Beschreibende Statistik

Teilskript zur Vorlesung “Stochastik” der Lehrerweiterbildung Mathematik

Ralph-Hardo Schulz

Aufgabe der beschreibenden Statistik (deskriptiven Statistik) ist die Analyse einer gegebenen Gruppe von Daten, zunächst ohne Schlüsse auf andere, insbesondere zu erwartende Daten bzw. Ereignisse.

1 Einige Grundbegriffe

Bei einer Datenerhebung werden ein oder mehrere **Merkmale** (Beobachtungs-Merkmale) untersucht. Die beobachteten Werte heißen **Merkmalsausprägungen**.

Beispiele:

Population	Merkmal	Merkmalsausprägung
Tiere einer Herde	Gewicht	g [kg] mit $g \in \mathbb{Q}^+$
	Geschlecht	m/w
	Gesundheitszustand	gut/schlecht
Kandidaten einer Bundestagswahl	Erststimmen-Anzahl	$n \in \mathbb{N}$
Menge von Bolzen	Durchmesser	d [mm] mit $d \in \mathbb{R}^+$

Dabei unterscheidet man zwischen *qualitativem* Merkmalstyp und *quantitativ/ metrischem* Typ (also solchem mit Ausprägungen, die durch Messen oder Zählen bestimmt werden können).

Die Menge der Merkmalsträger, z.B. die Tiere einer Herde oder eine Kollektion von Bolzen, heißt dann **Grundgesamtheit**, **Population**, Kollektiv oder Beobachtungseinheit.
Oft ist allerdings eine Untersuchung der gesamten interessierenden Population nicht möglich, man trifft dann eine (repräsentative ?) Teilauswahl, untersucht also eine **Stichprobe**.
Die empirisch gewonnen Daten werden in einer detaillierten Liste, der **Urliste**, eingetragen; erfasst werden (u.U. durchnummeriert) alle Elemente der Stichprobe mit den jeweiligen Merkmalausprägungen.

Bolzen-Nr.	Durchmesser [mm]	Nr.	Durchm. [mm]	Nr.	Durchm. [mm]	Nr.	Durchm. [mm]	Nr.	Durchm. [mm]
1	8,6	6	9,1	11	9,6	16	8,0	21	9,7
2	7,6	7	8,4	12	10,1	17	8,2	22	10,4
3	9,0	8	10,4	13	8,0	18	9,0	23	8,8
4	8,1	9	9,3	14	8,7	19	7,7	24	9,9
5	10,0	10	8,3	15	8,4	20	9,0	25	7,9

Tabelle 1: Beispiel einer Urliste (Beispiel 1)

2 Darstellung (“Visualisierung”) empirischer Daten (Beispiele)

Zur graphischen Darstellung eines (nicht zu großen) Datensatzes mit metrischem Merkmalstyp bieten sich u.a. an:

- ein **Scatter-Diagramm** (Streudiagramm, Punktdiagramm), d.h. Punkte mit den (kartesischen) Koordinaten (x, y), wobei
  x Variable für die Nr. des Merkmalträgers ist und
  y für die Ausprägung des Merkmals bei x steht
  (vgl. Abbildung 1)

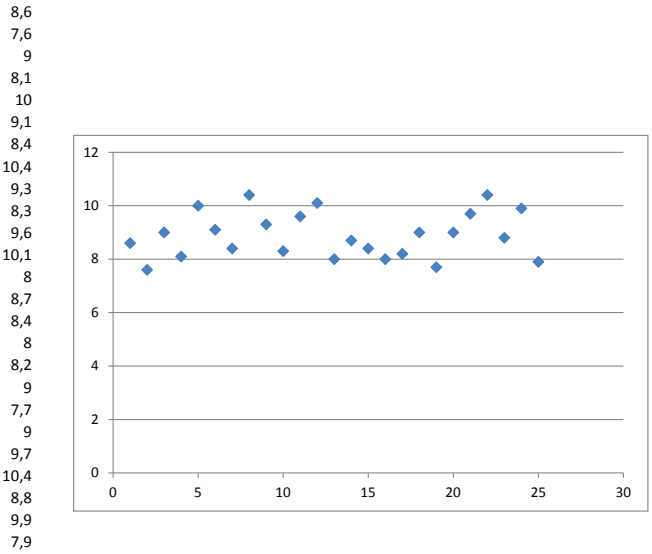

Abbildung 1: Scatter-Diagramm (erstellt mit Excel)

- ein **Balkendiagramm** (Stabdiagramm, Säulendiagramm) (mit Balken, die von der $x-$Achse zum Scatterpunkt führen, s. Abbildung 2)

- ein **“Tortendiagramm” (Kreisdiagramm)**[1], das z.B. zur Darstellung der Prozentzahlen bei einer Wahl (als relative Größe der “Tortenstücke”, also der Sektoren einer Kreisscheibe) (vgl. Abbildung 3) oder der Darstellung von anderen Prozentverteilungen dient.

