichproben werden verglichen, um zu entscheiden, ob sie als Werte von Zufallsvariablen mit identischem Erwartungswert betrachtet werden können.
Voraussetzungen: Die Differenzen der Daten sind unabhängig und identisch verteilt. Die Varianz der Differenzen wird aus den Daten geschätzt. Die Differenzen müssen normalverteilt sein (metrisch skalierte Variablen) (der Zweistichproben-t-Test ist mit dem Einstichproben-t-Test äquivalent, wenn in letzterem $\bar{D} := \bar{X}$ und $\bar{x}_0 = 0$ gesetzt wird)

$T = \frac{\bar{D}}{\frac{S_d}{\sqrt{n}}}$ ist $t_{n-1}-$verteilt.
D: Zufallsvariable mit dem Mittelwert d der Differenzen d_i als Wert
S_d: Zufallsvariable mit der empirischen Standardabweichung s_d der Differenzen als Wert
n: Anzahl Differenzen

Verbundener Zweistichproben-Vorzeichentest

Ziel: Zwei verbundene Stichproben werden verglichen, um zu entscheiden, ob sie zur selben Grundgesamtheit gehören.
Voraussetzungen: Die Differenzen der Daten sind unabhängig und identisch verteilt. (metrisch oder ordinal skalierte Variablen) (Der verbundene Zweistichproben-Vorzeichentest ist mit dem Einstichproben Vorzeichentest äquivalent, wenn die Differenzen als Daten betrachtet werden und der Median mit 0 identifiziert wird).

$X \sim B(n, 0.5)$
n: Anzahl der von 0 verschiedenen Differenzen

Verbundener Zweistichproben-Wilcoxon-Test

Ziel: Zwei verbundene Stichproben werden verglichen, um zu entscheiden, ob sie als Werte von Zufallsvariablen mit identischer Verteilung betrachtet werden können.
Voraussetzungen: Die Differenzen der Daten sind unabhängig und identisch verteilt. (metrisch skalierte Variablen) (es sollte nicht zuviele gleiche Differenzen geben und diese Gruppen von Differenzen sollten nicht zu gross sein). Die Differenzen sollten symmetrisch um ihr arithmetisches Mittel liegen (allerdings ist der Test bezüglich dieser Bedingung robust, so dass wir sie im folgenden nicht beachten).
Vorgehen: $|x_i - y_i|$ rangieren. Die Summe der Ränge mit $x_i - y_i < 0$ berechnen (oder die Summe der Ränge mit $x_i - y_i > 0$). In Tabelle Rangsumme mit kritischem Wert vergleichen (oder bei genügend Daten asymptotischen Testwert berechnen; s. oben Wilcoxon-Einstichprobentest).

Verbundener Zweistichproben-McNemar-Test

Ziel: Zwei verbundene Stichproben werden verglichen, um zu entscheiden, ob sich eine signifikante Änderungen der absoluten Häufigkeiten ergeben haben (Variablen mit zwei Ausprägungen).
Voraussetzungen: i.i.d. Zufallsvariablen

$X \sim B(n, 0.5)$
n : Anzahl der Objekte, bei denen eine Änderung erfolgte.

Unverbundener Zweistichproben-Varianz-Test

Ziel: Die Varianzen zweier Stichproben werden verglichen, um zu entscheiden, ob die Stichprobendaten als Werte von Zufallsvariablen mit identischer Varianz betrachtet werden können.
Voraussetzungen: Die Daten sind unabhängig und identisch normalverteilt (metrisch skaliert).

$\frac{S_1^2}{S_2^2}$ ist $F_{(n_1-1,n_2-1)}-$verteilt.
S_1^2 : Zufallsvariable mit s_1^2 als Wert
S_2^2 : Zufallsvariable mit s_2^2 als Wert
n_1 : Umfang der ersten Stichprobe
n_2 : Umfang der zweiten Stichprobe

Unverbundener Zweistichproben-t-Test (bei nicht signifikant verschiedene Varianzen)

Ziel: Es wird überprüft, ob die Werte zweier Stichproben als Werte von Zufallsvariablen mit identischen Erwartungswerten betrachtet werden können.
Voraussetzungen: Die Daten sind unabhängig und identisch normalverteilt (metrisch skaliert). Die Varianzen werden aus den Daten geschätzt und unterscheiden sich nicht signifikant.

$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ ist $t_{n_1+n_2-2}-$verteilt.
X_1: Zufallsvariable mit dem Mittelwert $\bar{x}_1$ der Stichpobe 1 als Wert
X_2: Zufallsvariable mit dem Mittelwert $\bar{x}_2$ der Stichprobe 2 als Wert
S_1^2: Zufallsvariable mit der Varianz s_1^2 der Stichprobe 1 als Wert.
S_2^2 :Zufallsvariable mit der Varianz s_2^2 der Stichprobe 2 als Wert
n_1 : Anzahl Daten der Stichprobe 1
n_2 : Anzahl Daten der Stichprobe 2

t-Test bei signifikant verschiedenen Varianzen (Welch-Test)

Ziel: Es wird überprüft, ob die Stichprobendaten als Werte von Zufallsvariablen mit identischen Erwartungswerten betrachtet werden können.
Voraussetzungen: Die Daten sind unabhängig und identisch normalverteilt (metrisch skaliert). Varianzen können sich unterscheiden.

