deskriptive Statistik

Silke Bott

Sommersemester 2023

Mehrdimensionale deskriptive Statistik

In vielen Anwendungen interessiert man sich nicht nur für ein einziges Merkmal sondern gleichzeitig für mehrere.
Bei der Untersuchen von zwei Merkmalen X und Y interessieren dabei grundsätzlich zwei Fragen:

- Beeinflussen sich die Merkmalsausprägungen von X und Y gegenseitig (Korrelationsanalyse)?
- Hängen die Ausprägungen eines Merkmals (etwa Y) funktional von den Ausprägungen des anderen Merkmals (also X) ab (Regressionsanalyse).

Mehrdimensionale deskriptive Statistik

Sonntagsumfrage (infratest/dimap 14.9.2017 bei 1200 Befragten)

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Sonst
	444	240	108	90	114	144	60

Aufgegliedert nach Ostdeutschland und Westdeutschland:

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Sonst
WD	359	197	55	76	92	77	44
OD	85	43	53	14	22	67	16

Mehrdimensionale deskriptive Statistik

Beispiel

Eine Befragung zu *Altersstufe* („jung“, „mittleres Alter“ und „älter“) und *Einkommensstufe* („niedrig“, „mittel“ und „hoch“) bei 100 Erwerbstätigen hat folgende Zahlen ergeben

	jg	ma	al
n	20	15	7
m	8	17	14
h	2	8	9

Mehrdimensionale deskriptive Statistik

Wir betrachten nun allgemeiner zwei Merkmale X und Y mit den möglichen Ausprägungen $a_1, \ldots, a_k$ für X und $b_1, \ldots, b_l$ für Y.

Definition

Die Werte $h_{i,j}, i = 1, \ldots, k, j = 1, \ldots, l$ heißen *gemeinsame Verteilung der Merkmale X und Y in absoluten Häufigkeiten* und die Tabelle

$$\begin{array}{c|ccc}
 & b_1 & \ldots & b_l \\
\hline
a_1 & h_{1,1} & \ldots & h_{1,l} \\
\vdots & \vdots & \ddots & \vdots \\
a_k & h_{k,1} & \ldots & h_{k,l} \\
\hline
\end{array}$$

heißt *Kontingenztabelle* der Merkmale X und Y.

Mehrdimensionale deskriptive Statistik

Definition (Randhäufigkeiten)

Die *Randhäufigkeiten des Merkmals Y* sind die Spaltensummen der Kontingenztabelle

$$h_{\bullet,j} = h_{1,j} + \cdots + h_{k,j} \qquad (j = 1, \ldots, l)$$

und die *Randhäufigkeiten des Merkmals X* sind die Zeilensummen der Kontingenztabelle

$$h_{i,\bullet} = h_{i,1} + \cdots + h_{i,l} \qquad (i = 1, \ldots, k)$$

Mehrdimensionale deskriptive Statistik

Die Kontingenztabelle wird oft mit den Randhäufigkeiten ergänzt und hat dann die folgende Gestalt

$$
\begin{array}{c|ccc|c}
 & b_1 & \dots & b_l & \\
\hline
a_1 & h_{1,1} & \dots & h_{1,l} & h_{1,\bullet} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
a_k & h_{k,1} & \dots & h_{k,l} & h_{k,\bullet} \\
\hline
 & h_{\bullet,1} & \dots & h_{\bullet,l} &
\end{array}
$$

Mehrdimensionale deskriptive Statistik

In unserem Beispiel mit der Sonntagsfrage etwa

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Sonst	
WD	359	197	55	76	92	77	44	900
OD	85	43	53	14	22	67	16	300
	444	240	108	90	114	144	60	1200

Mehrdimensionale deskriptive Statistik

In unserem Beispiel mit der Einkommensverteilung ergibt sich

	jg	ma	al	
n	20	15	7	42
m	8	17	14	39
h	2	8	9	19
	30	40	30	100

Mehrdimensionale deskriptive Statistik

Übung

Die Aufgliederung des Beschäftigungsverhältnisses nach Geschlecht ergibt folgende Daten (in Tausend)

	arbeitslos	Teilzeit	vollbeschäftigt
weiblich	500	4200	7300
männlich	500	800	12700

Bestimmen Sie die Randhäufigkeiten.

