Kapitel 3

Beschreibende Statistik mehrdimensionaler Daten

3.1 Gemeinsame Häufigkeiten und Randhäufigkeiten

Aufgabenstellung:

Es geht um die Analyse der gemeinsamen Verteilung von zwei oder mehreren Größen. Wir beschränken uns hier auf den zweidimensionalen Fall und die zugehörigen Methoden der beschreibenden Statistik. Eine solche Situation tritt etwa in der Sozialstatistik auf, wenn jedes Individuum hinsichtlich zweier Merkmale (X, Y) untersucht wird, oder im Laborversuch, in dem zwei Größen (X, Y) gemessen werden. Ziel ist die Analyse der Abhängigkeit der beiden Größen.

Bsp. 3.1 X = Note in Mathematik, Y = Note in Englisch oder X = Bewölkungsgrad, Y = Tagestemperatur.

Die Ausprägungen der Größe X seien $x_1, \ldots, x_k$, die von Y seien $y_1, \ldots, y_l$. Ein möglicher Messwert (x_i, y_j) stellt also einen Punkt in der (x, y)-Ebene dar.

Bivariater Datensatz: Dies ist die Menge (Gitter, Raster)

$$\{(x_i, y_j) : i = 1, \ldots, k, j = 1, \ldots, l\}.$$

Gemeinsame absolute und relative Häufigkeiten:

$$H_{ij} = H(X = x_i, Y = y_j), \quad h_{ij} = \frac{1}{n} H_{ij}$$

bei Stichprobenumfang n oder bei Klasseneinteilung in X- bzw. Y-Klassen auch

$$H_{ij} = H(X \in X\text{−Klasse } i,\ Y \in Y\text{−Klasse } j).$$

Bsp. 3.2 (Noten) $n = 30$, X = Note Mathematik, Y = Note Englisch.

$X \setminus Y$	1	2	3	4	5	
1	2				1	3
2	4	2				6
3	2	5	2	1		10
4	4	4	1			9
5		1	1			2
	12	12	4	1	1	30

Dabei werden in der i-ten Zeile, j-ten Spalte die

$$H_{ij} = H(\text{Mathematik} = i,\ \text{Englisch} = j) = H(X = x_i, Y = y_j)$$

eingetragen.

Randverteilung: Die Randspalte bzw. Randzeile stellt die Randverteilung dar:

i-te Zeile: $\quad H_{i\star} = H(X = x_i, Y \text{ beliebig}) = \sum_{j=1}^{l} H_{ij}$ (Zeilensumme).

j-te Spalte: $\quad H_{\star j} = H(X \text{ beliebig}, Y = y_j) = \sum_{i=1}^{k} H_{ij}$ (Spaltensumme).

Die Randverteilungen geben das Verhalten der Messgrößen X und Y einzeln wieder.

Bsp. 3.3 (Noten) Gegenüberstellung der Verteilungen der Mathematik-/Englischnoten im Boxplot, siehe Abb. 3.1.

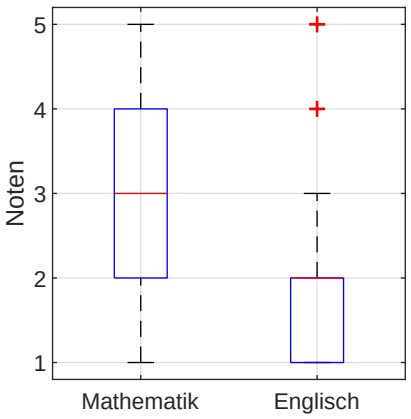

Abbildung 3.1: Boxplots für Mathematik- und Englischnoten.

Berechnung der Parameter der Größen X und Y:

$$\begin{aligned} \overline{x} = \tfrac{1}{n}\sum_{i=1}^{k} x_i H_{i\star} &= 3.0333, & \overline{y} = \tfrac{1}{n}\sum_{j=1}^{l} y_j H_{\star j} &= 1.9000, \\ s_x^2 = \tfrac{1}{n-1}\sum_{i=1}^{k}(x_i-\overline{x})^2 H_{i\star} &= 1.2057, & s_y^2 = \tfrac{1}{n-1}\sum_{j=1}^{l}(y_j-\overline{y})^2 H_{\star j} &= 0.9897. \end{aligned}$$

Allgemeine zweidimensionale Häufigkeitstabelle:

$X \setminus Y$	y_j			
	$\vdots$			
x_i	$\dots$	H_{ij}	$\dots$	$H_{i\star}$
	$\vdots$			
	$H_{\star j}$			n

$\sum_j$ (rechts neben $H_{i\star}$), $\sum_i$ (unter $H_{\star j}$)

Relative Häufigkeiten: Es gilt

$$h_{ij} = \frac{1}{n}H_{ij}, \quad h_{i\star} = \frac{1}{n}H_{i\star}, \quad h_{\star j} = \frac{1}{n}H_{\star j}.$$

Was kann noch über die gemeinsame Verteilung gesagt werden?
Es interessieren uns die wechselseitigen Abhängigkeiten.

