Formelsammlung in Statistik

Häufigkeitsverteilung und Summenverteilung

Zählindices

$i = 1, \ldots, n$ zählt die Elemente der Urliste
$j = 1, \ldots, m$ zählt die Merkmalsausprägungen
$k = 1, \ldots, K$ zählt die Klassen

Klassierung von Daten

x_k^u mit $x_{k+1}^u > x_k^u$: untere Klassengrenze
$x_k^o = x_{k+1}^u$ obere Klassengrenze
$x_k^* = \frac{1}{2}(x_k^u + x_k^o)$ Klassenmitte (manchmal auch einfach x_k)
$\Delta x_k = x_k^o - x_k^u$ Klassenbreite

absolute Häufigkeit

h_j bzw. h_k

relative Häufigkeit

$$f_j = \frac{h_j}{n} \quad \text{bzw.} \quad f_k = \frac{h_k}{n}$$

Häufigkeitsdichten

$$h_k^D = \frac{h_k}{\Delta x_k}, \quad f_k^D = \frac{f_k}{\Delta x_k} \quad \text{(nur klassierte Daten!)}$$

absolute Summenhäufigkeit

$$H_j = H(X \leq x_j) = \sum_{j'=1}^{j} h_{j'} \quad \text{(für Klassen analog mit } k\text{)}$$

$$H(X > x_j) = n - H(X \leq x_j)$$

relative Summenhäufigkeit

$$F_j = F(X \leq x_j) = \sum_{j'=1}^{j} f_{j'} = H_j/n$$

$$F(X > x_j) = 1 - F(X \leq x_j)$$

Empirische Verteilungsfunktion (nichtklassierte Daten)

$$F(x) = \begin{cases} 0 & \text{falls } x < x_1, \\ 1 & \text{falls } x > x_m, \\ F_j & \text{falls } x_j \leq x < x_{j+1} \end{cases}$$

Empirische Verteilungsfunktion (klassierte Daten)

$$F(x) = \begin{cases} 0 & \text{falls } x < x_1^u, \\ 1 & \text{falls } x \geq x_K^o, \\ F_{k-1} + \left(\frac{x - x_k^u}{\Delta x_k}\right) f_k & \text{falls } x_k^u \leq x < x_k^o \end{cases}$$

Empirische Dichtefunktion (nur klassierte Daten)

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \begin{cases} 0 & \text{falls } x < x_1^u \text{ oder } x > x_K^o, \\ f_k^D & \text{falls } x_k^u \leq x < x_k^o \end{cases}$$

Lagemaße (Mittelwerte & Co)

arithmetisches Mittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(aus der Urliste)}$$

$$= \frac{1}{n}\sum_{j=1}^{m} h_j x_j = \sum_{j=1}^{m} f_j x_j \quad \text{(aus den Merkmalsausprägungen)}$$

$$\approx \frac{1}{n}\sum_{k=1}^{K} h_k x_k^* = \sum_{k=1}^{K} f_k x_k^* \quad \text{(aus klassierten Daten)}$$

arithmetisches Mittel der Linearkombination Y=a+bx

$$\bar{y} = a + b\bar{x}$$

harmonisches Mittel

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \text{bzw.} \quad \frac{1}{\sum_{j=1}^{m} \frac{f_j}{x_j}} \quad \text{bzw.} \quad \frac{1}{\sum_{k=1}^{K} \frac{f_k}{x_k^*}}$$

geometrisches Mittel

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

Modus bzw. Dichtemittel (klassierte Daten)

$$\bar{x}_D = x_{\hat{k}}^u + \frac{f_{\hat{k}}^D - f_{\hat{k}-1}^D}{2f_{\hat{k}}^D - f_{\hat{k}-1}^D - f_{\hat{k}+1}^D}\Delta x_{\hat{k}}$$

$\hat{k}$: Klassenindex der Klasse mit maximaler Häufigkeitsdichte

Median bzw. Zentralwert (geordnete Urliste)

$$x_{0.5} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & (n \text{ ungerade}) \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & (n \text{ gerade}) \end{cases}$$