Mehrdimensionale deskriptive Statistik

Lösung:

Die um die Randhäufigkeiten erweiterte Datentabelle ist

	arbeitslos	Teilzeit	vollbeschäftigt	
weiblich	500	4200	7300	12 000
männlich	500	800	12700	14 000
	1000	5000	20 000	26 000

Mehrdimensionale deskriptive Statistik

Definition (relative Häufigkeiten)

Bei einer Stichprobengröße n heißt

$$f_{i,j} = f(a_i, b_j) = \frac{1}{n} \cdot h(a_i, b_j)$$

die *relative (gemeinsame) Häufigkeit* der Merkmalsausprägungen (a_i, b_j), und

$$f(a_i) = f_X(a_i) = \frac{1}{n} \cdot h_{i,\bullet}, \qquad f(b_j) = f_Y(b_j) = \frac{1}{n} \cdot h_{\bullet,j}$$

heißen relative Randhäufigkeiten der Merkmalsausprägungen a_i von X und b_j von Y.

Mehrdimensionale deskriptive Statistik

Bei der Sonntagsumfrage

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Sonst	
WD	0.299	0.164	0.046	0.063	0.077	0.064	0.037	0.75
OD	0.071	0.036	0.044	0.012	0.018	0.056	0.013	0.25
	0.370	0.200	0.090	0.075	0.095	0.12	0.050	1

Mehrdimensionale deskriptive Statistik

Beispiel

Beim Einkommensbeispiel erhalten wir

	jg	ma	al
n	0.20	0.15	0.07
m	0.08	0.17	0.14
h	0.02	0.08	0.09

Mehrdimensionale deskriptive Statistik

Übung

Bestimmen Sie die Tabelle der relativen Häufigkeiten für das Beschäftigungsbeispiel mit

	arbeitslos	Teilzeit	vollbeschäftigt	
weiblich	500	4200	7300	12 000
männlich	500	800	12700	14 000
	1000	5000	20 000	26 000

Mehrdimensionale deskriptive Statistik

Lösung:

Die Tabelle der relativen Häufigkeiten für das Beschäftigungsbeispiel ist

	arbeitslos	Teilzeit	vollbeschäftigt	
weiblich	0.0192	0.1615	0.2808	0.4615
männlich	0.0192	0.0308	0.4885	0.5385
	0.0385	0.1923	0.7692	1

Mehrdimensionale deskriptive Statistik

Definition (bedingete Häufigkeiten)

Die *bedingten Häufigkeiten von Y unter der Bedingung $X = a_i$*, kurz auch $Y|X = a_i$ sind gegeben durch

$$f_Y(b_1|a_i) = \frac{h_{i,1}}{h_{i,\bullet}}, \quad \ldots \quad , f_Y(b_l|a_i) = \frac{h_{i,l}}{h_{i,\bullet}}$$

Die *bedingten Häufigkeiten von X unter der Bedingung $Y = b_j$*, kurz auch $Y|X = a_i$ sind gegeben durch $f_X(a_1|b_j) = \frac{h_{1,j}}{h_{\bullet,j}}, \ldots, f_X(a_k|b_j) = \frac{h_{k,j}}{h_{\bullet,j}}$.
Die Merkmale X und Y heiße *empirisch unabhängig*, wenn gilt

$$\begin{aligned} f_Y(b_j|a_i) &= \frac{1}{n} \cdot h_{\bullet,j} && \text{für alle } j, i \\ f_X(a_i|b_j) &= \frac{1}{n} \cdot h_{i,\bullet} && \text{für alle } j, i \end{aligned}$$

Mehrdimensionale deskriptive Statistik

Bemerkung

Sind X und Y empirisch unabhängig, so gilt

$$h_{i,j} = \frac{h_{i,\bullet} \cdot h_{\bullet,j}}{n}$$

Deshalb nennen wir auch

$$\widetilde{h_{i,j}} = \frac{h_{i,\bullet} \cdot h_{\bullet,j}}{n}$$

die *Häufigkeiten, die zu erwarten sind, wenn kein Zusammenhang vorliegt*.