$T = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}$ ist näherungsweise t-verteilt mit
$df = \frac{(1+R)^2}{\frac{R^2}{(n-1)}+\frac{1}{m-1}}$, wobei $R = \frac{\frac{S_x^2}{n}}{\frac{S_y^2}{m}} = \frac{S_x^2 m}{S_y^2 n}$
$\bar{X} =$ ZV Mittelwert der Variable X
$\bar{Y} =$ ZV Mittelwert der Variable T
$S_x^2 =$ Zufallsvariable Varianz der Variable X
$S_y^2 =$ Zufallsvariable Varianz der Variable Y
$n =$ Stichprobengrösse der Stichprobe 1
$m =$ Stichprobengrösse der Stichprobe 2

Rangtest von Mann-Whitney

Ziel: Es wird überprüft, ob die Daten zweier Stichproben als Werte von Zufallsvariablen mit identischer Verteilung betrachtet werden können.
Voraussetzungen: Die Daten sind unabhängig und identisch verteilt (metrisch oder ordinal skaliert). Teststatistik:

$$U_1 := n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
R_i Zufallsvariable, die als Testwert $\sum_{x_j \in S_i} r_{x_j}$ annimmt

Die Zufallsvariable R_i nimmt also als Wert die Rangsumme der Daten x_j der i-ten Stichprobe an. Für U_1 gilt: $U_1 = n_1 n_2 - U_2$, wobei $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$.

n_2	9	10	11	12	13	14	15	16	17	18	19	20
n_1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

Tabelle 3: Kritische Werte für den U-Test von Mann und Whitney für den einseitigen Test bei $\alpha = 0.01$, für den zweiseitigen Test bei $\alpha = 0.02$; der kritische Wert gehört zum Verwerfungsbereich

n_2	9	10	11	12	13	14	15	16	17	18	19	20
n_1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Tabelle 4: Kritische Werte für den U-Test von Mann und Whitney für den einseitigen Test bei $\alpha = 0.025$, für den zweiseitigen Test bei $\alpha = 0.05$; der kritische Wert gehört zum Verwefungsbereich

n_2	9	10	11	12	13	14	15	16	17	18	19	20
n_1												
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

Tabelle 5: Kritische Werte für den U-Test von Mann und Whitney für den einseitigen Test bei $\alpha = 0.05$, für den zweiseitigen Test bei $\alpha = 0.1$; der kritische Wert gehört zum Verwerfungsbereich

Bei $n_1 + n_2 \geq 30$ ist die Verteilung der Teststatistik U_i näherungsweise normal.

$Z = \frac{U_i - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \sim N(0,1)$ (asymptotisch)
U_i : Zufallsvariable $= n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i \quad (i \in \{1,2\})$.
R_i : Zufallsvariable (Wert von $R_i = \sum_{j=1}^{n_i} r_i(x_j)$)
$r_i(x_j)$ = Rang des Datums x_j der Stichprobe i
n_1 : Umfang der ersten Stichprobe
n_2 : Umfang der zweiten Stichprobe

Bei verbundenen Rängen wird eine andere Standardabweichungsformel verwendet:

$Z = \frac{U_i - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n(n-1)} \left(\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} \right)}} \sim N(0,1)$
$t_i =:$ Anzahl der Daten mit dem jeweils selben Rang i
k : Anzahl der Datengruppen mit verbundenen Rängen
$n := n_1 + n_2$
Übrige Interpretationen wie oben.

Test auf Anteile bei zwei Ausprägungen

Ziel: Es sollen überprüft werden, ob die Anteile in zweier Stichproben zueinander passen (als Werte identisch verteilter Zufallsvariablen betrachtet werden können).

Voraussetzungen: Stichprobengrösse: $n_i \geq 50$. Die Daten sind Werte von unabhängigen, identsich verteilten Zufallsvariablen. Teststatistik:

$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1}+\frac{1}{n_2})}} \sim N(0,1)$ (asmyptotisch für $n_i \geq 50$)
$\frac{X_1}{n_1}$: Zufallsvariable mit dem Anteil $\frac{x_1}{n_1}$an der Stichprobe 1 als Wert
$\frac{X_2}{n_2}$: Zufallsvariable mit dem Anteil $\frac{x_2}{n_2}$an der Stichprobe 2 als Wert
n_1 : Grösse der Stichprobe 1
n_2 : Grösse der Stichprobe 2

Vertrauensintervalle

$]T_0, T_1[= \{b' : b'$ führt bei einem Testergebnis von b bei einem Signifikanzniveau von $\alpha = 1 - \beta$ **nicht** zur Verwerfung der Nullhypothese, dass die Stichprobe mit dem Testwert b aus einer Grundgesamtheit mit dem Parameter b' stammt$\}$ (bei zweiseitigem Testen)
$]T_0, T_1[$ wird das $\beta-$Vertrauensintervall oder $\beta-$Konfidenzintervall von b genannt. T_0, T_1 sind Zufallsvariablen, welche als konkrete Werte die Intervallgrenzen annehmen.