[1] Zu den Nachteilen des Kreisdiagramms s.z.B. http://de.wikipedia.org/wiki/Tortendiagramm!

8,6
7,6
9
8,1
10
9,1
8,4
10,4
9,3
9,6
10,1
8
8,7
8,4
8
8,2
9
7,7
9
9,7
10,4
8,8
9,9
7,9

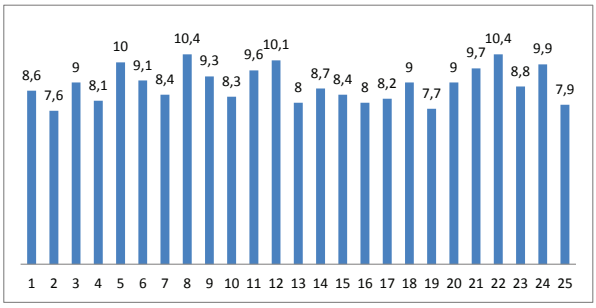

Abbildung 2: Balkendiagramm (erstellt mit Excel)

	Wahlausgang
CDU	32,10%
SPD	27,40%
Grüne	12,60%
Linke	8%
Piraten	7,20%
FDP	4,30%
sonstige	7,80%

Abbildung 3: ’dreidimensionales’ Tortendiagramm (Beispiel)

Aus der Urliste (oder direkt bei der Datenerhebung) kann man eine **Strichliste** mit den Häufigkeiten des Auftretens einer Merkmalausprägung erstellen:

Durchmesser	Häufigkeit	Durchmesser	Häufigkeit	Durchmesser	Häufigkeit
7,6	|	8,6	|	9,6	|
7,7	|	8,7	|	9,7	|
7,8		8,8	|	9,8	
7,9	|	8,9		9,9	|
8,0	||	9,0	|||	10,0	|
8,1	|	9,1	|	10,1	|
8,2	|	9,2		10,2	
8,3	|	9,3	|	10,3	
8,4	||	9,4		10,4	||
8,5		9,5			

Tabelle 2: Strichliste zur Urliste aus Tabelle 1

Aus dieser Liste (mit N Beobachtungseinheiten) kann man dann (evtl. nach Klassenbildung, s.u., Tabelle 3) ein Diagramm erstellen, bei dem die Häufigkeiten n_i der Merkmalsausprägungen i (bzw. deren relative Häufigkeiten $h_i = n_i/N$) aufgetragen werden (s. auch unten, Definition 1).

2.1 Darstellung mit Klassen

Eine Urliste enthält zwar alle benötigten Daten, ist aber meist nicht sehr übersichtlich bzw. infolge der Anzahlen der Merkmalträger oder Merkmalausprägungen nicht mehr sinnvoll darstellbar. Daher wählt man eine geeignete **Klasseneinteilung** der Messgröße, wodurch allerdings Information verloren geht; es entsteht ein **Gruppierungsfehler**. Die Wahl der Klassengrenzen bzw. Klassenbreite wird durch den Zweck der Untersuchung bestimmt und stellt einen Kompromiss zwischen Klarheit (wenige übersichtliche Gruppen) und Information (genaue Angabe der Einzelwerte) dar.
Faustregel: Man bestimmt eine Gesamtzahl von 5 bis 15 Gruppen derart, dass in der Klasse mit den meisten Werten etwa 20 bis 25% der Werte enthalten sind.

Daten, die auf eine Klassengenze fallen, werden je zur Hälfte beiden Klassen zugerechnet[2]; fällt eine ungerade Anzahl auf die Klassengrenze, so wird der überzählige Wert der Gruppe mit den kleineren Werten angerechnet.

[2]Dies wird aber nicht einheitlich so gehandhabt; oft werden halboffene Intervalle als Klassen genommen; z.B. werden bei der Erstellung von Histogrammen mit Excel (s. unten, Abbildung 5) die auf Klassengrenzen fallenden Daten zur kleineren Klasse gerechnet.