Mehrdimensionale deskriptive Statistik

Bei der Sonntagsumfrage wären die zu erwartenden Häufigkeiten bei empirischer Unabhängigkeit

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Sonst	
WD	333	180	81	67.5	85.5	108	45	900
OD	111	60	27	22.5	28.5	36	15	300
	444	240	108	90	114	144	60	1200

Die Merkmale sind also offensichtlich nicht empirisch unabhängig.

Mehrdimensionale deskriptive Statistik

Im Einkommensbeispiel mit den Daten

	jg	ma	al	
n	20	15	7	42
m	8	17	14	39
h	2	8	9	19
	30	40	30	100

erhalten wir als Tabelle der erwarteten Häufigkeiten bei empirischer Unabhängigkeit

	jg	ma	al	
n	12.6	16.8	12.6	42
m	11.7	15.6	11.7	39
h	5.7	7.6	5.7	19
	30	40	30	100

Auch hier sind die Merkmale offensichtlich nicht empirisch unabhängig.

Mehrdimensionale deskriptive Statistik

Übung

Bestimmen Sie die erwarteten Häufigkeiten im Beschäftigungsbeispiel mit

	arbeitslos	Teilzeit	vollbeschäftigt	
weiblich	500	4200	7300	12 000
männlich	500	800	12700	14 000
	1000	5000	20 000	26 000

Mehrdimensionale deskriptive Statistik

Lösung:

Die Tabelle der erwarteten Häufigkeiten bei empirischer Unabhängigkeit ist

	arbeitslos	Teilzeit	vollbeschäftigt
weiblich	461.54	2307.69	9230.77
männlich	538.46	2692.31	10 769.23

Auch hier sind die Daten offensichtlich nicht empirisch unabhängig (was man vor allem an der Spalte „Teilzeit“ sieht).

Kontingenzkoeffizienten

Definition (Kontingenzkoeffizienten)

Die Größe

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(h_{i,j} - \widetilde{h_{i,j}}\right)^2}{\widetilde{h_{i,j}}}$$

heißt χ^2*–Koeffizient von* X *und* Y.
Die Größe

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

heißt *Kontingenzkoeffizient von* X *und* Y.
Die Größe

$$K^* = \frac{K}{\sqrt{\frac{M-1}{M}}} \qquad \text{wobei } M = \min\{k, l\}$$

heißt *korrigierter Kontingenzkoeffizient von* X *und* Y.

Kontingenzkoeffizienten

Bemerkung

Es gilt

1. Die Kontingenzkoeffizienten bilden ein Maß dafür, ob die Merkmalsausprägungen für X mit denen für Y zusammenhängen.
2. Die Kontingenzkoeffizienten sind geeignet für die Untersuchung mindestens nominalskalierter Merkmale.
3. $\chi^2 \in [0, \infty[$, $K \in [0, \sqrt{\frac{M-1}{M}}]$ und $K^* \in [0, 1]$.
4. χ^2 hängt ab von der Skalierung der Merkmale (werden alle Häufigkeiten verdoppelt, so verdoppelt sich auch χ^2), K und K^* sind unabhängig davon. Ferner ist K^* normiert.
5. Genau dann sind die Kontingenzkoeffizienten 0, wenn X und Y empirisch unabhängig sind.
6. Je näher K bzw. K^* an $\sqrt{\frac{M-1}{M}}$ bzw. 1 liegen, desto stärker ist der Zusammenhang zwischen X und Y.