Vertrauensintervall für eine normalverteilte Statistik

$T_i = \bar{X} \pm \Phi^{-1}(\beta + \frac{1-\beta}{2})\frac{s_0}{\sqrt{n}}$
β = Wahrscheinlichkeit, dass wirklicher Wert in Intervall ist
$\bar{X}$ = Zufallsvariable mit dem Stichprobenmittel $\bar{x}$ als Wert
T_i = Zufallsvariable Intervallgrenze des Vertrauensintervalles,
die als Wert t_i annimt.
s_0 = Standardabweichung der ZV unter Nullhypothese
$\Phi^{-1}(\beta + \frac{1-\beta}{2}) = \beta + \frac{1-\beta}{2}-$Quantil der Standardnormalverteilung

Vertrauensintervall für eine t-verteilte Statistik

$$T_i = \bar{X} \pm F^{-1}_{t_{(n-1)}}(\beta + \frac{1-\beta}{2})\frac{S}{\sqrt{n}},$$

Stichprobentheorie

Satz 124. *Sei Y die Zufallsvariable, welche als Werte die Summe der y_j einer Stichprobe $\mathfrak{S}_i$ annimmt. Dann gilt:*

$$E(Y) = \frac{n}{N}\sum_{j=1}^{N} y_j$$

Satz 125. *Sei $\bar{Y}$ die Zufallsvariable, die Stichprobenmittel als Werte annimmt. Dann gilt:*

$$E(\bar{Y}) = \bar{y}_0$$

($\bar{y}_0$ Mittelwert der Gesamtpopulation).

Satz 126. $E(N\bar{Y}) = y$
(N = Anzahl Objekte der Gesamtpopulation; $\bar{Y}$ = Zufallsvariable mit den Mittelwerten der Stichproben der Grösse n als Werte; y = Total der Gesamtpopulation).

Satz 127. $V(\bar{Y}) = \frac{N-n}{nN}s_0^2 \qquad \left(= \left(\frac{1}{n} - \frac{1}{N}\right)s_0^2 = \frac{1}{n}(1 - \frac{n}{N})s_0^2\right)$
($\bar{Y}$ = Zufallsvariable mit Stichprobenmitteln als Werte; s_0^2 Varianz der Grundgesamtheit)

Satz 128. *Sei S^2 die Zufallsvariable, die als Werte konkrete Stichprobenvarianzen annimmt. Es gilt:*

$$E(S^2) = s_0^2$$

(s_0^2 = Varianz der Grundgesamtheit).

Satz 129. $\frac{1}{n}\left(1 - \frac{n}{N}\right)S^2$ *ist eine erwartungstreue Schätzstatistik der Varianz* $V(\bar{Y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)s_0^2$.
(n = Stichprobengrösse; N = Grösse der Gesamtpopulation; s_0^2 Varianz der Gesamtpopulation. Die Zufallsvariable S^2 nimmt im konkreten Fall den Wert $s_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ = Varianz der Stichprobe $\mathfrak{S}_i$ an).

Satz 130. *Für das Stichrpobenmittel ist das folgende Vertrauensintervall gegeben:*

$\bar{Y}_{g_i} = \bar{Y} \pm \Phi^{-1}(\beta + \frac{1-\beta}{2})\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right) S^2}$ *(mit Endlichkeitskorrektur)*
$\beta =$ *Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
$Y_{g_i} =$ *Zufallsvariable „Intervallgrenzen“ mit konkreten*
Intervallgrenzen $\bar{y}_g$ *des Vertrauensintervalls als Wert.*
$Y =$ *Zufallsvariable mit den Stichprobenmittelwerten* $\bar{y}_i$ *als Werten.*
$S =$ *Zuvallsvariable „Standardabweichung“ mit der*
Standardabweichung der Stichprobe s *als Wert.*
$n =$ *Stichprobengrösse*
$N =$ *Grösse der Grundgesamtheit*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 131. *Für einen gewünschten Präzisionsgrad des Mittelwertes (= Breite des Vertrauensintervalls) ist die folgende Stichprobengrösse gegeben:*

$n = \frac{N s_0^2 t^2}{s_0^2 t^2 + d^2 N}$ *(mit Endlichkeitskorrektur)*
$d =$ *die Hälfte der Breite des Vertrauensintervalles*
$s_0 =$ *Standardabweichung der Gesamtpopulation (gewöhnlich*
geschätzt durch frühere Stichprobe)
$n =$ *Stichprobengrösse*
$t := \Phi^{-1}(\beta + \frac{1-\beta}{2})$
$\beta =$ *Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 132. *Die Varianz des Totals* $N\bar{Y}$ *ist gegeben durch:* $V(N\bar{Y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) s_0^2$

