Formelsammlung

zur Vorlesung

Statistik I

PD Dr. C. Heumann

Häufigkeitsverteilungen

Darstellungsformen von Daten

- Rohdaten:

$$x_1, x_2, \ldots, x_n$$

n	Anzahl der Beobachtungen
x_i	Merkmalsausprägung des i-ten Objekts

- Geordnete Daten:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

n	Anzahl der Beobachtungen
$x_{(i)}$	i-t größte Merkmalsausprägung

- Häufigkeitstabelle für unterschiedliche Merkmalsausprägungen:

$$a_1, a_2, \ldots, a_k$$

k	Anzahl unterschiedlicher Merkmalsausprägungen
a_j	j-te Merkmalsausprägung
n_j	absolute Häufigkeit von a_j
$f_j = f(a_j) = \frac{n_j}{n}$	relative Häufigkeit von a_j

- Klassierte Daten (bei stetigen Merkmalen), z.B.:

$$| \underbrace{x_{(1)}, x_{(2)}, x_{(3)}}_{K_1} | \underbrace{x_{(4)}, x_{(5)}}_{K_2} | x_{(6)}, \ldots | \underbrace{\ldots, x_{(n)}}_{K_k} |$$

k	Anzahl der Klassen
K_j	j-te Klasse
e_{j-1}	untere Grenze der j-ten Klasse
e_j	obere Grenze der j-ten Klasse
$d_j = e_j - e_{j-1}$	Klassenbreite der j-ten Klasse
$a_j = \frac{1}{2}(e_j + e_{j-1})$	Klassenmitte der j-ten Klasse
n_j	absolute Häufigkeit in der j-ten Klasse
$f_j = \frac{n_j}{n}$	relative Häufigkeit in der j-ten Klasse

Empirische Verteilungsfunktion

Für Häufigkeitstabelle:

$$F(x) = \sum_{a_j \leq x} f(a_j)$$

Lagemaße

Modus

Für Häufigkeitstabelle bzw. bei gruppierten Daten:

$$\bar{x}_M \quad = \quad a_j \quad \Leftrightarrow \quad n_j = \max\{n_1, n_2, \ldots, n_k\}$$

Median

$$\tilde{x}_{0,5} \quad = \quad \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade,} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{falls } n \text{ gerade.} \end{cases}$$

Bei klassierten Daten:

$$\tilde{x}_{0,5} \quad = \quad e_{m-1} + \frac{0,5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m$$

wobei m folgendermaßen bestimmt ist:

$$\sum_{j=1}^{m-1} f_j < 0,5 \quad \text{und} \quad \sum_{j=1}^{m} f_j \geq 0,5$$

Quantile

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \text{ keine ganze Zahl ist,} \\ & k \text{ ist dann die kleinste,} \\ & \text{ganze Zahl} > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{falls } n\alpha \text{ ganzzahlig ist.} \end{cases}$$

Arithmetisches Mittel

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

Bei Häufigkeitstabelle bzw. bei gruppierten Daten:

$$\bar{x} \quad = \quad \frac{1}{n} \sum_{j=1}^{k} n_j a_j = \sum_{j=1}^{k} f_j a_j$$

Falls bekannt werden im klassierten Fall statt der Klassenmitten a_j die Klassenmittelwerte $\bar{x}_j$ verwendet.

Geometrisches Mittel

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

Für Häufigkeitstabelle bzw. bei gruppierten Daten:

$$\bar{x}_G = \sqrt[n]{\prod_{j=1}^{k} a_j^{n_j}} = \left(\prod_{j=1}^{k} a_j^{n_j} \right)^{\frac{1}{n}}$$

