Statistik Formelsammlung

Katharina Ring

11. März 2016

Inhaltsverzeichnis

1 Deskriptive Statistik

1.1 Kenngrößen (Parameter: Zufallsvariablen

1.2 Kenngrößen (Parameter): Stichprobe

1.2.1 Lagemaße

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

Median

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

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 $\in \mathbb{N}$, x $>$ nα

Minimum/Maximum

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

Arithmetisches Mittel

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

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

Rechenregeln:

- ⋆ $E(a + b \cdot X) = a + b \cdot E(X)$
- ⋆ $E(X \pm Y) = E(X) \pm E(Y)$

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

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

1.2.2 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:

- ⋆ $Var(aX + b) = a^2 \cdot Var(X)$
- ⋆ $Var(X \pm Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

(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

1.2.3 Konzentrationsmaße

Gini-Koeffizient

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

mit

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

Dies sind auch die Werte für die Lorenzkurve.

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

Lorenz-Münzner-Koeffizient (G normiert)

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

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

1.2.4 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 $(\sigma^2)^{\frac{2}{3}}$

(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

$$\gamma = k - 3$$

1.2.5 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: $0 \leq \Phi \leq \sqrt{\min(k,l) - 1}$

Cramérs 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_{korr}

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

Wertebereich: $0 \leq C_{korr} \leq 1$

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

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

$K = \sum_{i<m} \sum_{j<n} n_{ij} n_{mn}$ Anzahl konkordanter Paare
$D = \sum_{i<m} \sum_{j>n} n_{ij} n_{mn}$ Anzahl diskordanter Paare

Wertebereich: $-1 \leq \gamma \leq 1$

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

Spearmans Rangkorrelationskoeffizient

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

Entspricht ohne Bindungen:

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

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

Wertebereich: $-1 \leq R \leq 1$

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

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2 \qquad \text{bzw. } s_{xy} = \frac{S_{xy}}{n}$$

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

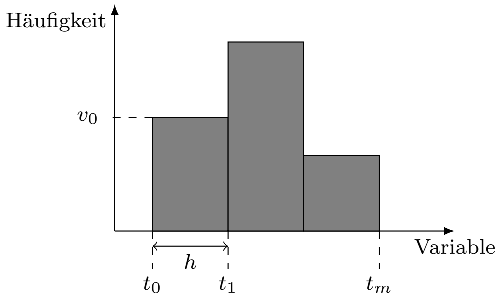
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{bzw. } s_{yy} = \frac{S_{yy}}{n}$$

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

1.3 Tabellen

1.4 Diagramme

1.4.1 Histogramm



Stichprobe: $X = \{x_1, x_2, ...; x_n\}$
k-te Klasse: $B_k = [t_k, t_{k+1}), k = \{0, 1, ..., m-1\}$
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)

1.4.2 QQ-Plot

1.4.3 Plot der Realisationen

1.4.4 Scatterplot

2 Wahrscheinlichkeit

2.1 Kombinatorik

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

Dabei gilt:
$n! = n \cdot (n-1) \cdot ... \cdot 1$
$\binom{n}{m} = \frac{n!}{m!(n-m)!}$

2.2 Wahrscheinlichkeitsrechnung

Laplace-Wahrscheinlichkeit

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

Axiome von Kolmogorov

(1) $0 \leq P(A) \leq 1$
(2) $P(\Omega) = 1$
(3) $P(A \cup B) = P(A) + P(B)$ (für A und B disjunkt)

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 $A_i, ..., A_n$ vollständige Zerlegung von Ω in paarweise disjunkte Ereignisse

Bedingte Wahrscheinlichkeit

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

Multiplikationssatz

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

Satz von der totalen Wahrscheinlichkeit

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

Satz von Bayes

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

Statistische Unabhängigkeit

$$\text{A, B unabhängig} \Leftrightarrow P(A \cap B) = P(A) + P(B)$$
$$\text{X, Y unabhängig} \Leftrightarrow f_{XY}(x,y) = f_X(x) \cdot f_Y(y) \qquad \forall x, y$$

3 Hypothesentests

4 Regression

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:
$Cov(x,y) = Cov(x, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}) = \hat{\beta}_1 Var(x)$
$\iff \hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$

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

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

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

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

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