Satz 133. *Für das Total* $N\bar{Y}$ *ist das folgende Vertrauensintervall gegeben:*

$Y_{g_i} = N\bar{Y} \pm \Phi^{-1}(\beta + \frac{1-\beta}{2})\sqrt{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2}$ *(mit Endlichkeitskorrektur)*
$Y_{g_i} =$ *Zufallsvariable „Intervallgrenze“ mit konkreten*
Intervallgrenzen als Wert.
$NY =$ *Zufallsvariable „Total“ mit konkreten Totalen als Wert*
$S =$ *Zufallsvariable „Standardabweichung“ mit den konkreten*
Stichprobenstandardabweichungen s_S *als Wert.*
$N =$ *Anzahl Objekte der Gesamtpopulation*
$n =$ *Anzahl Objekte der Stichprobe*
$\beta =$ *Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 134. *Für einen gewünschten Präzisionsgrad (= Breite des Vertrauensintervalls) des Totals ist die folgende Stichprobengrösse gegeben:*

$n = \frac{t^2 s_0^2 N^2}{d^2 + t^2 s_0^2 N}$ *(mit Endlichkeitskorrektur)*
$s_0^2 =$ *Varianz in der Grundgesamtheit, gewöhnlich geschätzt*
aus früheren Erhebungen
$d =$ *Hälfte der Vertrauensintervalllänge*
$N =$ *Anzahl Objekte der Gesamtpopulation*
$n =$ *Anzahl Objekte der Stichprobe*
$t = \Phi^{-1}(\beta + \frac{1-\beta}{2})$
$\beta =$ *Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 135. *Sei P_A die Zufallsvariable, die für eine Stichprobe $\mathfrak{S}_i$ den Stichprobenanteil $p^A_{\mathfrak{S}_i}$ annimmt und sei p^A_G der Anteil von A an G. Dann gilt*

$$E(P_A) = p^A_G.$$

Satz 136. $V(P_A) = \frac{N-n}{nN} s_0^2 = \frac{N-n}{N-1} \frac{p^A_G(1-p^A_G)}{n}$ *mit* $s_0^2 = \frac{N}{N-1} p^A_G(1-p^A_G)$

Satz 137. *Für Stichprobenanteile ist das folgende Vertrauensintervall gegeben:*

$P_{g_i} = P_A \pm \Phi^{-1}(\beta + \frac{1-\beta}{2}) \sqrt{(1-\frac{n}{N}) \frac{P_A(1-P_A)}{n-1}}$
mit Endlichkeitskorrektur
β *= Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
P_{g_i} *= Zufallsvariablen „Grenze des Vertrauensintervalls“ mit konkreten Grenzen als Wert.*
P_A *= Zufallsvariable mit den Anteilen* $p^A_{\mathfrak{S}_i}$ *an der Stichprobe als Werten*
N *= Grösse der Gesamtpopulation*
n *= Stichprobengrösse*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 138. *Für einen gewünschten Präzisionsgrad (= Breite des Vertrauensintervalls) des Anteils ist die folgende Stichprobengrösse gegeben:*

$n = \frac{N}{\frac{(N-1)d^2}{t^2 p^A_G(1-p^A_G)}+1}$ *(mit Endlichkeitskorrektur)*
d *= Hälfte der Vertrauensintervalllänge*
N *= Grösse der Gesamtpopulation*
n *= Stichprobengrösse*
$t = \Phi^{-1}(\beta + \frac{1-\beta}{2})$
β *= Wahrscheinlichkeit, dass Grösse im Vertrauensintervall liegt (z.B. 0.95)*
Φ^{-1} *Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung*

Satz 139. *Schätzstatistik des Mittelwertes einer geschichteten Stichprobe:*

$\bar{Y} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{Y}_h$
N *= Anzahl Objekte der Grundgesamtheit*
L *= Anzahl Schichten pro Schichtebene*
N_h *= Anzahl Objekte der Schicht h (Grundgesamtheit)*
Y_h *= Zufallsvariable mit Stichprobenmittelwert der Schicht h als Wert*