Median bzw. Zentralwert (klassierte Daten)

$$x_{0.5} = x_{k'}^u + \frac{0.5 - F_{k'-1}}{f_{k'}}\Delta x_{k'}$$

mit k' so dass $F_{k'-1} < 0.5$, aber $F_{k'} \geq 0.5$,
also die Klasse k', innerhalb der F den Wert 0.5 annimmt.

q-Quantil

$$x_q = x_{k'}^u + \frac{q - F_{k'-1}}{f_{k'}}\Delta x_{k'}$$

mit k' so dass $F_{k'-1} < q$, aber $F_{k'} \geq q$,
also die Klasse k', innerhalb der F den Wert q annimmt.

Streuungsmaße

Bei klassierten Daten (Klassen $k = 1, \ldots, K$) gelten alle Gleichungen nur näherungsweise.[1]

Varianz

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{bzw.} \quad \sum_{k=1}^{K} f_k(x_k^* - \bar{x})^2$$

In der Stichprobentheorie: zusätzlicher Faktor $n/(n-1)$

Varianz der Linearkombination Y=a+bx

$$s_y^2 = b^2 s_x^2$$

Standardabweichung

$$s_x = \sqrt{s_x^2}$$

Variationskoeffizient

$$V = \frac{s_x}{\bar{x}} \quad \text{(nur sinnvoll, wenn alle } x_i > 0\text{)}$$

Verschiebungssatz

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

$$\sum_{k=1}^{K}(x_k^* - \bar{x})^2 f_k = \sum_{k=1}^{K}(x_k^*)^2 f_k - \bar{x}^2$$

Spannweite

$$R = x_{\max} - x_{\min}$$

mittlere absolute Abweichung

$$s_{\text{MAD}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{0.5}|$$

Interquartilsabstand

$$s_{\text{IQ}} = x_{0.75} - x_{0.25}$$

Weitere Maße für die Verteilungsform

N-tes zentrales Moment

$$M_N = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^N \text{ (Urliste)}$$

$$M_N = \sum_{j=1}^{m}(x_j - \bar{x})^N f_j \text{ (Merkmalsausprägungen)}$$

$$M_N = \sum_{k=1}^{K}(x_k^* - \bar{x})^N f_k \text{ (klassierte Daten)}$$

Schiefe

$$\Gamma = \frac{M_3}{s_x^3}$$

Exzess ("Kurtosis")

$$K = \frac{M_4}{s_x^4} - 3$$

[1] Die Formeln hängen sehr von der Verteilung innerhalb der Klassen ab. Auch bei Gleichverteilung innerhalb der Klasen ist die aus klassierten Daten ermittelte Stichprobenvarianz *nicht* erwartungstreu, obwohl dies in den meisten Skripten und Lehrbüchern impliziert wird.

Konzentrationsmaße

Merkmalssumme $M = \sum_{i=1}^{n} x_i = n\bar{x} = \sum_{k=1}^{K} x_k^* h_k$ (letzteres für klassierte Daten)

relativer Anteil an M $p_i = \frac{x_i}{M}$ bzw. $p_k = \frac{x_k^* h_k}{M} = \frac{x_k^* f_k}{\bar{x}}$

kumulierter relativer Anteil $P_i = \sum_{i'=1}^{i} p_{i'}$ (Urliste und klassierte Daten)

Herfindahl-Index $K_H = \sum_{i=1}^{n} p_i^2$

Exponentialindex $K_E = \prod_{i=1}^{n} p_i^{p_i} = p_1^{p_1} p_2^{p_2} \dots p_n^{p_n}$

Punkte der Lorenzkurve $(0,0)$ und (F_i, P_i), $i = 1, \dots, n$ bzw. $1, \dots, K$
mit F_i der üblichen relativen Summenhäufigkeit.