Kontingenzkoeffizienten

Bei der Sonntagsumfrage gilt

$$\begin{aligned}\chi^2 &= 89.865 \\ K &= 0.264 \\ K^* &= 0.373\end{aligned}$$

Auch diese Zahlen zeigen, dass die die Merkmale nicht empirisch unabhängig sind.

Kontingenzkoeffizienten

Übung

Berechnen Sie den χ^2–Koeffizienten, den Kontingenzkoeffizienten und den korrigierten Kontingenzkoeffizienten der Daten des Einkommensbeispiels:

H	jg	ma	al	
n	20	15	7	42
m	8	17	14	39
h	2	8	9	19
	30	40	30	100

$\widetilde{H}$	jg	ma	al	
n	12.6	16.8	12.6	42
m	11.7	15.6	11.7	39
h	5.7	7.6	5.7	19
	30	40	30	100

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(h_{i,j} - \widetilde{h_{i,j}}\right)^2}{\widetilde{h_{i,j}}}, \quad K = \sqrt{\frac{\chi^2}{n+\chi^2}}, \quad K^* = \frac{K}{\sqrt{\frac{M-1}{M}}}$$

Kontingenzkoeffizienten

Lösung:

Es gilt

$$\begin{aligned} \chi^2 &= 13.1 \\ K &= 0.34 \\ K^* &= 0.42 \end{aligned}$$

Auch hier zeigen die Zahlen, dass die Merkmale nicht empirisch unabhängig sind.

Kontingenzkoeffizienten

Beispiel

Im Beispiel mit den Beschäftigungsverhältnissen gilt

$$\begin{aligned} \chi^2 &= 3637.7 \\ K &= 0.350 \\ K^* &= 0.496 \end{aligned}$$

Auch hier zeigen die Zahlen, dass die Merkmale nicht empirisch unabhängig sind.

Korrelationskoeffizienten

Wir betrachten jetzt wieder metrische skalierte Merkmale X und Y mit Merkmalsausprägungen $x_1, \ldots, x_n$ und $y_1, \ldots y_n$.

Definition (Korrelationskoeffizienten)

Die Größe

$$r = r_{X,Y} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}}$$

heißt *Bravais–Pearson–Korrelationskoeffizient* oder *empirischer Korrelationskoeffizient* von X und Y.

Korrelationskoeffizienten

Bemerkung

Die Größe

$$s^2_{X,Y} = \frac{1}{n-1} \cdot \sum (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

wird als *korrigierte empirische Kovarianz* der Stichprobe bezeichnet. Damit gilt

$$r_{X,Y} = \frac{s^2_{X,Y}}{s_X \cdot s_Y}$$

Genau dann ist $s^2_{X,Y} = 0$, wenn $r_{X,Y} = 0$.

Korrelationskoeffizienten

Bemerkung

Es gilt

1. $r_{X,Y} \in [-1, 1]$.
2. $r_{X,Y}$ untersucht einen linearen Zusammenhang zwischen X und Y: Genau dann gilt $r_{X,Y} = \pm 1$, wenn $y_i = ax_i + b$ für alle $i \in \{1, \dots, n\}$, wobei $a, b \in \mathbb{R}, a \neq 0$ fest sind. Dabei ist $r_{X,Y} = 1$ wenn $a > 0$ und $r_{X,Y} = -1$ wenn $a < 0$.
3. Ist $r_{X,Y} > 0$ so heiße X und Y positiv oder gleichsinnig linear korreliert, ist $r_{X,Y} < 0$, so sind sie negativ oder gegensinning linear korreliert. Ist $r_{X,Y} = 0$, so sind sie unkorreliert.