Satz 140. *Varianz der Statistik Mittelwert einer geschichteten Stichprobe:*

$V(\bar{Y}) = \sum_{h=1}^{L} \left(W_h^2 \cdot \frac{1}{n_h} \cdot (1 - \frac{n_h}{N_h}) \sigma_h^2 \right)$
L *= die Anzahl der Schichten einer Schichtebene*
$W_h = \frac{N_h}{N}$ *= Anteil der Schicht h in der Gesamtpopulation*
n_h *= Anzahl der Objekte der Schicht h in der Stichprobe*
N_h *= Anzahl der Objekte der Schicht h in der Gesamtpopulation*
$s_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \bar{y}_h)^2$ *die Varianz der Objekte der Schicht h in der Gesamtpopulation*
U_h *= Menge der Indizes der Objekte der Schicht h in der Gesamtpopulation*
$\bar{y}_h$ *= Mittelwert der Ausprägung* y_i *der Schicht h in der Gesamtpopulation*

s_h^2 *wird geschätzt durch* $S_h^2 = \frac{1}{n_h - 1} \sum_{i \in I_h} (y_i - \bar{Y}_h)^2$
I_h *= Menge der Indizes der Schicht h in der Stichprobe*
S_h^2 *= Zufallsvariable mit der Stichprobenvarianz der Schicht h als Wert.*
Y_h *= Zufallsvariable mit dem Stichprobenmittel der Schicht h als Wert.*
y_i *= Stichprobenwerte*

Wir erhalten das Vertrauensintervall:

$$\boxed{\bar{Y}_{g_i} = \bar{Y} \pm \Phi^{-1}(\beta + \tfrac{1-\beta}{2})\sqrt{V(\bar{Y})}}$$

Satz 141. *Für Totale erhalten wir die Schätzstatistik:*

$T = \sum_{i=1}^{L} N_i \bar{Y}_i$
$L =$ *Anzahl Schichten*
$N_i =$ *Anzahl Objekte der Schicht i in der Grundgesamtheit.*
$\bar{Y}_i =$ *Zufallsvariable mit Stichprobenmittel der Schicht i als Wert*

Die Varianz ist:

$$\boxed{V(T) = V(N\bar{Y}) = N^2 V(\bar{Y}) = N^2 \sum_{i=1}^{L} \left(W_i^2 \cdot \tfrac{1}{n_i} \cdot (1 - \tfrac{n_i}{N_i})\sigma_i^2 \right)}$$

und das Vertrauensintervall

$$\boxed{T_g = T \pm \Phi^{-1}(\beta + \tfrac{1-\beta}{2})\sqrt{V(T)}}$$

Satz 142. *Für Anteile erhält man bei geschichteten Stichproben:*

$P_A = \sum_{h=1}^{L} \frac{N_h P_h}{N}$ *ist eine erwartungstreue Schätzstatistik eines Anteils der Grundgesamtheit.*
$P_A =$ *Zufallsvariable, die einen konkreten Stichprobenanteil als Wert annimmt*
$N_h =$ *Grösse der Schicht in der Grundgesamtheit.*
$P_h =$ *Zufallsvariable, die einen konkreten Schicht-Stichprobenanteil annimmt.*
$N =$ *Grösse der Grundgesamtheit.*

und

$V(P_A) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h(1-P_h)}{n_h}$.
$P_A =$ *Zufallsvariable, die einen einen konkreten Stichprobenanteil als Wert annimmt*
$N_h =$ *Grösse der Schicht h in der Grundgesamtheit.*
$P_h =$ *Zufallsvariable, die einen konkreten Schicht-Stichprobenanteil annimmt.*
$N =$ *Grösse der Grundgesamtheit.*
$n_h =$ *Grösse der Schicht h in der Stichprobe.*

Dies ergibt für die Vertrauensintervalle:

$$\boxed{P_g = P_A \pm \Phi^{-1}(\beta + \tfrac{1-\beta}{2})\sqrt{V(P_A)}.}$$

Bivariate Einstichprobentests

χ^2-Unabhängigkeitstests:

Ziel: Es soll überprüft werden, ob zwei statistische Variablen unabhängig sind (d.h. ob die Verteilung der Häufigkeit von Objekten den laut Randverteilung erwarteten Häufigkeiten entspricht oder ob es bedeutsame Abweichungen von den erwarteten Häufigkeiten gibt).
Voraussetzungen: Die erwartete Häufigkeit pro Feld sollte nicht bei mehr als 20% der Felder kleiner sein als 5. (Vor allem für nominalskalierte Daten verwendet). Die Zufallsvariablen sind identisch verteilt und unabhängig.

$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(K_{i,j}-k_{i,j}^{e}\right)^2}{k_{i,j}^{e}}$ ist asymptotisch $\chi^2_{(I-1)\cdot(J-1)}-$verteilt für $H_{ij}^e < 5$ für
höchstens 20% der Felder
$K_{i,j}$: Zufallsvariable, die als Werte die empirischen absolute Häufigkeiten k_{ij}
annimmt.
$k_{i,j}^e$: erwartete absolute Häufigkeiten
I : Anzahl der Zeilen
J : Anzahl der Spalten

Gamma

Ziel: Testen eines proportionalen Zusammenhangs in Kreuztabellen mit ordinalskalierten Variablen
Voraussetzungen: grosse Stichproben; Daten können als Werte unabhängiger und identisch verteilter Zufallsvariablen betrachtet werden.
(i) beide Variablen sind ordinalskaliert oder
(ii) eine der Variablen ist ordinalskaliert während die andere eine nominalskalierte Variable mit nur zwei Ausprägungen ist (dichotom skalierte Variable)