Gini-Koeffizient $G = 1 - \sum_{i=1}^{n} (P_i + P_{i-1}) \frac{1}{n} = 1 - \frac{1}{n} \left(2 \sum_{i=1}^{n-1} P_i + P_n \right)$ (Urliste)

$= 1 - \sum_{k=1}^{K} (P_k + P_{k-1}) f_k$ (klassierte Daten)

Verhältnis- und Indexzahlen

Wachstumsfaktor $I_t = \frac{x_t}{x_{t-1}}$, **Wachstumsrate** $r_t = I_t - 1$

Preisindex von Laspeyres $P_{0t}^{(\mathrm{L})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(0)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$ mit $p_i(t)$ den Preisen und $q_i(t)$ den Mengen

Preisindex von Paasche $P_{0t}^{(\mathrm{P})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(t)}$

Mengenindices von Laspeyres und Paasche $Q_{0t}^{(\mathrm{L})} = \frac{\sum_{i=1}^{n} p_i(0) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$, $Q_{0t}^{(\mathrm{P})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(t) q_i(0)}$

Wertindex $W_{0t} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$

Bivariate Analyse: Kreuztabelle klassierter Daten

Relative Häufigkeit	$f(x_i, y_j) = \dfrac{h(x_i, y_j)}{n}$
absolute Randhäufigkeiten	$h(x_i) = \sum_{j=1}^{K_y} h(x_i, y_j),\ h(y_j) = \sum_{i=1}^{K_x} h(x_i, y_j)$
relative Randhäufigkeiten	$f(x_i) = \dfrac{h(x_i)}{n},\ f(y_j) = \dfrac{h(y_j)}{n}$
bedingte relative Häufigkeiten	$f(x_i \mid y_j) = \dfrac{h(x_i, y_j)}{h(y_j)} = \dfrac{f(x_i, y_j)}{f(y_j)}$
empirische Unabhängigkeit	$f(x_i, y_j) = f(x_i) f(y_j)$ für alle i, j
Arithmetische Mittelwerte	$\bar{x} = \dfrac{1}{n} \sum_{i=1}^{K_x} x_i^* h(x_i) = \sum_{i=1}^{K_x} x_i^* f(x_i)$ $\bar{y} = \dfrac{1}{n} \sum_{j=1}^{K_y} y_j^* h(y_j) = \sum_{j=1}^{K_y} y_j^* f(y_j)$

Bivariate Analyse II: Regressionsanalyse:

(Empirische) Kovarianz

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \text{ (Urliste)}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}(x_i^* - \bar{x})(y_j^* - \bar{y})h(x_i, y_j) \text{ (Kreuztabelle)}$$

Verschiebungssatz für die Kovarianz

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \quad \text{bzw.}$$

$$\sum_i \sum_j (x_i^* - \bar{x})(y_j^* - \bar{y})h(x_i, y_j) = \sum_i \sum_j x_i^* y_j^* h(x_i, y_j) - n\bar{x}\bar{y}$$

lineare Regression

$$\hat{y}(x) = a + bx, \quad a = \bar{y} - b\bar{x}$$

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} && \text{(allgemein)} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} && \text{(Urliste)} \end{aligned}$$

nichtlineare Regression

$$\text{Minimiere } F = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}(x, a, b, \ldots)\right)^2$$

$$\text{bezüglich } a, b, \ldots \Rightarrow \frac{\partial F}{\partial a} = 0, \frac{\partial F}{\partial b} = 0, \ldots$$

Grenzfunktion (="absolute Elastizitätsfunktion")

$$g(x) = \frac{\mathrm{d}\hat{y}}{\mathrm{d}x}$$

(relative) Elastizitätsfunktion

$$\epsilon_{yx} = \frac{x}{\hat{y}}\frac{\mathrm{d}\hat{y}}{\mathrm{d}x}$$

Aufspaltung in erklärte und nichterklärte Schwankung

$$(y_i - \bar{y}) = \Delta_i + e_i \text{ mit } \Delta_i = \hat{y}_i - \bar{y}, \quad e_i = y_i - \hat{y}_i$$