Falls bekannt können im klassierten Fall statt der Klassenmitten a_j auch die klassenspezifischen geometrischen Mittel verwendet werden.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} = \sqrt[n]{\frac{B_n}{B_0}}$$

mit

$$B_n = B_0 \cdot x_1 \cdot \ldots \cdot x_n$$

B_0	Anfangsbestand
B_i	Bestand zum Zeitpunkt $i = 1, \ldots, n$
B_n	Endbestand
$x_i = \frac{B_i}{B_{i-1}}$	i-ter Wachstumsfaktor

Harmonisches Mittel

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

Für Häufigkeitstabelle bzw. bei gruppierten Daten:

$$\bar{x}_H = \frac{n}{\sum_{j=1}^{k} \frac{n_j}{a_j}} = \frac{1}{\sum_{j=1}^{k} \frac{f_j}{a_j}}$$

Falls bekannt können im klassierten Fall statt der Klassenmitten a_j auch die klassenspezifischen harmonischen Mittel verwendet werden.

Streuungsmaße

Spannweite

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

Quartilsabstand

$$d_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

(Empirische) Varianz

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

Für Häufigkeitstabelle bzw. bei gruppierten Daten:

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(a_j-\bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k} n_j a_j^2 - \bar{x}^2$$

Streuungszerlegung:

$$s^2 = s^2_{zwischen} + s^2_{innerhalb}$$

mit

$$s^2_{zwischen} = \frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j-\bar{x})^2$$

$$s^2_{innerhalb} = \frac{1}{n}\sum_{j=1}^{k} n_j s_j^2$$

$$s_j^2 = \frac{1}{n_j}\sum_{x_i\in K_j}(x_i-\bar{x}_j)^2$$

$$\bar{x}_j = \frac{1}{n_j}\sum_{x_i\in K_j} x_i$$

(Empirische) Standardabweichung

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

Variationskoeffizient

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

Konzentrationsmaße

Lorenzkurve

$$u_0 = 0, \quad v_0 = 0$$

$$u_i = \frac{i}{n}, \quad i = 1,\ldots,n$$

$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \quad i = 1,\ldots,n$$

Bei gruppierten Daten:

$$\tilde{u}_0 = 0, \quad \tilde{v}_0 = 0$$

$$\tilde{u}_i = \sum_{j=1}^{i} f_j, \quad i = 1,\ldots,k$$

$$\tilde{v}_i = \frac{\sum_{j=1}^{i} f_j a_j}{\sum_{j=1}^{k} f_j a_j} = \frac{\sum_{j=1}^{i} n_j a_j}{n\bar{x}}, \quad i = 1,\ldots,k$$

Falls bekannt werden im gruppierten Fall statt der Klassenmitten a_j die Klassenmittelwerte $\bar{x}_j$ verwendet.

Gini-Koeffizient

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

Bei gruppierten Daten:

$$G = 1 - \frac{1}{n}\sum_{j=1}^{k} n_j(\tilde{v}_{j-1}+\tilde{v}_j)$$

Wertebereich:

$$0 \le G \le \frac{n-1}{n}$$

Normierter Gini-Koeffizient (Lorenz-Münzner-Koeffizient):

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

Wertebereich:

$$0 \le G^+ \le 1$$

Maßzahlen für den Zusammenhang zweier Merkmale

Schema einer $k\times l$ - Kontingenztafel

		Merkmal Y					
		y_1		y_j		y_l	$\sum$
Merkmal X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	$\sum$	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

Bei ordinalen Merkmalen:

$K = \sum_{i<m}\sum_{j<n} n_{ij}n_{mn}$ Anzahl konkordanter Paare
$D = \sum_{i<m}\sum_{j>n} n_{ij}n_{mn}$ Anzahl diskordanter Paare
$T_X = \sum_{i=m}\sum_{j<n} n_{ij}n_{mn}$ Anzahl Bindungen bzgl. X
$T_Y = \sum_{i<m}\sum_{j=n} n_{ij}n_{mn}$ Anzahl Bindungen bzgl. Y

Schema einer 2×2 - Kontingenztafel

		Merkmal Y		
		y_1	y_2	$\sum$
Merkmal X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	$\sum$	$a+c$	$b+d$	n