Korrelationskoeffizienten

Beispiel

Betrachte $X = \textit{Größe}$ und $Y = \textit{Gewicht}$ und folgende Daten

Proband	1	2	3	4
Größe	175	184	179	182
Gewicht	82	81	96	81

Es gilt

$$\overline{x} = 180, \quad \overline{y} = 85$$

und damit

$$r_{x,y} = \frac{-20}{\sqrt{46 \cdot 162}} = -\frac{20}{\sqrt{7452}} = -0.232$$

Damit sind X und Y hier leicht negativ korreliert.

Korrelationskoeffizienten

Übung

Bestimmen Sie $r_{X,Y}$ für $X = \textit{Größe}$ und $Y = \textit{Gewicht}$ und folgende Daten

Proband	1	2	3	4	5
Größe	170	175	180	185	190
Gewicht	68	76	74	84	88

Zur Erinnerung:

$$r = r_{X,Y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Korrelationskoeffizienten

Übung

Bestimmen Sie $r_{X,Y}$ für $X = \textit{Größe}$ und $Y = \textit{Gewicht}$ und folgende Daten

Proband	1	2	3	4	5
Größe	170	175	180	185	190
Gewicht	68	76	74	84	88

Lösung:

Es gilt

$$\overline{x} = 180, \quad \overline{y} = 78$$

und

$$r_{x,y} = \frac{240}{\sqrt{250 \cdot 256}} = \frac{240}{\sqrt{64\,000}} = 0.948$$

Damit sind X und Y hier stark positiv korreliert.

Korrelationskoeffizienten

Ein alternativer Korrelationskoeffizient für mindestens ordinal skalierte Merkmale ist möglich, wenn wir von den ursprünglichen Werten von X und Y zu ihren *Rängen* übergehen.

Definition

Die Größe

$$r_{SP} = \frac{\sum_{i=1}^{n} (rg(x_i) - \overline{rg}_X) \cdot \left(rg(y_i) - \overline{rg}_y\right)}{\sqrt{\sum_{i=1}^{n} (rg(x_i) - \overline{rg}_X)^2 \cdot \sum_{i=1}^{n} (rg(y_i) - \overline{rg}_Y)^2}}$$

heißt *Spearman–Korrelationskoeffizient*.

Korrelationskoeffizienten

Bemerkung

Es gilt

1. $r_{SP} \in [-1, 1]$.
2. r_{SP} ist der Bravais–Pearson–Korrelationskoeffizient der Rangausprägungen $(rg(x_i), rg(y_i)$ der Merkmale X und Y.
3. Genau dann ist $r_{SP} > 0$, wenn es einen gleichsinnigen montonen Zusammenhang zwischen X und Y gibt.
4. Genau dann ist $r_{SP} < 0$, wenn es einen gegensinnigen montonen Zusammenhang zwischen X und Y gibt.
5. $r_{SP} \approx 0$ wenn es keinen monotonen Zusammenhang zwischen X und Y gibt.

Korrelationskoeffizienten

Beispiel

Für $X =$ *Größe* und $Y =$ *Gewicht* aus dem ersten Beispiel gilt

Proband	1	2	3	4
rg_X	4	1	3	2
rg_Y	2	3.5	1	3.5

Es gilt

$$\overline{rg_X} = 2.5, \quad \overline{rg_Y} = 2.5$$

und damit

$$r_{SP} = \frac{-3.5}{\sqrt{5 \cdot 4.5}} = -\frac{20}{\sqrt{22.5}} = -0.738$$

Damit sind X und Y relativ stark negativ rangkorreliert.