$\frac{\Gamma}{\sqrt{\frac{S}{n}}} \sim N(0,1)$ (asymptotisch; für n genügend gross)
S ist die Zufallsvariable, die den Wert $s = \frac{16}{(\pi_c+\pi_d)^4}\cdot\sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}(\pi_d h_{ij}^c - \pi_c h_{ij}^d)^2$ annimt:
$h_{ij} = \frac{k_{ij}}{n}$
k_{ij} = absolute Häufigkeit der Zelle i,j
n = Stichprobengrösse
$h_{ij}^c = \sum_{a<i}\sum_{b<j} h_{ab} + \sum_{a>i}\sum_{b>j} h_{ab}$
$h_{ij}^d = \sum_{a<i}\sum_{b>j} h_{ab} + \sum_{a>i}\sum_{b<j} h_{ab}$
$\pi_c = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}h_{ij}^c$
$\pi_d = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}h_{ij}^d$
Γ = Zufallsvariable, die konkrete γ als Werte annimmt.

Einfaktorielle Varianzanalyse

Ziel: Die einfaktorielle Varianzanalyse untersucht den Zusammenhang einer nominalskalierten Variable und einer metrisch skalierten Variable. Es wird überprüft, ob sich die Mittelwerte der Daten bezüglich der Ausprägungen der nominalskalierten Variable unterscheiden.
Voraussetzungen: Die Daten entsprechen unabhängigen Zufallsvariablen. Die Variable, deren Mittelwerte $\bar{x}_j$ verglichen werden, ist metrisch skaliert. Die Abweichungen der Daten von den jeweiligen Mittelwerten $\bar{x}_j$ sind normalverteilt mit $N(0,\sigma_j^2)$, wobei näherungsweise $\sigma_j^2 = \sigma_l^2$ für alle j und l.(j,l entsprechen den Ausprägungen der nomalskalierten Variable). Die Varianzanalyse wird auch ANOVA genannt (Analysis of Variance).

$\frac{\frac{SQM}{m-1}}{\frac{SQR}{n-m}}$ ist $F_{m-1,n-m}-$verteilt (sofern die Residuen gemäss $N(0,\sigma_i^2)$ verteilt sind
m: Anzahl Gruppen
n: Anzahl Daten
SQM: $\sum_{j=1}^{m} k_j \left(\bar{x}_j - \bar{x}_t\right)^2$
SQR: $\sum_{j=1}^{m}\sum_{i=1}^{k_j} \left(x_{ij} - \bar{x}_j\right)^2$
k_j : Anzahl Daten der Gruppe j
$\bar{x}_j$: Mittelwert der Daten der Gruppe j
$\bar{x}_t$: Mittelwert aller Daten.
$s_i^2 = s_1^2 = ... = s_m^2$ (die Varianzen der Residuen der Gruppen müssen identisch sein)

Kruskal-Wallis-Test

Ziel: Der Kruskal-Wallis-Test untersucht den Zusammenhang einer nominalskalierten Variable und einer metrisch oder ordinal skalierte Variable. Es wird überprüft, ob sich die Verteilung der Daten bezüglich der Ausprägungen der nominalskalierten Variable unterscheiden.
Voraussetzungen: Die Variable, deren Verteilungen verglichen werden, ist metrisch oder ordinal skaliert. Es gibt nicht zuviele Bindungen.

Beim Kruskal-Wallis-Test werden die Daten der Grösse nach geordnet. Ihnen werden dann in bekannter Manier (siehe Wilcoxon-Test) Ränge zugeordnet. Man kann die folgende χ^2-verteilte Teststatistik entwickeln:

$K = \frac{12}{n(n+1)} \sum_{j=1}^{m} k_j(\bar{R}_j - \frac{n+1}{2})^2$ ist asymptotisch $\chi^2_{m-1}-$verteilt.
$\bar{R}_j = \frac{1}{k_j} \sum_{i=1}^{k_j} r_{ij}$ (Durchschnittlicher Rang der Gruppe j)
r_{ij} = Rang des Datums x_{ij}
n: Anzahl der Daten
k_j : Anzahl der Daten der Gruppe j
m: Anzahl der Gruppen

Bindungskorrektur: Multiplikation von K mit:

$$\frac{1}{1 - \frac{1}{n^3-n}\sum_{i=1}^{g}(t_j^3 - t_j)}$$

wobei g die Anzahl der Gruppen von verbundenen Ränge ist; t_j Anzahl der Daten pro Gruppe j verbundener Daten.

Test für den Pearson-Korrelationskoeffizienten

Ziel: Es soll der lineare Zusammenhang zwischen zwei metrisch skalierten Variablen X und Y untersucht werden.
Voraussetzung: Die Daten können als Ausprägungen unabhängiger, identisch normalverteilter Zufallsvariablen betrachtet werden. Metrisch skaliert. Die Daten liegen in etwa auf einer Geraden und liegen ungefähr punktsymmetrisch zum Punkt $(\bar{x}, \bar{y})$.