χ^2-Statistik

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

Wertebereich:

$0 \le \chi^2 \le n(min(k,l)-1)$

Spezialfall: 2×2-Tafeln:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Phi-Koeffizient

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

Wertebereich:

$0 \le \Phi \le \sqrt{min(k,l)-1}$

Spezialfall: 2×2-Tafeln:

$$\Phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Cramers V

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

Wertebereich:

$0 \le V \le 1$

Kontingenzkoeffizient C

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

Wertebereich:

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

Korrigierter Kontingenzkoeffizient C_{korr}

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

Wertebereich:

$0 \le C_{korr} \le 1$

Odds-Ratio

$$OR = \frac{ad}{bc}$$

Wertebereich:

$0 \le OR < \infty$

Gamma nach Goodman und Kruskal

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

Wertebereich:

$-1 \le \gamma \le 1$

Kendalls τ_b

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

Wertebereich:

$-1 \le \tau_b \le 1$

Kendalls/Stuarts τ_c

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

Wertebereich:

$-1 \le \tau_c \le 1$

Rangkorrelationskoeffizient nach Spearman

- Ohne Bindungen:

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

mit

$R(x_i)$	Rang der i-ten Beobachtung von X
$R(y_i)$	Rang der i-ten Beobachtung von Y
$d_i = R(x_i) - R(y_i)$	Rangdifferenz

- Mit Bindungen:

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

mit

J	Anzahl unterschiedl. Merkmalsausprägungen bei X
K	Anzahl unterschiedl. Merkmalsausprägungen bei Y
b_j	abs. Häufigkeit der j-ten Merkmalsausprägung bei X
c_k	abs. Häufigkeit der k-ten Merkmalsausprägung bei Y

Wertebereich:

$$-1 \leq R \leq 1$$

Korrelationskoeffizient nach Bravais-Pearson

$$\begin{aligned} r &= \frac{Covar(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \end{aligned}$$

mit

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

und

$$\begin{aligned} Var(X) &= \frac{1}{n} \cdot S_{xx} \\ Var(Y) &= \frac{1}{n} \cdot S_{yy} \\ Covar(X,Y) &= \frac{1}{n} \cdot S_{xy} \end{aligned}$$

Wertebereich:

$$-1 \leq r \leq 1$$

Lineare Einfachregression

Modell

$$Y = a + bX + e$$

KQ-Schätzer

$$\begin{aligned} \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

Eigenschaften der Regressionsgeraden

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

Streuungszerlegung

$$SQ_{Total} = SQ_{Regression} + SQ_{Residual}$$

$$\begin{aligned} SQ_{Total} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ SQ_{Regression} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ SQ_{Residual} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2 \\ &= S_{yy} - \hat{b}^2 S_{xx} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \end{aligned}$$

Es gilt: $S_{yy} = SQ_{Total}$

Bestimmtheitsmaß

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}} = r^2$$

Wertebereich:

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

Verhältniszahlen und Indizes

Indexzahlen

Einfache Indexzahl (oder Messzahl):

$$I_{0t} = \frac{x_t}{x_0}$$

mit

x_0	Wert der Maßzahl in der Basisperiode
x_t	Wert der Maßzahl in der Berichtsperiode t

Veränderung des Basisjahres:

$$I_{kt} = \frac{x_t}{x_k} = \frac{\frac{x_t}{x_0}}{\frac{x_k}{x_0}} = \frac{I_{0t}}{I_{0k}}$$

