Statistik Zusammenfassung

Nils Weiß - Alexander Strobl

29. Juni 2014

Inhaltsverzeichnis

1 Grundlagen

1.1 Häufigkeitsverteilungen

Eigenschaft	Beschreibung
Merkmalsträger	Objekt von Interesse bei empirischer Untersuchung
Gesamtheit	Menge der relevanten Merkmalsträger. Die Anzahl nennt man Umfang der Gesamtheit
Mikrodaten	Daten, welche ausgewertet werden sollen
Häufigkeitsverteilung	Ausprägungen der einzelnen Merkmalsträger

Abbildung 1: Begriffserklärungen: Häufigkeitsverteilung

Merkmalsausprägung x_i	absolute Häufigkeiten n_i	relative Häufigkeiten f_i
1	6	0.3 / 30%
2	7	0.35 / 35%
3	4	0.2 / 20%
4	2	0.1 / 10 %
5	1	0.05 / 5 %
$\sum$	20	1 / 100 %

Abbildung 2: Beispiel: Häufigkeitsverteilung von Noten

1.2 Kummulierte Häufigkeiten

1.2.1 Absolut

Summe der "ersten"i absoluten Häufigkeiten

$$N_i = \sum_{j=1}^{i} n_j$$

1.2.2 Relativ

Summe der "ersten"i relativen Häufigkeiten

$$F_i = \sum_{j=1}^{i} f_j$$

1.2.3 Empirische Verteilungsfunktion

Summe über alle i, für die $x_i \leqslant x$ ist
Es werden die relativen Häufigkeiten f_i all jener Ausprägungen summiert, die höchsten gleich x sind

$$F(x) = \sum_{\{i|x_i \leqslant x\}} f_i$$

Klasse i x_i	Klassen-obergrenze x_i^o	abs. Häufigkeiten n_i	rel. Häufigkeiten f_i	emp. Verteilungsfunktion a. d. Klassenobergrenze $F(x_i^0)$
1	29	7	0.01165	1.17 %
2	39	59	0.09817	10.98 %
3	49	127	0.21131	32.11 %
4	54	120	0.19967	52.08 %
5	59	146	0.24293	76.37 %
6	64	112	0.18636	95.01 %
7	73	30	0.04992	100.00 %
$\sum$		601	1	

Abbildung 3: klassierte Altersverteilung

1.3 Maßzahlen zur Beschreibung einer Verteilung

1.3.1 Modus

(auch: Modalwert, häufigster Wert)

Bezeichnet das Merkmal x_i mit der größten absoluten Häufigkeit n_i bzw. der größten relativen Häufigkeit f_i.

1.3.2 Median - Quantile

Median: $x_{0,5}$= 50% Quantil

Wichtige Quantile: $x_{0,25}$, $x_{0,75}$

1.3.3 Arithmetisches Mittel

(auch: Mittelwert, Durchschnittswert)

$\overline{x} = \frac{\textbf{Summe aller Merkmalswerte}}{\textbf{Anzahl aller Merkmalswerte}}$

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

$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i$

$\overline{x} = \sum_{i=1}^{k} x_i f_i$

$\overline{x} = \frac{1}{n} \sum_{i=1}^{r} \overline{x_i} n_i$

$\overline{x} = \sum_{i=1}^{r} \overline{x_i} f_i$

2 Streuung und Konzentration

2.0.4 Spannweite

(Einfachstes Streuungsmaß, Differenz zwischen größtem und kleinstem auftretenden Merkmalswert)

Spannweite $= max(x_i) - min(x_i)$

2.0.5 Quartilsabstand

(Spannweite der mittleren 50% der Merkmalsträger)

Quartilsabstand $= x_{0,75} - x_{0,25}$

2.0.6 Varianz

(Mittlere quadratische Abweichung vom Mittelwert)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \sum_{i=1}^{k}(x_i - \overline{x})^2 f_i$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i - \overline{x}^2 = \sum_{i=1}^{k}x_i^2 f_i - \overline{x}^2$$

$$s^2 = \overline{x^2} - \overline{x}^2$$

Varianz bei Vorliegen von Teilgesamtheiten

$$s^2 = \frac{1}{n}\sum_{i=1}^{r}s_i^2 n_i + \frac{1}{n}\sum_{i=1}^{r}(\overline{x_i} - \overline{x})^2 n_i$$

2.0.7 Standardabweichung

$s = \sqrt{s^2}$

2.0.8 Standardisierte Merkmale

Ein Merkmal, für dessen Verteilung $\overline{x} = 0$ und $s^2 = 1$ gilt, heißt standardisiert.