Nach Erstellen der entsprechenden Liste kann man diese dann analog zu Abschnitt 1.2 graphisch darstellen. Dabei werden die Anzahlen der jeweiligen Klassenmitte zugeordnet (in der Annahme, dass die Daten innerhalb der Klasse gleichmäßig verteilt sind). Sinnvoll und üblich ist aber auch eine Darstellung als **Histogramm**. Bei diesem werden direkt nebeneinander liegende Rechtecke von der Breite der jeweiligen Klasse gezeichnet, deren Flächeninhalt der Klassenhäufigkeit entspricht. Bei gleicher Klassenbreite stellt die Länge des Rechtecks also die Häufigkeit dar.

Die absolute Häufigkeit einer Merkmalausprägung allein ist allerdigs nicht sehr informativ, wenn man nicht gleichzeitig den Umfang N des Datensatzes kennt. Üblich ist daher die Angabe der relativen Häufigkeit, also der Quotient der absoluten Häufigkeit und N (siehe unten, Definition 1.1). In der graphischen Darstellung bedeutet der Übergang von der absoluten zur relativen Häufigkeit lediglich eine lineare Veränderung des Ordinatenmaßstabs (Multiplikation mit $\frac{1}{N}$).

Klassen-Nr.	Klasse	Klassenmitte	absolute Häufigkeit	relative Häufigkeit
1	7,5–8,0	7,75	4	0,16
2	8,0–8,5	8,25	6	0,24
3	8,5–9,0	8,75	5	0,20
4	9,0–9,5	9,25	3	0,12
5	9,5–10,0	9,75	4	0,16
6	10,0–10,5	10,25	3	0,12

Tabelle 3: Klasseneinteilung und Klassenhäufigkeit bei Beispiel 1 (s. Tabelle 1)

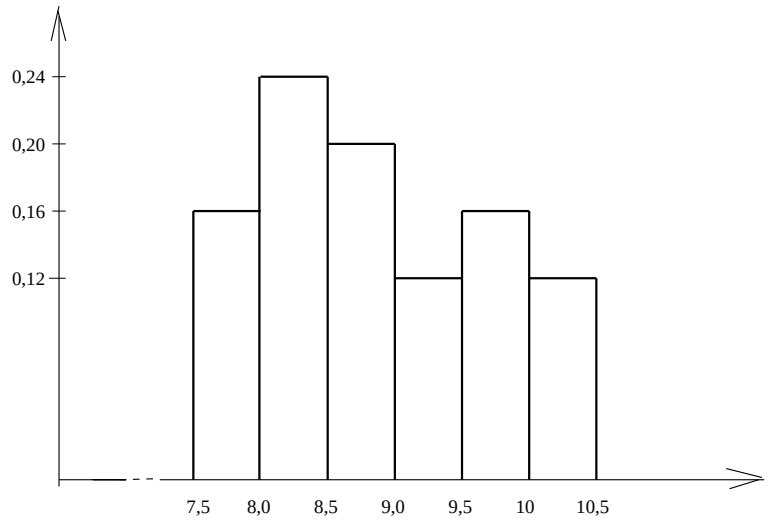

Abbildung 4: Histogramm zu Beispiel 1 (s. Tabelle 3)

Klasse	Häufigkeit
7	0
7,5	0
8	5
8,5	5
9	6
9,5	2
10	4
10,5	3
11	0
und größer	0

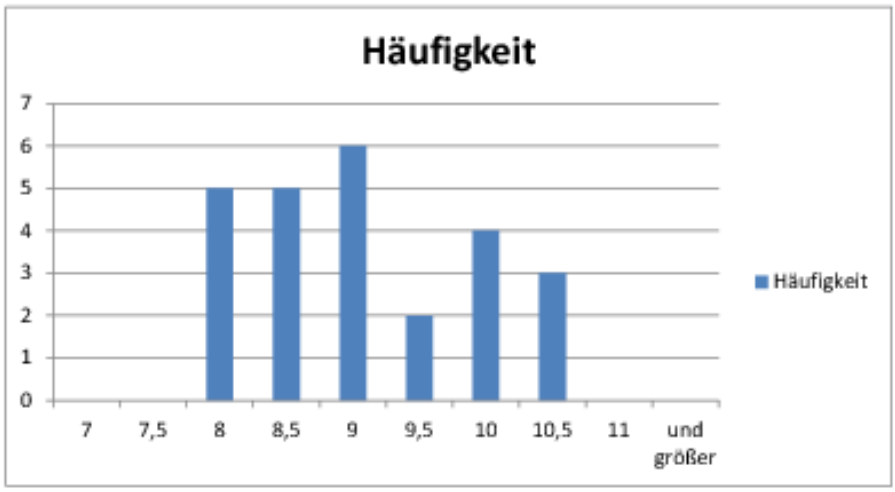

Abbildung 5: (Alternatives) “Histogramm” zum Beispiel 1 (erstellt mit Excel)