Zunächst der Extremfall der **Unabhängigkeit**. Dies sollte bedeuten, dass – ganz gleich, was der Wert für X ist – die Verteilung von Y stets dieselbe ist, und umgekehrt. Daraus folgt aber, dass die Zeilen und Spalten in der Häufigkeitsmatrix proportional sein müssen.

Bsp. 3.4 Ein Beispiel zweier unabhängiger Merkmale:

$X \setminus Y$	1	2	3	4	5	
1	1	2			1	4
2	3	6			3	12
3	2	4			2	8
4						0
5	1	2			1	4
	7	14	0	0	7	28

Insbesondere muss jede Zeile ein Vielfaches c der Randverteilungszeile sein. Es gilt dann etwa für die erste Zeile:

$$H_{1j} = c \cdot H_{\star j}, \quad j = 1, \dots, l \qquad (\text{hier } c = 1/7).$$

Summation über j ergibt

$$H_{1\star} = \sum_{j=1}^{l} H_{1j} = \sum_{j=1}^{l} c \cdot H_{\star j} = c \cdot n, \quad \text{bzw.} \quad c = \frac{1}{n} \cdot H_{1\star}.$$

Daraus folgt

$$H_{1j} = c \cdot H_{\star j} = \frac{1}{n} \cdot H_{1\star} \cdot H_{\star j}, \quad j = 1, \ldots, l$$

bzw. nach Division durch n

$$h_{1j} = h_{1\star} \cdot h_{\star j}, \quad j = 1, \ldots, l.$$

Das Entsprechende gilt für alle Zeilen. Damit ist geklärt, dass für die Unabhängigkeit notwendig und hinreichend ist, dass die relativen Häufigkeiten h_{ij} die Produkte der jeweiligen relativen Randhäufigkeiten sind:

$$h_{ij} = h_{i\star} \cdot h_{\star j}, \quad i = 1, \ldots, k, \; j = 1, \ldots, l.$$

Es ist klar, dass die Unabhängigkeit eine sehr starke Eigenschaft ist (bei Schulnoten wohl kaum jemals zutreffend, wie das Notenbeispiel zeigt). Es werden daher abgestuftere Begriffe benötigt.

3.2 Kovarianz und Korrelation

Betrachten wir die Produkte

$$(x_i - \overline{x}) \cdot (y_j - \overline{y})$$

der Abweichungen zu den jeweiligen Mittelwerten. Diese Abweichungen sind positiv, wenn gleichzeitig $x_i > \overline{x},\ y_j > \overline{y}$ oder $x_i < \overline{x},\ y_j < \overline{y}$ ist. Sie sind negativ, wenn $x_i > \overline{x},\ y_j < \overline{y}$ oder $x_i < \overline{x},\ y_j > \overline{y}$ ist.

Bsp. 3.5 (Noten) Das Ergebnis dieser Produkte für das Notenbeispiel aus Bsp. 3.2 steht in der folgenden Tabelle. In den Klammern sind die Häufigkeiten $\neq 0$ angegeben.

	y_i	1	2	3	4	5
	$y_j - \overline{y}$	-0.9000	0.1000	1.1000	2.1000	3.1000
x_i	$x_i - \overline{x}$					
1	−2.0333	1.8300 (2)	−0.2033	−2.2367	−4.2700	−6.3033 (1)
2	−1.0333	0.9300 (4)	−0.1033 (2)	−1.1367	−2.1700	−3.2033
3	−0.0333	0.0300 (2)	−0.0033 (5)	−0.0367 (2)	−0.0700 (1)	−0.1033
4	0.9667	−0.8700 (4)	0.0967 (4)	1.0633 (1)	2.0300	2.9967
5	1.9667	−1.7700	0.1967 (1)	2.1633 (1)	4.1300	6.0967

Im Falle der "Unkorreliertheit" sollten sich diese mit den Häufigkeiten (H_{ij}) multiplizierten Produkte in Summe weg heben. Ist dies nicht der Fall, so treten also tendenziell Abweichungen von den Mittelwerten in derselben Richtung auf.

Empirische Kovarianz: Sie ist definiert als

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \overline{x})(y_j - \overline{y})\, H_{ij}.$$

Wenn $s_{xy} = 0$ ist, so heißen X und Y **unkorreliert**, andernfalls **korreliert**.