Verkettungsregel:

$$I_{0t} = I_{0k} \cdot I_{kt}$$

Definitionen

$$\begin{aligned} \mathbf{p}_0' &= (p_0(1), \ldots, , p_0(n)) \\ \mathbf{p}_t' &= (p_t(1), \ldots, p_t(n)) \\ \mathbf{q}_0' &= (q_0(1), \ldots, q_0(n)) \\ \mathbf{q}_t' &= (q_t(1), \ldots, q_t(n)) \end{aligned}$$

mit

$p_0(i)$	Preis von Gut $i = 1,\ldots,n$ in der Basisperiode
$p_t(i)$	Preis von Gut $i = 1,\ldots,n$ in der Berichtsperiode t
$q_0(i)$	Menge von Gut $i = 1,\ldots,n$ in der Basisperiode
$q_t(i)$	Menge von Gut $i = 1,\ldots,n$ in der Berichtsperiode t

Preisindex nach Laspeyres

$$P_{0t}^L = \frac{\mathbf{p}_t'\mathbf{q}_0}{\mathbf{p}_0'\mathbf{q}_0} = \frac{\sum_{i=1}^n p_t(i)q_0(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

Preisindex nach Paasche

$$P_{0t}^P = \frac{\mathbf{p}_t'\mathbf{q}_t}{\mathbf{p}_0'\mathbf{q}_t} = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_t(i)}$$

Mengenindex nach Laspeyres

$$Q_{0t}^L = \frac{\mathbf{p}_0'\mathbf{q}_t}{\mathbf{p}_0'\mathbf{q}_0} = \frac{\sum_{i=1}^n p_0(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

Mengenindex nach Paasche

$$Q_{0t}^P = \frac{\mathbf{p}_t'\mathbf{q}_t}{\mathbf{p}_t'\mathbf{q}_0} = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_t(i)q_0(i)}$$

Umsatzindex (Wertindex)

$$W_{0t} = \frac{\mathbf{p}_t'\mathbf{q}_t}{\mathbf{p}_0'\mathbf{q}_0} = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

Erweiterung des Warenkorbs

Einführung des neuen Guts (mit Nummer $n+1$) zum Zeitpunkt t':

$$\begin{aligned} P_{t',t'+1}^L(\text{erweitert}) &= \frac{\mathbf{p}_{t'+1}'\mathbf{q}_0 + p_{t'+1}(n+1)q_{t'}(n+1)}{\mathbf{p}_{t'}'\mathbf{q}_0 + p_{t'}(n+1)q_{t'}(n+1)} \\ P_{0,t'+1}^L(\text{verkettet}) &= P_{0,t'}^L \cdot P_{t',t'+1}^L(\text{erweitert}) \end{aligned}$$

Subindizes

$$P_{0t}^L = w^I P_{0t}^L(I) + w^{II} P_{0t}^L(II)$$

mit

$$\begin{aligned} P_{0t}^L(I) &= \frac{\sum_{i=1}^m p_t(i)q_0(i)}{\sum_{i=1}^m p_0(i)q_0(i)} \\ P_{0t}^L(II) &= \frac{\sum_{i=m+1}^n p_t(i)q_0(i)}{\sum_{i=m+1}^n p_0(i)q_0(i)} \\ U &= \sum_{i=1}^n p_0(i)q_0(i) \\ w^I &= \frac{\sum_{i=1}^m p_0(i)q_0(i)}{U} \\ w^{II} &= 1 - w^I = \frac{\sum_{i=m+1}^n p_0(i)q_0(i)}{U} \end{aligned}$$

Zeitreihen

Additives Komponentenmodell

$$y_t = g_t + s_t + r_t, \quad t = 1,\ldots,T$$

mit

y_t	Beobachtung zum Zeitpunkt t
g_t	glatte Komponente zum Zeitpunkt t
s_t	saisonale Komponente zum Zeitpunkt t
r_t	Restkomponte zum Zeitpunkt t

Gleitende Durchschnitte

ungerader Ordnung $2k+1$:

$$y_t^* = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j}$$

gerader Ordnung $2k$:

$$y_t^* = \frac{1}{2k}\left(\frac{1}{2}y_{t-k} + \sum_{j=-k+1}^{k-1} y_{t+j} + \frac{1}{2}y_{t+k}\right)$$