3 Quantifizierung von Verteilungen

3.1 Relative Häufigkeit

Wir holen folgende Definitionen nach:

3.1.1 Definition

Kommt ein Wert a genau k mal in einem Datensatz vor, so heißt k die **absolute Häufigkeit** $[a]$ von a; enthält der Datensatz N Daten, so heißt

$$h(a) := \frac{[a]}{N}$$

die **relative Häufigkeit** von a. Nach der Klassenbildung spricht man von absoluter bzw. relativer **Klassenhäufigkeit**. Ist x die Variable für den Wert der Merkmalausprägung, so schreibt man auch $h(x = a)$ statt $h(a)$.

3.1.2 Hilfssatz

Sind $a_1, a_2, \ldots, a_s$ die möglichen Werte des betrachteten Datensatzes vom Umfang N (ohne Berücksichtigung der Vielfachheiten), so gilt

$$0 \leq h(a_i) \leq 1 \text{ für } \quad i = 1, \ldots s \quad \text{und} \quad \sum_{i=1}^{s} h(a_i) = 1.$$

Beweisskizze:
Man dividiere $0 \leq [a_i] \leq N$ durch N; ferner erhält man $\sum h(a_i) = \frac{1}{N}\sum [a_i] = \frac{N}{N} = 1$.
$\square$.

3.2 Verteilung und Summenhäufigkeit

3.2.1 Definition

(i) Die Funktion $h : a_i \mapsto h(x = a_i)$ mit Definitionsbereich $\{a_1, a_2, \ldots, a_s\}$ und Wertebereich $W \subseteq [0,1]$ heisst **Häufigkeitsfunktion (Verteilung)** der Variablen (**Größe**) x in dem betrachteten Datensatz.

(ii) Neben der Häufigkeit ist auch die (diskrete) **Summenhäufigkeit** von (zunächst theoretischer) Bedeutung:

$$\boxed{h(x \leq t) := \sum_{a_j \leq t} h(x = a_j)}.$$

Die Summenhäufigkeit ist also die Häufigkeit, dass die Klassenmitte kleiner oder gleich t ist.

(iii) Dabei heißt $F : \mathbb{R} \to [0,1]$ mit

$$F(t) = h(x \leq t)$$

die **Verteilungsfunktion** der Merkmalausprägung beim betrachteten Datensatz.

Zum Beispiel 1 mit Klassenbildung gemäß Tabelle 3 ergibt sich so:

a_i	$h(x = a_i)$	$h(x \leq a_i)$
7,75	0,16	0,16
8,25	0,24	0,40
8,75	0,20	0,60
9,25	0,12	0,72
9,75	0,16	0,88
10,25	0,12	1,00

und

$$F(t) = \begin{cases} 0 & t < a_1 \\ 0,16 & a_1 \leq t < a_2 \\ 0,4 & a_2 \leq t < a_3 \\ 0,6 & \text{falls } a_3 \leq t < a_4 \\ 0,72 & a_4 \leq t < a_5 \\ 0,88 & a_5 \leq t < a_6 \\ 1 & a_6 \leq t \end{cases}$$

Ein Historamm zu dieser Verteilung ist in Abbildung 6 dargestellt.