Additionsregel bei linearer Regression

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\Delta_i^2 + \sum_{i=1}^{n} e_i^2$$

Bestimmtheitsmaß

$$B = 1 - U = 1 - \frac{\sum_{i=1}^{n} e_i^2}{n s_y^2} = 1 - \frac{s_e^2}{s_y^2} \text{ (allgemein)}$$

$$B = \frac{s_\Delta^2}{s_y^2} = \frac{\sum_{i=1}^{n}\Delta_i^2}{n s_y^2} \text{ (lineare Regression)}$$

Bivariate Analyse III: Korrelationsanalyse

(Maß-)Korrelationskoeffizient (nach Pearson)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Zusammenhang mit Bestimmtheitsmaß der linearen Regression

$$B = r_{xy}^2$$

Rangkorrelationskoeffizient nach Spearman

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} \left(R_i^x - R_i^y\right)^2}{n(n^2-1)}$$

Zeitreihen

Additive Aufspaltung

$$Y_i = T_i + K_i + S_i + U_i = G_i + S_i + U_i$$

T = Trend, K = Konjunkturkomponente,
$G = T + K$ = glatte Komponente, S = Saisonkomponente,
U = Rest.

multiplikative Aufspaltung

$$Y_i = T_i K_i S_i U_i = G_i S_i U_i$$

gleitender Durchschnitt der Ordnung τ

$$\bar{y}_i^{(\tau)} = \begin{cases} \frac{1}{\tau} \sum\limits_{j=i-m}^{i+m} y_j & \tau \text{ ungerade, } m = \frac{\tau-1}{2}, \\ \frac{1}{\tau}\left(\frac{y_{i-m} + y_{i+m}}{2} + \sum\limits_{j=i-m+1}^{i+m-1} y_j\right) & \tau \text{ gerade, } m = \frac{\tau}{2}. \end{cases}$$

exponentielle Glättung

$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1},\ \hat{y}_0 = y_0$
mit dem Glättungsparameter $\alpha = 1 - e^{-\frac{1}{\tau}}$

Berechnung der Saisonfigur

$$\tilde{S}_j = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \bar{y}_{ij}^{(\tau)}), \quad S_j = \tilde{S}_j - \frac{1}{\tau}\sum_{j'=1}^{\tau}\tilde{S}_{j'},$$

i = Index der Perioden, j = Index innerhalb jeder Periode.
Bei trendfreien Zeitreihen ist $\tilde{S}_j = S_j$ sowie $\bar{y}_{ij}^{(\tau)} = \bar{y}$.

Zufall und Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit	$P(A\|B) = P(A,$ falls B eingetreten ist$)$
Stochastische Unabhängigkeit	$P(A\|B) = P(A)$ bzw. $P(B\|A) = P(B)$
Additionstheorem	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiplikationstheorem	$P(A \cap B) = P(B)P(A\|B) = P(A)P(B\|A)$
Totale Wahrscheinlichkeit bei sich einander ausschließenden Ereignissen A_k	$P(B) = \sum_k P(B\|A_k)P(A_k)$
Theorem von Bayes	$P(A_k\|B) = \frac{P(B\|A_k)P(A_k)}{P(B)}$