Korrelationskoeffizienten

Übung

Bestimmen Sie r_{sp} für $X = \textit{Größe}$ und $Y = \textit{Gewicht}$ und folgende Daten

Proband	1	2	3	4	5
Größe	170	175	180	185	190
Gewicht	68	76	74	84	88

Korrelationskoeffizienten

Lösung:

Wir ermitteln zunächst die Rangtabelle

Proband	1	2	3	4	5
rg_X	5	4	3	2	1
rg_Y	5	3	4	2	1

Es gilt

$$\overline{rg_X} = 3, \quad \overline{rg_Y} = 3$$

und

$$r_{x,y} = \frac{9}{\sqrt{10 \cdot 10}} = \frac{9}{10} = 0.90$$

Damit sind X und Y hier stark positiv rangkorreliert.

lineare Regression

Lineare Regression für Beobachtungen $(x_1, y_1), \ldots, (x_n, y_n)$:

Finde Gerade $y = f(x) = a \cdot x + b$, so dass $\sum_{i=1}^{n} (y_i - f(x_i))^2$ minimal wird.

Setze

$$\widehat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad \widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x}$$

lineare Regression

Definition (Lineare Regression)

Die Gerade

$$f(x) = \widehat{a} \cdot x + \widehat{b}$$

heißt *lineare Einfachregression* der Merkmale X und Y.
Die Werte

$$\varepsilon_i = y_i - f(x_i)$$

heißen *Residuen* der Einfachregression.

lineare Regression

Definition

Die Größe

$$R^2 = \frac{\sum_{i=1}^{n} (f(x_i) - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

heißt Bestimmtheitsmaß der linearen Regression.

Bemerkung

$$R^2 = r_{X,Y}^2$$

Bemerkung

$$\sum_{i=1}^{n} \left(y_i - \widehat{a} \cdot x_i - \widehat{b}\right)^2 = \min \left\{ \sum_{i=1}^{n} (y_i - a \cdot x_i - b)^2 \mid a, b \in \mathbb{R} \right\}$$

lineare Regression

Beispiel

Wir betrachten zwei Merkmale X und Y mit den folgenden Ausprägungen

k	1	2	3	4	5	6	7
x_k	3	5	6	7	8	10	12
y_k	24	30	31	30	28	44	42

Es gilt

$$\overline{x} = 7.2875, \qquad \overline{y} = 32.7142$$

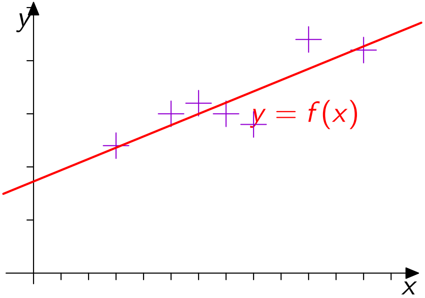
und damit

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 2.1211 \quad \text{und} \quad \widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x} = 17.2603$$

lineare Regression

Es ist $R^2 = 0.757$ (also eine recht ordentliche Erklärung), und die lineare Regressionsgerade ist gegeben durch

$$f(x) = 17.2603 + 2.1211 \cdot x$$



lineare Regression

Übung

Bestimmen Sie die Regressionsgerade für $X = \textit{Größe}$, $Y = \textit{Gewicht}$ und

Proband	1	2	3	4
Größe	160	170	180	190
Gewicht	57	72	81	86

Lösung:

Zunächst ist

$$\overline{x} = 175, \qquad \overline{y} = 74$$

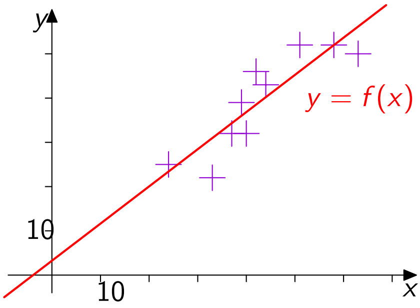
und damit

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.96 \quad \text{und} \quad \widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x} = -94$$

lineare Regression

Es ist und $R^2 = 0.95$, also eine sehr gute Erklärung, und die lineare Regressionsgerade ist gegeben durch

$$f(x) = -94 + 0.96 \cdot x$$



lineare Regression

An der großen Skala ist schwer zu erkennen, wie gut die Gerade, die Punkte erklärt, aber auch nähere Betrachtung zeigt ein gutes Bild: