

Statistik Formelsammlung

Katharina Ring

4. April 2019

Inhaltsverzeichnis

1 Deskriptive Statistik	3	3 Hypothesentests	9
1.1 Kenngrößen (Parameter): Stichprobe	3	3.1 Tests für Einstichprobenprobleme	9
1.1.1 Lagemaße	3	3.1.1 Normalverteilung	9
1.1.2 Streuungsmaße	3		
1.1.3 Konzentrationsmaße	3	4 Regression	9
1.1.4 Gestaltmaße	4	4.1 Annahmen	9
1.1.5 Zusammenhangsmaße	4	4.2 Verfahren	9
1.2 Tabellen	5	4.2.1 Kleinste Quadrate (OLS)	9
1.3 Diagramme	5	4.2.2 Maximum Likelihood	10
1.3.1 Histogramm	5	4.3 Modell	10
1.3.2 QQ-Plot	5	4.3.1 lineare Einfachregression	10
1.3.3 Plot der Realisationen	5	4.3.2 Multivariate lineare Regression	10
1.3.4 Scatterplot	5	4.4 ANOVA (Streuungszerlegung)	10
		4.5 Gütemaße	10
2 Wahrscheinlichkeit	5	4.5.1 Bestimmtheitsmaß	10
2.1 Kombinatorik	5	5 Klassifikation	10
2.2 Wahrscheinlichkeitsrechnung	5	5.1 Diskriminanzanalyse (Bayes)	10
2.3 Zufallsvariablen	6		
2.4 Zufallsvektoren	6	6 Clusteranalyse	10
2.5 Verteilungen	7	7 Bayessche Statistik	10
2.5.1 Diskrete Verteilungen	7	7.0.1 Grundlagen	11
2.5.2 Stetige Verteilungen	7	7.0.2 Markov Chain / Monte Carlo	11
2.5.3 Exponentialfamilie	8		
2.6 Grenzwertsätze	9		

1 Deskriptive Statistik

1.1 Kenngrößen (Parameter): Stichprobe

1.1.1 Lagemaße

Modus Häufigster Wert von x_i . Auch zwei oder mehr Modi sind möglich (bimodal).

Median

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{falls } n \text{ gerade} \end{cases}$$

Quantile

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{falls } n\alpha \text{ ganzzahlig} \end{cases}$$

mit

$$k = \min \{x \in \mathbb{N}, \quad x > n\alpha\}$$

Minimum/Maximum

$$x_{\min} = \min_{i \in \{1, \dots, N\}} (x_i) \quad x_{\max} = \max_{i \in \{1, \dots, N\}} (x_i)$$

1.1.2 Streuungsmaße

Spannweite

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

Quartilsabstand

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

(Empirische) Varianz

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Schätzer für das zweite zentrierte Moment, inkl.

Varianzverschiebungssatz

Rechenregeln:

$$\star \operatorname{Var}(aX + b) = a^2 \cdot \operatorname{Var}(X)$$

1.1.3 Konzentrationsmaße

Gini-Koeffizient

$$G = \frac{2 \sum_{i=1}^n i x_{(i)} - (n+1) \sum_{i=1}^n x_{(i)}}{n \sum_{i=1}^n x_{(i)}} = 1 - \frac{1}{n} \sum_{i=1}^n (v_{i-1} + v_i)$$

mit

Arithmetisches Mittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Schätzer für den Erwartungswert $\mu = E[X]$
(erstes Verteilungsmoment)

Rechenregeln:

$$\star E(a + b \cdot X) = a + b \cdot E(X)$$

$$\star E(X \pm Y) = E(X) \pm E(Y)$$

Geometrisches Mittel

$$\bar{x}_G = \sqrt[n]{\sum_{i=1}^n x_i}$$

Für Wachstumsfaktoren: $\bar{x}_G = \sqrt[n]{\frac{B_n}{B_0}}$

Harmonisches Mittel

$$\bar{x}_H = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}}$$

$$\star \operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$$

(Empirische) Standardabweichung

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

Variationskoeffizient

$$\nu = \frac{s}{\bar{x}}$$

Mittlere absolute Abweichung

$$e = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Schätzer für das erste absolute zentrierte Moment

$$u_i = \frac{i}{n}, \quad v_i = \frac{\sum_{j=1}^i x_{(j)}}{\sum_{j=1}^n x_{(j)}} \quad (u_0 = 0, \quad v_0 = 0)$$

Dies sind auch die Werte für die Lorenzkurve.

$$\text{Wertebereich: } 0 \leq G \leq \frac{n-1}{n}$$

Lorenz-Münzner-Koeffizient (G normiert)

$$G^+ = \frac{n}{n-1}G$$

Wertebereich: $0 \leq G^+ \leq 1$

1.1.4 Gestaltmaße

(Empirische) Schiefe

$$\nu = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Schätzer für das dritte zentrierte Moment, normiert durch $(\sigma^2)^{\frac{3}{2}}$

(Empirische) Wölbung/Kurtosis

$$k = \left[n(n+1) \cdot \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3(n-1) \right] \cdot \frac{n-1}{(n-2)(n-3)} + 3$$

Schätzer für das vierte zentrierte Moment, normiert durch $(\sigma^2)^2$

Exzess

$$\gamma = k - 3$$

1.1.5 Zusammenhangsmaße

Für zwei nominale Variablen

χ^2 -Statistik

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}} = n \left(\sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i+}n_{+j}} - 1 \right)$$

Wertebereich: $0 \leq \chi^2 \leq n(\min(k, l) - 1)$

Phi-Koeffizient

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

Wertebereich: $0 \leq \Phi \leq \sqrt{\min(k, l) - 1}$

Cramér's V

$$V = \sqrt{\frac{\chi^2}{\min(k, l) - 1}}$$

Wertebereich: $0 \leq V \leq 1$

Kontingenzkoeffizient C

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

Wertebereich: $0 \leq C \leq \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}$

Korrigierter Kontingenzkoeffizient C_{corr}

$$C_{\text{corr}} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Wertebereich: $0 \leq C_{\text{corr}} \leq 1$

Odds-Ratio

$$OR = \frac{ad}{bc} = \frac{n_{ii}n_{jj}}{n_{ij}n_{ji}}$$

Wertebereich: $0 \leq OR < \infty$

Für zwei ordinale Variablen

Gamma nach Goodman und Kruskal

$$\gamma = \frac{K - D}{K + D}$$

Kendalls τ_b

$$\tau_b = \frac{K - D}{\sqrt{(K + D + T_X)(K + D + T_Y)}}$$

mit

$T_X = \sum_{i=m} \sum_{j < n} n_{ij}n_{mn}$ Anzahl Bindungen bzgl. X

$T_Y = \sum_{i < m} \sum_{j=n} n_{ij}n_{mn}$ Anzahl Bindungen bzgl. Y

Wertebereich: $-1 \leq \tau_b \leq 1$

Kendalls/Stuarts τ_c

$$\tau_c = \frac{2 \min(k, l)(K - D)}{n^2(\min(k, l) - 1)}$$

Wertebereich: $-1 \leq \tau_c \leq 1$

Spearman's Rangkorrelationskoeffizient

$$\rho = \frac{n(n^2 - 1) - \frac{1}{2} \sum_{j=1}^J b_j(b_j^2 - 1) - \frac{1}{2} \sum_{k=1}^K c_k(c_k^2 - 1) - 6 \sum_{i=1}^n d_i^2}{\sqrt{n(n^2 - 1) - \sum_{j=1}^J b_j(b_j^2 - 1)} \sqrt{n(n^2 - 1) - \sum_{k=1}^K c_k(c_k^2 - 1)}}$$

oder

$$\rho = \frac{s_{rg_x} r_{g_y}}{\sqrt{s_{rg_x} r_{g_x} s_{rg_y} r_{g_y}}}$$

Entspricht ohne Bindungen:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

mit

$d_i = R(x_i) - R(y_i)$ Rangdifferenz

Wertebereich: $-1 \leq \rho \leq 1$

Für zwei metrische Variablen

Korrelationskoeffizient nach Bravais-Pearson

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

mit

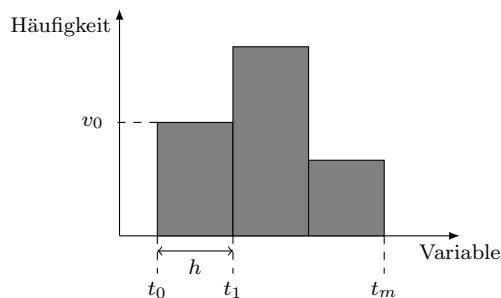
$$\begin{aligned} S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) & \text{bzw. } s_{xy} &= \frac{S_{xy}}{n} \\ S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 & \text{bzw. } s_{xx} &= \frac{S_{xx}}{n} \\ S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 & \text{bzw. } s_{yy} &= \frac{S_{yy}}{n} \end{aligned}$$

Wertebereich: $-1 \leq r \leq 1$

1.2 Tabellen

1.3 Diagramme

1.3.1 Histogramm



Stichprobe: $X = \{x_1, x_2, \dots, x_n\}$
 k -te Klasse: $B_k = [t_k, t_{k+1})$, $k = \{0, 1, \dots, m-1\}$
 Anzahl Beobachtungen in der k -ten Klasse: v_k
 Klassenbreite: $h = t_{k+1} - t_k, \forall k$

Scotts Regel

$$h^* \approx 3.5\sigma n^{-\frac{1}{3}}$$

Für annähernd normalverteilte Daten (min MSE)

1.3.2 QQ-Plot

1.3.3 Plot der Realisationen

1.3.4 Scatterplot

2 Wahrscheinlichkeit

2.1 Kombinatorik

	ohne Wiederholung	mit Wiederholung
Permutationen	$n!$	$\frac{n!}{n_1! \dots n_s!}$
Kombinationen: ohne Reihenfolge	$\binom{n}{m}$	$\binom{n+m-1}{m}$
mit Reihenfolge	$\binom{n}{m} m!$	n^m

Dabei gilt:

$$n! = n \cdot (n-1) \cdot \dots \cdot 1$$

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

2.2 Wahrscheinlichkeitsrechnung

Laplace-Wahrscheinlichkeit

$$P(A) = \frac{|A|}{|\Omega|}$$

Axiome von Kolmogorov

mathematische Definition von Wahrscheinlichkeit

- (1) $0 \leq P(A) \leq 1 \quad \forall A \in \mathcal{A}$
- (2) $P(\Omega) = 1$
- (3) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
 $\forall A_i \in \mathcal{A}, i = 1, \dots, \infty$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$

Folgerungen:

- $P(\bar{A}) = 1 - P(A)$

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(B) = \sum_{i=1}^n P(B \cap A_i)$, für A_1, \dots, A_n vollständige Zerlegung von Ω in paarweise disjunkte Ereignisse

Mises' Wahrscheinlichkeitsbegriff

frequentistische Definition von Wahrscheinlichkeit

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A(n)}{n}$$

mit n Anzahl der Wiederholungen eines Zufallsexperiments und $n_A(n)$ Anzahl an Ereignissen A

Bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{für } P(B) > 0$$

Multiplikationssatz

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

Satz von der totalen Wahrscheinlichkeit

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

2.3 Zufallsvariablen

Definition

$$Y : \Omega \rightarrow \mathbb{R}$$

Die Untermenge möglicher Werte von \mathbb{R} heißt Träger
Notation: Realisationen von Y werden als Kleinbuchstaben dargestellt. $Y = y$ bedeutet, dass Y die Realisation y angenommen hat.

Stetige und diskrete Zufallsvariablen

Ist der Träger überabzählbar unendlich, so heißt die Zufallsvariable *stetig*, sonst heißt sie *diskret*.

- **Dichte $f(\cdot)$:**

Für stetige Variablen: $P(Y \in [a, b]) = \int_a^b f_Y(y) dy$

Für diskrete Variablen lässt sich die Dichte (und andere Funktionen) wie die gleichen Funktionen für den stetigen Fall aufschreiben, wenn man $\int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y} := \sum_{k: k \leq y} P(Y = k)$ definiert. Diese Notation wird hier verwendet.

- **Verteilungsfunktion $F(\cdot)$:** $F_Y(y) = P(Y \leq y)$

Zusammenhang:

$$F_Y(y) = \int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y}$$

Ist der Träger endlich oder abzählbar unendlich, so heißt die Zufallsvariable *diskret*.

2.4 Zufallsvektoren

Dichte und Verteilungsfunktion

$$F(y_1, \dots, y_q) = P(Y_1 \leq y_1, \dots, Y_q \leq y_q)$$

$$\begin{aligned} P(a_1 \leq Y_1 \leq b_1, \dots, a_q \leq Y_q \leq b_q) \\ = \int_{a_1}^{b_1} \dots \int_{a_q}^{b_q} f(y_1, \dots, y_q) dy_1 \dots dy_q \end{aligned}$$

Marginale Dichte

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, \dots, y_k) dy_2 \dots dy_k$$

Bedingte Dichte

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f(y_1, \dots, y_2)}{f(y_2)} \quad \text{für } f(y_2) > 0$$

Satz von Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{für } P(A), P(B) > 0$$

Stochastische Unabhängigkeit

$$A, B \text{ unabhängig} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

$$X, Y \text{ unabhängig} \Leftrightarrow f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \quad \forall x, y$$

Momente

- **Erwartungswert (1. Moment):** $\mu = E(Y) = \int y f_Y(y) dy$

- **Varianz (2. zentriertes Moment):**

$$\sigma^2 = \text{Var}(Y) = E(\{Y - E(Y)\}^2) = \int (y - E(Y))^2 f(y) dy$$

$$\text{Varianzverschiebungssatz: } E(\{Y - \mu\}^2) = E(Y^2) - \mu^2$$

Beweis:

$$\begin{aligned} E(\{Y - \mu\}^2) &= E(Y^2 - 2Y\mu + \mu^2) = E(Y^2) - 2\mu^2 + \mu^2 = \\ &= E(Y^2) - \mu^2 \end{aligned}$$

- **k. Moment:** $E(Y^k) = \int y^k f_Y(y) dy$,

- **k. zentrales Moment:** $E(\{Y - E(Y)\}^k)$

Momenterzeugende Funktion

$$M_Y(t) = E(e^{tY})$$

$$\text{mit } \left. \frac{\partial^k M_Y(t)}{\partial t^k} \right|_{t=0} = E(Y^k)$$

$$\text{kumulanterzeugende Funktion } K_Y(t) = \log M_Y(t)$$

Eine Zufallsvariable ist durch ihre momenterzeugende Funktion eindeutig definiert und andersherum (solange die Momente und Kumulanten endlich sind).

Iterierter Erwartungswert

$$E(Y) = E_X(E(Y|X))$$

Beweis:

$$E(Y) = \int y f(y) dy = \int \int y f(y|x) dy f_X(x) dx = E_X(E(Y|X))$$

$$\text{Var}(Y) = E_X(\text{Var}(Y|X)) + \text{Var}_X(E(Y|X))$$

Beweis:

$$\begin{aligned}
 \text{Var}(Y) &= \int (y - \mu_Y)^2 f(y) dy \\
 &= \int (y - \mu_Y)^2 f(y|x) f(x) dy dx \\
 &= \int (y - \mu_{Y|x} + \mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx \\
 &= \int (y - \mu_{Y|x})^2 f(y|x) f(x) dy dx + \\
 &\quad \int (\mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx + \\
 &\quad 2 \int (y - \mu_{Y|x})(\mu_{Y|x} - \mu_Y) f(y|x) f(x) dy dx \\
 &= \int \text{Var}(Y|x) f(x) dx + \int (\mu_{Y|x} - \mu_Y)^2 f(x) dx \\
 &= E_X(\text{Var}(Y|X)) + \text{Var}_X(E(Y|X))
 \end{aligned}$$

2.5 Verteilungen

2.5.1 Diskrete Verteilungen

Diskrete Gleichverteilung

$$\begin{aligned}
 Y &\sim U(\{y_1, \dots, y_k\}), y \in \{y_1, \dots, y_k\} \\
 P(Y = y_i) &= \frac{1}{k}, i = 1, \dots, k \\
 E(Y) &= \frac{k+1}{2}, \text{Var}(Y) = \frac{k^2-1}{12}
 \end{aligned}$$

Binomialverteilung Erfolge in unabhängigen Versuchen

$$\begin{aligned}
 Y &\sim \text{Bin}(n, \pi) \text{ mit } n \in \mathbb{N}, \pi \in [0, 1], y \in \{0, \dots, n\} \\
 P(Y = y|\pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\
 E(Y|\pi, n) &= n\pi, \text{Var}(Y|\pi, n) = n\pi(1-\pi)
 \end{aligned}$$

Poissonverteilung Zählmodelle für seltene Ereignisse

Immer nur ein Ereignis pro Zeitpunkt, Eintreten der Ereignisse ist unabhängig von bisheriger Geschichte, mittlere Anzahl der Ereignisse pro Zeit ist konstant und proportional zur Länge des betrachteten Zeitintervalls.

$$Y \sim \text{Po}(\lambda) \text{ mit } \lambda \in [0, +\infty], y \in \mathbb{N}_0$$

$$P(Y = y|\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

$$E(Y|\lambda) = \lambda, \text{Var}(Y|\lambda) = \lambda$$

Häufig wird die Varianz durch das Poisson-Modell unterschätzt, es liegt Überdispersion vor.

Approximation der Binomialverteilung für kleine p

Geometrische Verteilung

$$Y \sim \text{Geom}(\pi) \text{ mit } \pi \in [0, 1], y \in \mathbb{N}_0$$

$$P(Y = y|\pi) = \pi(1-\pi)^{y-1}$$

$$E(Y|\pi) = \frac{1}{\pi}, \text{Var}(Y|\pi) = \frac{1-\pi}{\pi^2}$$

Negative Binomialverteilung

$$Y \sim \text{NegBin}(\alpha, \beta) \text{ mit } \alpha, \beta \geq 0, y \in \mathbb{N}_0$$

$$P(Y = y|\alpha, \beta) = \binom{\alpha+y-1}{\alpha-1} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y$$

$$E(Y|\alpha, \beta) = \frac{\alpha}{\beta}, \text{Var}(Y|\alpha, \beta) = \frac{\alpha}{\beta^2}(\beta+1)$$

2.5.2 Stetige Verteilungen

Stetige Gleichverteilung

$$\begin{aligned}
 Y &\sim U(a, b) \text{ mit } a, b \in \mathbb{R}, a \leq b, y \in [a, b] \\
 p(y|a, b) &= \frac{1}{b-a} \\
 E(Y|a, b) &= \frac{a+b}{2}, \text{Var}(Y|a, b) = \frac{(b-a)^2}{12}
 \end{aligned}$$

Univariate Normalverteilung symmetrisch mit μ und σ^2

$$Y \sim N(\mu, \sigma^2) \text{ mit } \mu \in \mathbb{R}, \sigma^2 > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \mu, \text{Var}(Y|\mu, \sigma^2) = \sigma^2$$

Multivariate Normalverteilung symmetrisch mit μ und Σ

$$Y \sim N(\mu, \Sigma) \text{ mit } \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \text{ s.p.d., } y \in \mathbb{R}^d$$

$$p(y|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)$$

$$E(Y|\mu, \Sigma) = \mu, \text{ Var}(Y|\mu, \Sigma) = \Sigma$$

Log-Normalverteilung

$$Y \sim \text{LogN}(\mu, \sigma^2) \text{ mit } \mu \in \mathbb{R}, \sigma^2 > 0, y > 0$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}y} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$\text{Var}(Y|\mu, \sigma^2) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

Zusammenhang: $\log(Y) \sim N(\mu, \sigma^2) \Rightarrow Y \sim \text{LogN}(\mu, \sigma^2)$

Nichtzentrale Studentverteilung statistische Tests für μ mit unbekannter (geschätzter) Varianz und ν Freiheitsgraden

$$Y \sim t_\nu(\mu, \sigma) \text{ mit } \mu \in \mathbb{R}, \sigma^2, \nu > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\sqrt{\nu\pi\sigma})} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

$$E(Y|\mu, \sigma^2, \nu) = \mu \text{ für } \nu > 1,$$

$$\text{Var}(Y|\mu, \sigma^2, \nu) = \sigma^2 \frac{\nu}{\nu-2} \text{ für } \nu > 2$$

Zusammenhang: $Y|\theta \sim N(\mu, \frac{\sigma^2}{\theta}), \theta \sim \text{Ga}(\frac{\nu}{2}, \frac{\nu}{2}) \Rightarrow Y \sim t_\nu(\mu, \sigma)$

Betaverteilung

$$Y \sim \text{Be}(a, b) \text{ mit } a, b > 0, y \in [0, 1]$$

$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

$$E(Y|a, b) = \frac{a}{a+b},$$

$$\text{Var}(Y|a, b) = \frac{ab}{(a+b)^2(a+b+1)},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{a+b-2} \text{ für } a, b > 1$$

2.5.3 Exponentialfamilie

Definition

Zur Exponentialfamilie gehören alle Verteilungen, deren Dichte wie folgt geschrieben werden kann:

$$f_Y(y, \theta) = \exp^{t^T(y)\theta - \kappa(\theta)} h(y)$$

mit $h(y) \geq 0$, $t(y)$ Vektor der kanonischen Statistiken, θ Parametervektor und $\kappa(\theta)$ Normalisationskonstante.

Normalisierungskonstante

$$1 = \int \exp^{t^T(y)\theta} h(y) dy \exp^{-\kappa(\theta)}$$

$$\Leftrightarrow \kappa(\theta) = \log \int \exp^{t^T(y)\theta} h(y) dy$$

$\kappa(\theta)$ ist die kumulanterzeugende Funktion, somit $\frac{\partial \kappa(\theta)}{\partial \theta} = E(t(Y))$ und $\frac{\partial^2 \kappa(\theta)}{\partial \theta^2} = \text{Var}(t(Y))$

Gammaverteilung

$$Y \sim \text{Ga}(a, b) \text{ mit } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by)$$

$$E(Y|a, b) = \frac{a}{b},$$

$$\text{Var}(Y|a, b) = \frac{a}{b^2},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{b} \text{ für } a \geq 1$$

Invers-Gammaverteilung

$$Y \sim \text{IG}(a, b) \text{ mit } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-\frac{b}{y})$$

$$E(Y|a, b) = \frac{b}{a-1} \text{ für } a > 1,$$

$$\text{Var}(Y|a, b) = \frac{b^2}{(a-1)^2(a-2)} \text{ für } a \geq 2,$$

$$\text{mod}(Y|a, b) = \frac{b}{a+1}$$

Zusammenhang: $Y^{-1} \sim \text{Ga}(a, b) \Leftrightarrow Y \sim \text{IG}(a, b)$

Exponentialverteilung Zeit zwischen Poisson-Ereignissen

$$Y \sim \text{Exp}(\lambda) \text{ mit } \lambda > 0, y \geq 0$$

$$p(y|\lambda) = \lambda \exp(-\lambda y)$$

$$E(Y|\lambda) = \frac{1}{\lambda}, \text{ Var}(Y|\lambda) = \frac{1}{\lambda^2}$$

Chi-Quadrat-Verteilung quadrierte standardnormalverteilte Zufallsvariablen mit ν Freiheitsgraden

$$Y \sim \chi^2(\nu) \text{ mit } \nu > 0, y \in \mathbb{R}$$

$$p(y|\nu) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

$$E(Y|\nu) = \nu, \text{ Var}(Y|\nu) = 2\nu$$

Mitglieder

- **Poissonverteilung**
- **Geometrische Verteilung**
- **Exponentialverteilung**
- **Normalverteilung** $t(y) = \left(-\frac{y^2}{2}, y\right)^T, \theta = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)^T$,
 $h(y) = \frac{1}{\sqrt{2\pi}}, \kappa(\theta) = \frac{1}{2} \left(-\log \frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)$
- **Gammaverteilung**
- **Chi-Quadrat-Verteilung**
- **Betaverteilung**

2.6 Grenzwertsätze

Gesetz der großen Zahlen

Zentraler Grenzwertsatz

$$Z_n \xrightarrow{d} N(0, \sigma^2)$$

mit $Z_n = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$ und Y_i i.i.d. mit $\mu = 0$ und Varianz σ^2

Beweis:

Für eine normalverteilte Zufallsvariable $Z \sim N(\mu, \sigma^2)$ gilt $K_Z(t) = \mu t + \frac{1}{2}\sigma^2 t^2$. Die ersten beiden Ableitungen $\left. \frac{\partial^k K_Z(t)}{\partial t^k} \right|_{t=0}$ entsprechen μ und σ . Alle anderen Momente sind null.

Für $Z_n = (Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}$ gilt:

$$\begin{aligned} M_{Z_n}(t) &= E\left(e^{t(Y_1+Y_2+\dots+Y_n)/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}} \cdot e^{tY_2/\sqrt{n}} \cdot \dots \cdot e^{tY_n/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}}\right) E\left(e^{tY_2/\sqrt{n}}\right) \dots E\left(e^{tY_n/\sqrt{n}}\right) \\ &= M_Y^n(t/\sqrt{n}) \end{aligned}$$

Analog gilt: $K_{Z_n}(t) = nK_Y(t/\sqrt{n})$.

$$\begin{aligned} \left. \frac{\partial K_{Z_n}(t)}{\partial t} \right|_{t=0} &= \frac{n}{\sqrt{n}} \left. \frac{\partial K_Y(t)}{\partial t} \right|_{t=0} = \sqrt{n}\mu \\ \left. \frac{\partial^2 K_{Z_n}(t)}{\partial t^2} \right|_{t=0} &= \frac{n}{n} \left. \frac{\partial^2 K_Y(t)}{\partial t^2} \right|_{t=0} = \sigma^2 \end{aligned}$$

Mithilfe der Taylorreihe können wir $K_{Z_n}(t) = 0 + \sqrt{n}\mu t + \frac{1}{2}\sigma^2 t^2 + \dots$ schreiben, wobei die Terme in ... alle für $n \rightarrow \infty$ gegen 0 gehen.

Damit gilt $K_{Z_n}(t) \xrightarrow{n \rightarrow \infty} K_Z(t)$ mit $Z \sim N(\sqrt{n}\mu, \sigma^2)$.

3 Hypothesentests

3.1 Tests für Einstichprobenprobleme

3.1.1 Normalverteilung

4 Regression

μ gesucht, σ^2 bekannt (Einfacher Gauß-Test)

4.1 Annahmen

4.2 Verfahren

4.2.1 Kleinste Quadrate (OLS)

KQ-Schätzer (Einfachregression)

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r \sqrt{\frac{S_{yy}}{S_{xx}}}$$

Beweis:

$$\begin{aligned} Cov(x, y) &= Cov(x, \hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \hat{\beta}_1 Var(x) \\ &\iff \hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} \end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Beweis:

$$E[y] = E[\hat{\beta}_0 + \hat{\beta}_1 x + \hat{e}] \iff \hat{\beta}_0 = E[y] - \hat{\beta}_1 E[x]$$

4.2.2 Maximum Likelihood

4.3 Modell

4.3.1 lineare Einfachregression

Theoretisches Modell

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Empirisches Modell

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

Eigenschaften der Regressionsgeraden

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ \hat{e}_i &= y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - (\bar{y} + \hat{\beta}_1 (x_i - \bar{x})) \\ \sum_{i=1}^n \hat{e}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) \\ &= n\bar{y} - n\bar{y} - \hat{\beta}_1 (n\bar{x} - n\bar{x}) = 0 \\ \bar{\hat{y}} &= \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} (n\bar{y} + \hat{\beta}_1 (n\bar{x} - n\bar{x})) = \bar{y}\end{aligned}$$

4.3.2 Multivariate lineare Regression

4.4 ANOVA (Streuungszerlegung)

$$SS_{Total} = SS_{Explained} + SS_{Residual}$$

mit

$$\begin{aligned}SS_{Total} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ SS_{Explained} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ SS_{Residual} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = S_{yy} - \hat{\beta}^2 S_{xx}\end{aligned}$$

4.5 Gütemaße

4.5.1 Bestimmtheitsmaß

$$R^2 = \frac{SS_{Explained}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}} = r^2$$

Wertebereich: $0 \leq R^2 \leq 1$

5 Klassifikation

5.1 Diskriminanzanalyse (Bayes)

6 Clusteranalyse

7 Bayessche Statistik

7.0.1 Grundlagen

Bayes-Formel

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{für } P(A), P(B) > 0$$

oder allgemeiner:

$$\begin{aligned} f(\theta|X) &= \frac{f(X|\theta) \cdot f(\theta)}{\int f(X|\tilde{\theta})f(\tilde{\theta})d\tilde{\theta}} \\ &= C \cdot f(X|\theta) \cdot f(\theta) \quad \text{wähle } C \text{ so, dass } \int f(\theta|X) = 1 \\ &\propto f(X|\theta) \cdot f(\theta) \end{aligned}$$

Punktschätzer

Kreditibilitätsintervall

Sensitivitätsanalyse

Prädiktive Posteriori

$$f(x_Z|\mathbf{x}) = \int f(x_Z, \lambda|\mathbf{x})d\lambda = \int f(x_Z|\lambda)p(\lambda|\mathbf{x})$$

Uninformative Priori

$f(\theta) = \text{const.}$ für $\theta > 0$, damit: $f(\theta|X) = C \cdot f(X|\theta)$
(Da $\int f(\theta) = 1$ so nicht möglich, ist das eigentlich keine Dichte)

Konjugierte Priori

Wenn die Priori- und die Posteriori-Verteilung denselben Typ hat für eine gegebene Likelihoodfunktion, so nennt man sie konjugiert.

Binomial-Beta-Modell:

- Priori $\sim Be(\alpha, \beta)$
- $X \sim Binom(n, p, k)$
- Posteriori $\sim Be(\alpha + k, \beta + n - k)$

7.0.2 Markov Chain / Monte Carlo