Zufallsvariable (ZV) und Wahrscheinlichkeitsfunktion

Wahrscheinlichkeitsfunktion diskreter ZV	$p(x_i) = P(X = x_i) = p_i$
Verteilungsfunktion diskreter ZV	$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$
Wahrscheinlichkeits*dichte* stetiger ZV	$f(x) = \frac{dF}{\mathrm{d}x}$
Verteilungsfunktion stetiger ZV	$F(x) = P(X \leq x) = \int_{x'=-\infty}^{x} f(x')\ \mathrm{d}x'$
Wahrscheinlichkeit eines Intervalls (X diskret oder stetig)	$P(a \leq X \leq b) = F(b) - F(a)$ $P(X > a) = 1 - F(a)$
Erwartungswert	$E(X) = \sum_i x_i p(x_i)$ (diskrete Zufallsvariable) $= \int_{x=-\infty}^{\infty} x f(x)\ \mathrm{d}x$ (stetige Zufallsvariable)
Varianz	$V(X) = E(X - E(X))^2$ $= \sum_i [x_i - E(X)]^2 p(x_i)$ (diskrete Zufallsvariable) $= \int_{x=-\infty}^{\infty} [x - E(X)]^2 f(x)\ \mathrm{d}x$ (stetige Zufallsvariable)
Kovarianz	$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$
	$E(XY) = \sum_i \sum_j x_i y_j p(x_i, y_j)$ (diskrete Zufallsvariable) $= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y)\ \mathrm{d}x\ \mathrm{d}y$ (ststige Zufallsvariable)

Diskrete theoretische Verteilungen

Fakultät
$$n! = n \cdot (n-1) \cdot (n-2) \cdots (1)$$

Binomialkoeffizient
$$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N(N-1)\ldots(N-n+1)}{n!} = \frac{N!}{n!(N-n)!}$$

Binomialverteilung
$X \sim B(n;\theta)$
$$P(X=x) = p_B^{(n,\theta)}(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

Hypergeometrische Verteilung $X \sim H(N;n;M)$
$$P(X=x) = p_H^{(N,n,M)}(x) = \frac{\left(\begin{array}{c} M \\ x \end{array} \right) \left(\begin{array}{c} N-M \\ n-x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$$

Poissonverteilung
$X \sim \mathrm{Po}(\mu)$
$$P(X=x) = p_{\mathrm{P}}^{(\mu)}(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Stetige theoretische Verteilungen

Dichte der Gleichverteilung
$X \sim G(a,b)$
$$f_G^{(a,b)}(x) = \begin{cases} \frac{1}{b-a} & \text{falls } a \le x \le b, \\ 0 & \text{sonst.} \end{cases}$$

Exponentialverteilung
$X \sim E(\lambda)$
$$f_E^{(\lambda)}(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{sonst,} \end{cases} \quad E(X) = \sqrt{V(X)} = \tfrac{1}{\lambda}$$

Zusammenhang Exponentialverteilung mit Poissonverteilung
$n \sim \mathrm{Po}(\mu) \Leftrightarrow \Delta \sim E(\mu/T)$
n = Zahl der Ereignisse im Zeitraum T
Δ = Abstände zweier Ereignisse

Normalverteilung
$X \sim N(\mu, \sigma^2)$
$$f_N^{(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \; E(X) = \mu, \; V(X) = \sigma^2$$

Standardnormalverteilung
$$Z = \frac{X-\mu}{\sigma} \sim N(0;1), \; F(z) = F_N^{(0,1)}(x) =: \Phi(z)$$

Rechenregeln zur Anwendung der Tabelle von Φ
$$F_N^{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
$$\Phi(-x) = 1 - \Phi(x)$$

Funktionen von ZV und Zentraler Grenzwertsatz

Wahrscheinlichkeitsdichte einer (monoton steigenden) Funktion $Z = g(X)$ einer ZV

$$f_Z(z) = \left[\frac{f_X(x)}{g'(x)}\right]_{x=g^{-1}(z)}$$

Wahrscheinlichkeitsdichte der Summe $Z = X_1 + X_2$ zweier unabhängiger ZV

$$f_Z(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x)\mathrm{d}x$$

mit $f_i(x)$ den Wahrscheinlichkeitsdichten von X_i

Verteilungsfunktion der Summe $Z = X_1 + X_2$ zweier unabhängiger ZV

$$F_Z(z) = \int_{-\infty}^{\infty} f_1(x) F_2(z-x)\mathrm{d}x = \int_{-\infty}^{\infty} F_1(x) f_2(z-x)\mathrm{d}x$$

mit $F_i(x)$ den Verteilungsfunktionen von X_i

Sonderfall Normalverteilung

$$X_1 \sim N(\mu_1, \sigma_1^2),\ \ X_2 \sim N(\mu_2, \sigma_2^2)\ \Rightarrow\ Z \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