Bsp. 3.6 Im Notenbeispiel ist

$$s_{xy} = 0.0379,$$

d.h. die Leistungen in Englisch und Mathematik sind korreliert.

In MATLAB erhalten wir mit `cov(x,y)` die **Kovarianzmatrix**

$$\begin{bmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{bmatrix} = \begin{bmatrix} 1.2057 & 0.0379 \\ 0.0379 & 0.9897 \end{bmatrix},$$

wobei für die Diagonalelemente $s_{xx} = s_x^2$ und $s_{yy} = s_y^2$ gilt. Dies sind gerade die Varianzen der Randverteilungen.

Pearson-Korrelationskoeffizient: Um ein Maß für die Korrelation zu erhalten, das die Größenordnung der gesamten Schwankungsbreite herausskaliert, führen wir den Pearson-Korrelationskoeffizient R ein:

$$R = r_{xy} = \frac{s_{xy}}{s_x\, s_y}.$$

Die Extremfälle $R = 1$ bzw. $R = -1$ erhält man, wenn Y eine lineare Funktion von X mit positivem bzw. negativem Anstieg ist.

Bsp. 3.7 Im Notenbeispiel ist $R = 0.0347$, es liegt also eine (sehr schwache) positive Korrelation vor. Eher deutet dies darauf hin, dass in der betreffenden Schulklasse die Mathematik-/Englischnoten fast unkorreliert sind.

In MATLAB erhalten wir mit `corrcoef(x,y)` die symmetrische Matrix

$$\begin{bmatrix} r_{xx} & r_{xy} \\ r_{yx} & r_{yy} \end{bmatrix} = \begin{bmatrix} 1 & 0.0347 \\ 0.0347 & 1 \end{bmatrix},$$

wobei natürlich immer $r_{xx} = r_{yy} = 1$ gilt. In höheren Dimensionen erhält man z.B. für drei Spaltenvektoren `x`, `y` und `z` mit `corrcoeff([x,y,z])` die symmetrische Matrix

$$\begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix}.$$

Den einzelen Wert r_{xy} erhält man direkt mit `corr(x,y)`.

Bemerkung: Unabhängige Größen sind unkorreliert!

Wir werden dies im Kapitel 5 beweisen. Die Umkehrung gilt jedoch nicht.

Bsp. 3.8 Ein Beispiel eines bivariaten Datensatzes, der unkorreliert, aber nicht unabhängig ist:

$X \setminus Y$	1	2	3	4	5	
1		2			1	3
2						0
3	4	2	9	4	3	22
4		4			2	6
5						0
	4	8	9	4	6	31

Bsp. 3.9 Wir betrachten den Bewölkungsgrad X (in %) und die mittlere Tagestemperatur Y (in °C) im Jänner 2002 in Innsbruck (Tabelle 2.1 oder 3.1). Die Abbildung 3.2 zeigt das so genannte Streudiagramm (Scatterplot) der beiden Messgrößen.

Ein Streudiagramm erhält man mit MATLAB für Daten in `x` und `y` mit `plot(x,y,'o')` oder `scatter(x,y)`.

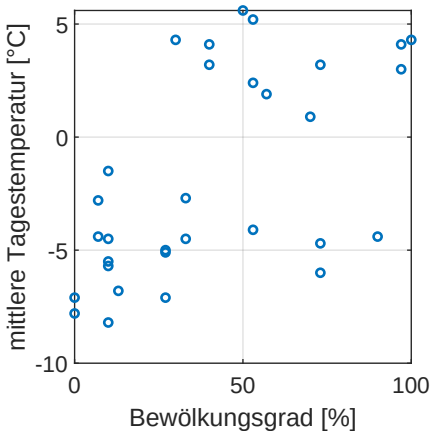

Abbildung 3.2: Streudiagramm Bewölkungsgrad/Temperatur, Innsbruck, Jänner 2002.

Offensichtlich gibt es eine Tendenz zu höheren Temperaturen bei höherem Bewölkungsgrad. Um diese Tendenz zu untermauern bzw. zu quantifizieren, berechnen wir den Pearson-Korrelationskoeffizienten R. Seinen Wert erhalten wir aus

$$s_x = 31.0579,\ s_y = 4.5788,\ s_{xy} = 80.8498,\ R = r_{xy} = \frac{s_{xy}}{s_x\, s_y} = 0.5685.$$

und belegt eine deutliche Korrelation im positiven Sinn.

Bsp. 3.10 Altersabhängige Unfallshäufigkeiten. Die Tabelle unten enthält die im Innsbrucker Straßenverkehr im Jahre 1992 Verunfallten nach Altersgruppe (Quelle: Statistisches Jahrbuch der Stadt Innsbruck 1992).

Altersklasse	Klassenmitte	Verunfallte	Gesamtzahl	Anteil
15 – 25	20	442	19296	0.02291
25 – 35	30	367	21492	0.01708
35 – 45	40	166	15671	0.01059
45 – 55	50	203	15621	0.01299
55 – 65	60	81	10557	0.00767
65 –	75	113	19083	0.00592

Um einen möglichen Zusammenhang feststellen zu können, vergleichen wir die beiden Größen X = Gruppenmitte und Y = Anteil, dazu das Streubild in Abb. 3.3.

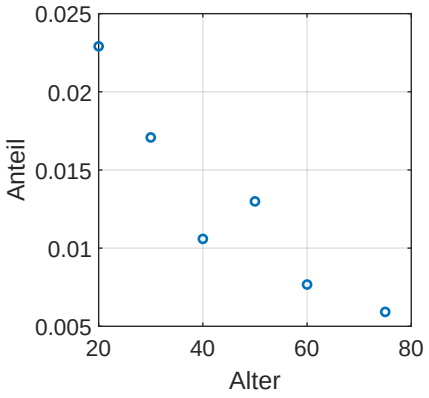

Abbildung 3.3: Streubild Altersgruppe/Unfallhäufigkeit.

Die statistischen Kennwerte sind:

$$s_x = 20.1039,\ s_y = 0.0063,\ s_{xy} = -0.1177,\ R = r_{xy} = -0.9280.$$

Es liegt also tatsächlich eine stark negative Korrelation vor.

3.3 Rangkorrelation

Die Rangstatistik setzt sich zum Ziel, den Einfluss der Skalenwahl zu entfernen, also zu erreichen, dass die statistischen Indikatoren nicht von den absoluten Zahlenwerten der Messungen abhängen. Dies erfolgt, indem die Messwerte durch ihre Rangnummern nach

Größenordnung ersetzt werden. Nicht ganzzahlige Ränge ergeben sich durch Mittelbildung der Ränge bei mehrfachem Auftreten gleicher Messwerten (Bindungen).

Diese Vorgangsweise gewinnt erst bei zwei- und mehrdimensionalen Messgrößen an Bedeutung. Abhängigkeiten werden dann allein durch Beziehungen zwischen den Rängen dargestellt und können durch den **Spearman-Rangkorrelationskoeffizienten** ρ bewertet werden. Der Spearman-Koeffizient ρ ist nichts anderes als der Pearson-Korrelationskoeffizient R, berechnet aus den Rängen des Datensatzes.

Bsp. 3.11 (Fortsetzung Bsp. 3.9) Die Ränge der Messgrößen Bewölkungsgrad und Temperatur sind in Tabelle 3.1 ersichtlich. Der aus diesen Rängen berechnete Pearson-Koeffizienten ergibt den Spearman-Rangkorrelationskoeffizienten

$$\rho = 0.5702.$$

Dieser unterscheidet sich hier nicht wesentlich vom in Bsp. 3.9 ausgerechneten Pearson-Koeffizienten $R = 0.5685$.

Mit MATLAB erhalten wir ρ mit `corr(x,y,'type','Spearman')`.

Bedeckung	Temperatur	Rang Bedeckung	Rang Temperatur	Bed.	Temp.	Rang Bed.	Rang Temp.
13	-6.8	10.0	5.0	7	-2.8	3.5	17.0
0	-7.1	1.5	3.5	90	-4.4	28.0	14.5
0	-7.8	1.5	2.0	73	-4.7	26.0	11.0
10	-8.2	7.0	1.0	70	0.9	24.0	20.0
27	-7.1	12.0	3.5	53	2.4	21.0	22.0
10	-5.7	7.0	7.0	57	1.9	23.0	21.0
27	-5.0	12.0	10.0	97	3.0	29.5	23.0
10	-4.5	7.0	12.5	53	5.2	21.0	30.0
7	-4.4	3.5	14.5	50	5.6	19.0	31.0
27	-5.1	12.0	9.0	97	4.1	29.5	26.5
73	-6.0	26.0	6.0	100	4.3	31.0	28.5
53	-4.1	21.0	16.0	30	4.3	14.0	28.5
33	-2.7	15.5	18.0	40	4.1	17.5	26.5
10	-1.5	7.0	19.0	73	3.2	26.0	24.5
10	-5.5	7.0	8.0	40	3.2	17.5	24.5
33	-4.5	15.5	12.5				

Tabelle 3.1: Bewölkungsgrad, mittlere Tagestemperaturen und deren Ränge. Innsbruck, Jänner 2002.