2.0.9 Kovarianz

(von X und Y)

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{x}\ \overline{y}$$

2.0.10 Korrelationskoeffizient

(nach Bravais-Pearson)

$$\frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Es gilt: $-1 \leqslant r \leqslant 1$

$r = 0$: kein linearer Zusammenhang $r = 1$: steigende Gerade $r = -1$: fallende Gerade

2.0.11 Regressionsgerade

$\hat{y} = \hat{\alpha} + \hat{\beta}x$

$\hat{\beta} = \frac{s_{XY}}{s_X}$

$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$

3 Wahrscheinlichkeitstheorie

Eigenschaft	**Formel**
Wahrscheinlichkeitsraum	Ω
LaPlace-Wahrscheinlichkeit	$P(A) = \frac{\text{Anz. Elem in A}}{\text{Anz. Elem in } \Omega} = \frac{\lvert A\rvert}{\lvert\Omega\rvert}$
Bedingte Wahrscheinlichkeiten	$P(B\lvert A) = \frac{P(A \cap B)}{P(A)}$
Multiplikationssatz	$P(A \cap B) = P(A) * P(B\lvert A)$
Additionssatz (bel. Ereignisse)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Additionssatz (ausschließende Ereignisse)	$P(A \cup B) = P(A) + P(B)$ // $\mathbf{(\ A \cap B = \emptyset)}$
Abhängigkeit	$P(A \cup B) = P(A) + P(B) - P(A)P(B)$
Unabhängigkeit	$P(A \cap B) = P(A)P(B)$
Totale Wahrscheinlichkeit	$P(B) = \sum_{i=1}^{m} P(B \cap A_) = P(A_i)P(B\lvert A_i)$
Satz von Bayes	$P(A_j\lvert B) = \frac{P(B\lvert A_j)P(A_j)}{\sum_{i=1}^{m} P(B\lvert A_i)P(A_i)}$

Tabelle 1: Begriffserklärungen: Wahrscheinlichkeitstheorie

4 Diskrete Zufallsvariablen

Eigenschaft	Formel	Beschreibung
Verteilungsfunktion	$F_x(x) = P(X \leq x)$	definiert die Wahrscheinlichkeit der Zufallsvariable X, dass X höchstens den Wert x annimmt
Unabhängigkeit	$P(X_1 = x_1, X_2 = x_2)$ $= P(X_1 = x_1) * P(X_2 = x_2)$	gilt ebenfalls für andere Operationen wie z.B. $\leq$
Erwartungswert	$\mathrm{E(X)} = \mu_x = \mu$ $= \sum_{i=1}^{k} x_i p_i = \sum_{i=1}^{k} x_i * P(X = x_i)$	Ist der Mittelwert von X
		Weitere Rechenregeln:
	$\mathrm{E(Y)} = \mathrm{E(g(X))} = \sum_i g(x_i) p_i$	Wenn g(x) eine reelle Funktion und Y = g(X)
	$\mathrm{E(X + Y)} = \mathrm{E(X)} + \mathrm{E(Y)}$ $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$	
Varianz	$Var(X) = E((X - \mu_x)^2)$	
	$= \sum_{i=1}^{k} (x_i - \mu_x)^2 * p_i$	
	$= E(X^2) - E(X)^2$	

Tabelle 2: Begriffserklärungen: Diskrete Zufallsvariablen

5 Stetige Zufallsvariablen

Eigenschaft	Formel	Beschreibung
Definition	$F_x(x) = \int_{-\infty}^{x} f(t)dt$	Verteilungsfunktion, $f(t)$ = Dichtefunktion
	$P(X = x) = 0$	Wahrscheinlichkeit für einen Wert gleich x ist immer 0
	$P(x_1 \leq X \leq x_2) = F_x(x_2) - F_x(x_1)$	$F_x^{'}(x) = f_x(x)$ Die Dichtefunktion ist die Ableitung der Verteilungsfunktion
Erwartungswert μ_x	$E(X) = \int_{-\infty}^{+\infty} x * f(x)dx$	Die Dichtefunktion $f(x)$ wird nie verändert! $E(\frac{1}{X}) = \int \frac{1}{X} * f(x)dx$
Rechenregeln	$E(Y) = E(g(X)) = \int_{a}^{b} g(x) * f(x)dx$	$g(x)$ ist eine reelle Funktion
	$E(aX + b) = a * E(X) + b$	lineare Transformation
	$E(X + Y) = E(X) + E(Y)$	
Modus	$F_x(x_p) = p$	Die Wahrscheinlickeit, dass X höchstens den Wert x_p annimmt, ist mind. $p/100\%$
Varianz	$Var(X) = \int_{-\infty}^{+\infty} (x - \mu_x)^2 * f(x)dx$	Standardabweichung ist $\sqrt{Var(X)}$
Rechenregeln	vgl. Diskrete Zufallsvariablen	