Erwartungswert und Varianz von $Z = aX + b$

$$E(Z) = aE(X) + b, \quad V(Z) = a^2 V(X)$$

Erwartungswert und Varianz von $Z = aX + bY$

$$\begin{aligned} E(Z) &= aE(X) + bE(Y) \\ V(Z) &= a^2V(X) + b^2V(Y) + 2ab\mathrm{Cov}(X,Y) \end{aligned}$$

Erwartungswert und Varianz der Summe $Z_n = \sum_{i=1}^n a_i X_i$ unabhängiger ZV

$$\begin{aligned} E(Z_n) &= \sum_{i=1}^{n} a_i E(X_i) \\ V(Z_n) &= \sum_{i=1}^{n} a_i^2 V(X_i) \end{aligned}$$

Zentraler Grenzwertsatz für die Summe $Z_n = \sum_{i=1}^n a_i X_i$

$$Z_n \approx N(\mu, \sigma^2) \text{ mit } \ \mu = E(Z_n),\ \ \sigma^2 = V(Z_n)$$

Voraussetzungen: Alle X_i unabhängig und kein Einzelbetrag der Varianz größer als $\sigma^2/30$. Die X_i können beliebige (!!) diskrete oder stetige ZV mit endlicher Varianz sein

Induktive Statistik: Parameterschätzung

Schätzer für den Mittelwert $\hat{\mu} = \bar{X}$

Schätzer für den Anteilswert $\hat{\theta}_i = f_i = \dfrac{h_i}{n}$

Schätzer für die Varianz (vgl. die Fußnote auf S. 3) $\hat{\sigma}^2 = S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ bzw. $\dfrac{n}{n-1}\sum_{k=1}^{K} f_k(x_k^* - \bar{x})^2$

Schätzer für die Koeffizienten der linearen Regression $Y = aX + b$

$$\begin{aligned}\hat{a} &= \bar{Y} - \hat{b}\bar{X} \\ \hat{b} &= \frac{\sum_{i=1}^{n}(X_iY_i) - \bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \bar{X}^2},\end{aligned}$$

Gauß-Statistik (Mittelwert bei bekannter Varianz und Anteilswert) $Z = \dfrac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0;1)$

t-Statistik (Mittelwert bei unbekannter Varianz) $T = \dfrac{\bar{X} - \mu}{S}\sqrt{n} \sim T(n-1)$

χ^2-Statistik (Varianz) $Q = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Konfidenzintervall (KI) für den Mittelwert bei bekannter Varianz *Hier und im Folgenden werden "Rezepte" wie KIs und Tests als Realisierungen (also mit kleinem xquer, s etc) und nicht als Zufallsgroessen (Xquer, S, etc) formuliert*

$$\mu \in \bar{x} \pm z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$z_{1-\alpha/2}$	Tabelliertes Quantil der Gauß-Statistik,
α	Fehlerwahrscheinlichkeit (z.B. $\alpha = 5\% \Rightarrow z_{1-\alpha/2} = z_{0.975} = 1.96$)
$\sqrt{\frac{N-n}{N-1}}$	Korrekturfaktor für endliche Grundgesamtheiten mit N Elementen (=1, falls $N \gg n$)

Konfidenzintervall für den Anteilswert

$$\theta \in f \pm z_{1-\alpha/2}\,\sqrt{\frac{f(1-f)}{n}}\sqrt{\frac{N-n}{N-1}}$$

mit f der relativen Häufigkeit in der Stichprobe. Bedingung: $nf(1-f) \geq 9$

Konfidenzintervall für den Mittelwert bei unbekannter Varianz

$$\mu \in \bar{x} \pm t_{1-\alpha/2}^{(n-1)}\frac{s}{\sqrt{n}}$$

mit $t_q^{(n-1)}$ dem tabellierten q-Quantil der t-Statistik mit $(n-1)$ "Freiheitsgraden"