Es kann gezeigt werden, dass für *normalverteilte* Zufallsvariablen gilt:

$\frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}$ ist $t_{n-2}-$verteilt.
R_{xy} : Zufallsvariable mit konkreten Pearson-Korrelationskoeffizienten r_{xy} als Wert
n : Stichprobengrösse

Test für die Spearman-Rangkorrelation

Ziel: Es soll der Zusammenhang zwischen zwei metrisch oder ordinal skalierten Variablen X und Y untersucht werden.
Voraussetzung: Die Daten können als Werte identisch verteilter, unabhängiger Zufallsvariablen betrachtet werden. Sie sind metrisch oder ordinal skaliert. Es liegen nicht zuviele Bindungen vor.

Als Teststatistik wird die "Hotelling-Pabst-Statistik" verwendet:

$$\boxed{D = \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2}$$

mit $R(x_i)$ als Rang von x_i und $R(y_i)$ als Rang von y_i. Für die Verteilung dieses Testwertes gelten ($n \leq 30$) die folgenden kritischen Werte

	$\alpha =$			
n	0.01	0.025	0.05	0.1
4			2	2
5	2	2	4	6
6	4	6	8	14
7	8	14	18	26
8	16	24	32	42
9	28	38	50	64
10	44	60	74	92
11	66	86	104	128
12	94	120	144	172
13	130	162	190	226
14	172	212	246	290
15	224	270	312	364
16	284	340	390	450
17	356	420	480	550
18	438	512	582	664
19	532	618	696	790
20	638	738	826	934
21	758	870	972	1092
22	892	1020	1134	1270
23	1042	1184	1312	1464
24	1208	1366	1510	1678
25	1390	1566	1726	1912
26	1590	1786	1960	2168
27	1808	2024	2216	2444
28	2046	2284	2494	2744
29	2306	2564	2796	3068
30	2584	2868	3120	3416

Tabelle 6: (kritischer Werte der Hotelling-Pabst-Statistik; Nullhypothese wird verworfen, wenn Testwerg kleiner-gleich kritischer Wert

Für $n > 30$ gilt näherungsweise:

$$\frac{D - E(D)}{\sqrt{Var(D)}} \sim N(0,1).$$

mit: D wie oben.
n = Anzahl Daten pro Variable
$R(x_i)$ = Rang von x_i

$$E(D) = \frac{1}{6}(n^3 - n) - \frac{1}{12}\sum_{j=1}^{m}(d_j^3 - d_j) - \frac{1}{12}\sum_{j=1}^{p}(h_j^3 - h_j)$$

m = Anzahl der unterschiedlichen Werte x_i in der Messreihe X.
p = Anzahl der unterschiedlichen Werte y_i in der Messreihe Y.
d_j = Anzahl der Beobachtungen, die den Rang $R(x_i)$ aufweisen.
h_j = Anzahl der Beobachtungen, die den Rang $R(y_i)$ aufweisen.
Für die Berechnung von

$$-\frac{1}{12}\sum_{j=1}^{m}(d_j^3 - d_j) - \frac{1}{12}\sum_{j=1}^{p}(h_j^3 - h_j)$$

genügt es, die Ausprägungen mit gleichen Rängen zu betrachten, da sonst

$$(d_j^3 - d_j) = 1^3 - 1 = 0 = (h_j^3 - h_j).$$

Liegen keine Bindungen - dh. gleiche Ränge - vor , gilt somit:

$$E(D) = \frac{1}{6}(n^3 - n).$$

$$Var(D) = \frac{(n-1)(n+1)^2 n^2}{36}\left(1 - \frac{\sum_{j=1}^{m}(d_j^3 - d_j)}{n^3 - n}\right)\left(1 - \frac{\sum_{j=1}^{p}(h_j^3 - h_j)}{n^3 - n}\right)$$

Auch hier gilt, dass wir für $\sum_{j=1}^{m}(d_j^3 - d_j)$ und $\sum_{j=1}^{p}(h_j^3 - h_j)$ nur Daten mit gleichen Rängen berücksichtigen müssen.

Test für die partielle Korrelation

Ziel: Es soll der Zusammenhang zwischen zwei Variablen (unter Ausschluss des Einflusses einer dritten Variable) getestet werden.

Voraussetzungen: alle drei Variablen sind metrisch skaliert (unabhängig, identisch normalverteilt). Die paarweisen zweidimensionalen Punktewolken liegen in etwa auf einer Geraden und ungefähr punktsymmetrisch zu den jeweiligen Punkten, welche die Mittelwerte der beiden Variablen als Komponenten enthalten.

$\frac{R_{xy/z}\cdot\sqrt{n-3}}{\sqrt{1-R^2_{xy/z}}}$ ist $t_{n-3}-$verteilt.
$R_{xy/z}$: Zufallsvariable mit partiellen Korrelationen von X und Y unter Ausschluss
des Einflusses von Z als Werte.
n: Stichprobengrösse.