Tabelle 3: Begriffserklärungen: Stetige Zufallsvariablen

6 Diskrete Verteilungen

diskrete Verteilungen

Verteilungsname	Wahrscheinlichkeitsgewicht/ Zähldichte	Erwartungs -wert $E(X)$	Varianz Var(X)	Anwendung
Bernoulli-Verteilung Parameter $0 < p < 1$	$P(X = 1) = p$, $P(X = 0) = 1 - p$	p	$p \cdot (1 - p)$	$X = 1$ = Erfolg, $X = 0$ = Misserfolge z.B. beim einmaligen Werfen eines Würfels eine 6 geworfen (=Erfolg), hier $p = \frac{1}{6}$.
Binomialverteilung Parameter $0 < p < 1$	$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$ für $k \in \{0, 1, 2, 3, \ldots, n\}$	$n \cdot p$	$n \cdot p \cdot (1 - p)$	X = Anzahl der Erfolge bei n identischen Bernoulli-Experimenten z.B. X = Anzahl geworfener 6en beim n-maligen Wurf eines fairen Würfels (hier $p = \frac{1}{6}$). z.B. X = Anzahl gezogener roter Kugeln, beim Ziehen **mit** Zurücklegen von n Kugeln aus einer Urne M roten und $N-M$ sonstigen Kugeln, wobei $p = \frac{M}{N}$
Diskrete Gleichverteilung auf $\{1, 2, 3, \ldots, n\}$	$P(X = k) = \frac{1}{n}$ für $k \in \{1, 2, 3, \ldots, n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	z.B. ein Wurf mit einem Würfel beschreibt X die geworfene Augenzahl, hier $n = 6$.
Geometrische Verteilung Parameter $0 < p < 1$	$P(X = k) = (1 - p)^{k-1} \cdot p$ für $k \in \{1, 2, 3, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	X beschreibt die Wartezeit auf den ersten Erfolg, beim fortgesetzten Ausführen eines Bernoulli-Experimentes z.B. beim Würfeln warten auf die erste 6, d.h. $X = k$ bedeutet die erste 6 wurde im k-ten Wurf geworfen.
Hypergeometrische Vert. N Anzahl Kugeln in der Urne M Anzahl roter Kugeln n Anzahl zu ziehende Kugeln k Anzahl roter Kugeln unter den gezogenen Kugeln	$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$ für $k \in \{\max(0, n–(N–M)), \ldots, \min(n, M)\}$	$n \cdot \frac{M}{N}$	$n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}$	X = Anzahl gezogener roter Kugeln, beim Ziehen **ohne** Zurücklegen von n Kugeln aus einer Urne M roten und $N - M$ sonstigen Kugeln
Poisson-Verteilung Parameter $\lambda > 0$	$P(X = k) = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}$ für $k \in \{0, 1, 2, 3, \ldots\}$	λ	λ	Anzahl Ereignisse in einem vorgegebenen Zeitintervall z.B. Anzahl radioaktiver Zerfälle, Anzahl Blitzschläge auf einer gegebenen Fläche,...

7 Stetige Verteilungen

stetige Verteilungen

Verteilungsname	Dichte	Verteilungsfunktion	Median	$E[X]$	$\mathrm{Var}(X)$	Anwendung
stetige Gleichverteilung auf $[a,b]$	$f(x)=\begin{cases}\frac{1}{b-a} & \text{für } x\in[a,b]\\ 0 & \text{sonst}\end{cases}$	$F(x)=\begin{cases}0 & \text{für } x<a\\ \frac{x-a}{b-a} & \text{für } x\in[a,b]\\ 1 & \text{für } x>b\end{cases}$	$\frac{a+b}{2}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	stetiges Analogon zur diskreten Gleichverteilung z.B. jede reelle Zahl aus dem Intervall $[a,b]$ wird mit gleicher Wahrscheinlichkeit gewählt.
Exponentialverteilung Parameter $\alpha>0$	$f(x)=\begin{cases}\alpha\cdot \mathrm{e}^{-\alpha\cdot x} & \text{für } x\geq 0\\ 0 & \text{sonst}\end{cases}$	$F(x)=\begin{cases}0 & \text{für } x<0\\ 1-\mathrm{e}^{-\alpha\cdot x} & \text{für } x\geq 0\end{cases}$	$\frac{\ln(2)}{\alpha}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$	stetiges Analogon zur geometrischen Verteilung Warten auf das erste/nächste Eintreffen eines Ereignisses z.B. Warten auf den Ausfall einer Glühbirne
Normalverteilung Parameter $\mu\in\mathbb{R}$ und $\sigma>0$	$f(x)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot\exp(-\frac{(x-\mu)^2}{2\cdot\sigma^2})$	$F(x)$ kann nicht als Funktion hingeschrieben werden, vgl. Tabelle	μ	μ	σ^2	Wenn auf etwas viele verschiedene zufällige Einflussfaktoren einwirken, ist das Ergebnis in etwa normalverteilt, z.B. die Körpergröße von Männern (Ernährung, Veranlagung,...) Wird auch zur Approximation von Binomial- und Poissonverteilungen verwendet