Konfidenzintervalle für den Regressionskoeffizienten b

$$b \in \hat{b} \pm t_{1-\alpha/2}^{(n-2)}\,\hat{\sigma}_b, \quad \hat{\sigma}_b^2 = \frac{\hat{\sigma}_R^2}{ns_x^2}, \quad \hat{\sigma}_R^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(x_i)\Big)^2$$

Konfidenzintervall der Regressionsfunktion $y = a + bx$

$$y \in \hat{a} + \hat{b}x \pm t_{1-\alpha/2}^{(n-2)}\,\frac{\hat{\sigma}_R}{\sqrt{n}}\sqrt{1 + \frac{(x-\bar{x})^2}{s_x^2}}$$

Induktive Statistik: Parametrische Tests

Testvariable für Test des Mittelwertes auf Wert μ_0 bei bekannter Varianz

$Z = \frac{\bar{X}-\mu_0}{\sigma}\sqrt{n} \sim N(0,1)$; Realisierung: z

Test des Anteilswertes auf Wert θ_0

$Z = \frac{f-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\sqrt{n}\sqrt{\frac{N-1}{N-n}}$

Test des Mittelwertes bei unbekannter Varianz

$T = \frac{\bar{X}-\mu_0}{S}\sqrt{n} \sim T(n-1)$; Realisierung: t

Test der Varianz

$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$; Realisierung: q

Test der Korrelation $\rho(x,y)$ auf Wert 0

$T_\rho = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}}} \sim T(n-2)$

Test des Koeffizienten b der linearen Regression $\hat{y}(x) = ax + b$ auf Wert b_0 sowie die Regression selbst auf den Wert $\hat{y_0}(x)$

$T_b = \sqrt{n}\left(\hat{B} - b_0\right)\frac{s_x}{\hat{\sigma}_R} \sim T(n-2)$

$T_{y(x)} = \left(\hat{Y}(x) - y_0(x))\right)\frac{s_x\sqrt{n}}{\hat{\sigma}_R\sqrt{s_x^2+(x-\bar{x})2}} \sim T(n-2)$

$\hat{A} = \bar{Y} - \hat{B}\bar{x}, \quad \hat{B} = \frac{1}{ns_x^2}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)(x_i - \bar{x})$

$\hat{\sigma}_R^2 = \frac{1}{n-2}\sum_{n=1}^n \left(Y_i - \hat{Y}(x_i)\right)^2$, $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

Zweiseitiger Test

Nullhypothese $\mu = \mu_0$ bzw. $\theta = \theta_0$ bzw. $\rho_{xy} = 0$ kann abgelehnt werden, falls

$\lvert z\rvert > z_{1-\alpha/2}$	(Mittelwert bei bekannter Varianz und Anteilswert)
$\lvert t\rvert > t_{1-\alpha/2}^{(n-1)}$	(Mittelwert bei unbekannter Varianz)
$q \notin \left[q_{\alpha/2}^{(n-1)}, q_{1-\alpha/2}^{(n-1)}\right]$	(Varianz)
$\lvert t\rvert > t_{1-\alpha/2}^{(n-2)}$	(Korrelation).

Einseitiger Test auf "größer"

Nullhypothese $\mu > \mu_0$ bzw. $\theta > \theta_0$ bzw. $\rho_{xy} > 0$ kann abgelehnt werden, falls

$z < -z_{1-\alpha}$	(Mittelwert bei bekannter Varianz und Anteilswert)
$t < -t_{1-\alpha}^{(n-1)}$	(Mittelwert bei unbekannter Varianz)
$q < q_\alpha^{(n-1)}$	(Varianz)
$t < -t_{1-\alpha}^{(n-2)}$	(Korrelation).