Regression

Berechnung der Koeffizienten:

$$\boxed{a = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}}$$

und

$$\boxed{b = \bar{y} - a\bar{x}}$$

Testtheorie: ANOVA

Ziel: Es soll überprüft werden, ob die Regressionsgerade Varianz „erklärt“.

Voraussetzungen: Die Daten können als Realisierungen unabhängiger und identisch verteilter Zufallsvariablen betrachtet werden. Die Residuen sind normalverteilt.

$\frac{SQE}{\frac{1}{n-2}SQR}$ ist $F_{1,n-2}-$verteilt.
$SQE = \sum_{i=1}^{n}(f(X_i) - \bar{Y})^2$ (erklärte Streuung)
$SQR = \sum_{i=1}^{n}(Y_i - f(X_i))^2$ (restliche Streuung)
n = Anzahl Datenpaare (X_i, Y_i)
$(X_1, ..., X_n)$ = Zufallsvektor, der die Werte $(x_1, ..., x_n)$ annimmt.
$(Y_1, ..., Y_n)$ = Zufallsvektor, der die Werte $(y_1, ..., y_n)$ annimmt
$f(X_i) = AX_i + B$ (Regressionsgerade)
$\bar{Y}$ = Zufallsvariable Mittelwerte der Zufallsvariablen Y_i

Testen der Koeffizienten:

$\frac{A}{\sqrt{V(A)}}$ ist $t_{n-2}-$verteilt.
A: Zufallsvariable mit der Steigung a als Wert
$V(A)$ geschätzt durch $\frac{V(R_i)}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$; $V(A)$ = Varianz der Zufallsvariable A
$V(R_i)$ geschätzt durch $\frac{1}{n-2}\sum_{i=1}^{n} R_i^2$ (mit $R_i := Y_i - f(X_i)) := Y_i - (AX_i + B)$)
n: Anzahl Paare Zufallsvariablen (X_i, Y_i)

$\frac{B}{\sqrt{V(B)}}$ ist $t_{n-2}-$verteilt.
B: Zufallsvariable mit der Steigung b als Wert
$V(B)$ geschätzt durch $\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right) V(R_i)$ ($V(B)$ =Varianz der Zufallsvariable B)
f: Regressionsgerade
$V(R_i)$ geschätzt durch $\frac{1}{n-2}\sum_{i=1}^{n} R_i^2$ (mit $R_i = Y_i - f(X_i)$) (Varianz der i.i.d. ZV $R_i = R_1$)
n: Anzahl Paare von Zufallsvariablen (X_i, Y_i)

Vertrauensintervalle:

$A_g = A \pm F^{-1}_{t_{(n-2)}}(\beta + \frac{1-\beta}{2})\sqrt{V(A)}$
β = Wahrscheinlichkeiten, dass faktisches a im Vertrauensintervall ist

$B_g = B \pm F^{-1}_{t_{(n-2)}}(\beta + \frac{1-\beta}{2})\sqrt{V(B)}$
β = Wahrscheinlichkeit, dass faktische b im Vertrauensintervall ist

Multivariate Regression

Ziel: Es wird eine Zielvariable als affine oder lineare Funktion mehrerer Einflussvariablen (= unabhängige Variablen) ausgedrückt. Die Zielvariable muss metrisch skaliert sein. Die unabhängigen Variablen sind gewöhnlich metrisch skaliert. Es können aber auch dichotome Variablen (d.h. Variablen mit zwei Ausprägungen) verwendet werden. Nominalskalierte Variablen können mit Hilfe von

Hilfsvariablen (Dummyvariablen) in dichotome Variablen verwandelt werden (Die Variable „Farbe“ mit den Ausprägungen „rot“, „grün“ und „andere“ wird z.B. verwandelt in zwei Variablen mit je zwei Ausprägungen: die Variable „Rot_Oder_Nicht“ mit den Ausprägungen „rot“ und „nicht_rot“ und die Variable „Gruen_Oder_Nicht“ mit den Ausprägungen „grün“ und „nicht_grün“). Mit Hilfe einer ANOVA wird untersucht, ob die Varianz durch das Modell signifikant reduziert wird. Mit Hilfe der Tests der Koeffizienten wird untersucht, ob diese signifikant von 0 verschieden sind.

Voraussetzungen: Die Daten können als Werte von i.i.d. ZV verstanden werden. Die Unabhängigen Variablen sind paarweise nicht korreliert. Die Residuen sind normalverteilt. $n \geq 40$. Die unabhängigen Variablen korrelieren nicht paarweise mit den Residuen. Die jeweils zweidimensionalen Punktwolken (x_{ik}, y_i) der unabhängigen Variable X_k und der Zielvariable Y liegen ungefähr auf einer Geraden und sind punktsymmetrisch. Die Varianz der Residuen ist überall identisch.