Einseitiger Test auf "kleiner"

Nullhypothese $\mu < \mu_0$ bzw. $\theta < \theta_0$ bzw. $\rho_{xy} < 0$ kann abgelehnt werden, falls

$z > z_{1-\alpha}$	(Mittelwert bei bekannter Varianz und Anteilswert)
$t > t_{1-\alpha}^{(n-1)}$	(Mittelwert bei unbekannter Varianz)
$q > q_{1-\alpha}^{(n-1)}$	(Varianz)
$t > t_{1-\alpha}^{(n-2)}$	(Korrelation).

Induktive Statistik: Nichtparametrische Tests

Testvariable für den χ^2-Anpassungstest

$$Q = \sum_{k=1}^{K} \left(\frac{(h_k - h_k^e)^2}{h_k^e} \right) = \sum_{k=1}^{K} \left(\frac{h_k^2}{h_k^e} \right) - n \sim \chi^2(K-1-r)$$

r	Zahl der zu schätzenden Parameter,
$n = \sum_k h_k$	Zahl der Beobachtungen,
K	Zahl der Klassen,
h_k^e	theoretischer Mittelwert für abs. Häufigkeit h_k bei Zutreffen der Nullhypothese

Bedingung für den χ^2-Anpassungstest

$h_k^e \geq 5$ für alle Klassen k

Testergebnis

H_0 nicht verwerfbar, falls für die Realisierung q von Q gilt: $q < q_{1-\alpha}^{(K-1-r)}$

Testvariable für den χ^2-Unabhängigkeitstest

$$Q = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \left(\frac{(h_{ij} - h_{ij}^{\mathrm{e}})^2}{h_{ij}^{\mathrm{e}}} \right) = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \left(\frac{h_{ij}^2}{h_{ij}^{\mathrm{e}}} \right) - n \sim \chi^2(m)$$

K_x, K_y	Zahl der Klassen,
$m = (K_x - 1)(K_y - 1)$	Zahl der Freiheitsgrade,
$h_{ij}^{\mathrm{e}} = \frac{h(x_i)h(y_j)}{n}$	Abs. Häufigkeit bei Zutreffen der Nullhypothese
$h(x_i), h(y_j)$	Spalten- und Zeilensumme
$n = \sum_j h(y_j)$	Gesamtzahl der Tabellenelemente

χ^2-Homogenitätstest (Test auf gleiche Grundgesamtheit)

Q wie beim Unabhängigkeitstest mit

K_x	Zahl der Klassen der Verteilung,
$K_y = M$	Zahl der Stichproben

Testvariable für den Kolmogorow-Smirnow-Anpassungstest (KS-Test)

$$D = \max_x \left| F(x) - F^{(0)}(x) \right| \sim D(n)$$

$F(x)$	Verteilungsfunktion der Stichprobe,
$F^{(0)}(x)$	Verteilungsfunktion bei Zutreffen der Nullhypothese
n	Stichprobenumfang

Entscheidung beim KS-Test

H_0 verwerfbar, falls für die Realisierung d von D gilt:

$$d > d_{n,1-\alpha} \approx \frac{c(\alpha)}{\sqrt{n}} = \sqrt{\frac{-\ln(\alpha/2)}{2n}}$$

Testvariable für den Zwei-Stichproben-KS-Homogenitätstest

$$D = \max_x |F_1(x) - F_2(x)| \sim D(n, m)$$

$F_1(x), F_2(x)$	Verteilungsfunktionen der zwei Stichproben,
n, m	zu $F_1(x), F_2(x)$ gehörige Stichprobenumfänge

Entscheidung beim Zwei-Stichproben KS-Test

H_0 verwerfbar, falls für die Realisierung d von D gilt:

$$d > d_{n,m,1-\alpha} \approx c(\alpha)\sqrt{\frac{n+m}{nm}} = \sqrt{\frac{-\ln(\alpha/2)(n+m)}{2nm}}$$