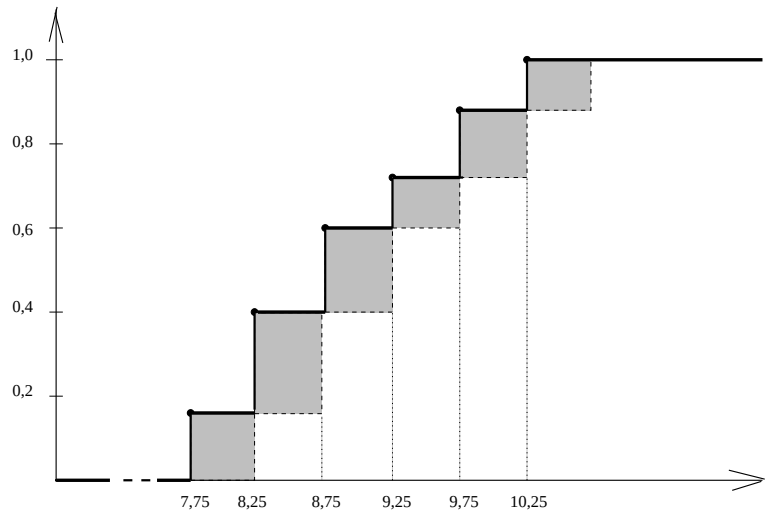

Abbildung 6: Histogramm der Summenhäufigkeit zu Beispiel 1 (Tabelle 3)
(Die schraffierten Rechtecke entsprechen dabei denjenigen aus dem Histogramm der relativen Häufigkeiten in Abbildung 4.)

3.3 Parameter empirischer Verteilungen

Die Beschreibung einer empirischen Stichprobe durch Häufigkeitsliste und Schaubild ist oft zu umständlich. Für viele praktische Probleme genügt es, gewisse charakteristische Eigenschaften des Zahlenmaterials durch entsprechende Kennzahlen, den sogenannten **Parametern**, auszudrücken. Wir betrachten hier die wichtigsten: den (emprischen) Mittelwert, den Median, die Quartile, die Streuung und die Varianz der Stichprobe.

3.3.1 Definition

Sind $x_1, \ldots, x_N$ die beobachteten Stichprobenwerte für die Größe x , so heißt

$$(*) \qquad \boxed{\overline{x} := \frac{1}{N} \sum_{j=1}^{N} x_j}$$

empirische Mittelwert für die Größe x bei der betrachteten Stichprobe. Nimmt x dabei die Zahlenwerte $a_1, \ldots, a_s$ mit den relativen Häufigkeiten $h(a_i)$ an, so ist

$$(**) \qquad \boxed{\overline{x} = \sum_{i=1}^{s} a_i \, h(x = a_i).}$$

Anmerkung zum Zusammenhang zwischen $(*)$ und $(**)$: Der empirische Mittelwert ist das (nach Häufigkeiten gewichtete) arithmetische Mittel der Klassen(mitten)werte: Ist n_i die absolute Häufigkeit des Auftretens von a_i, so gilt

$$\begin{aligned}\overline{x} &= \sum_i a_i \cdot \frac{[a_i]}{N} = \frac{1}{N}\sum_i a_i n_i \\ &= \frac{\overbrace{a_1+a_1+\ldots+a_1}^{n_1\text{mal}}+\overbrace{a_2+a_2+\ldots+a_2}^{n_2\text{mal}}+\ldots+\overbrace{a_s+a_s+\ldots+a_s}^{n_s\text{mal}}}{n_1+n_2+\ldots+n_s}.\end{aligned}$$

Anmerkung: Bei Klassenbildung nimmt man, wie schon in (2.1) erwähnt, meist die Klassenmitten als die Werte von x. *Beispiel* zum Datensatz von Tabelle 3:
$\overline{x} = 7,75\cdot 0,16+8,25\cdot 0,24+8,75\cdot 0,2+9,25\cdot 0,12+9,75\cdot 0,16+10,25\cdot 0,12$
$= 8,87$ [mm]
Durch die Klasseneinteilung ist in unserem Beispiel $\overline{x}$ allerdings nicht gleich dem arithmetischen Mittel der Werte der Urliste! Im Beispiel ist das arithmetische Mittel der Werte aus der Urliste gleich $8,888$.
Es kann sein, dass einzelne "Ausreisser" den Mittelwert zu stark beeinflussen.
Beispiel:[3] Von 10 Diplomkandidat(inn)en eines Jahrgangs haben sich 2 mit 11 Semestern, 3 mit 12 Semestern, 2 mit 13 und 2 mit 14 Semestern zum Examen gemeldet, aber einer mit 51 Semestern (ist so ähnlich vorgekommen!). Der Mittelwert liegt dann bei 16,3 Semestern, obwohl 90% das Studium wesentlich schneller geschafft haben (und nährt den Vorwurf überlangen Studiums). Wie vermeidet man diese Verfälschung?

3.3.2 Definition

Nummeriert man die N Werte des Datensatzes der Größe nach (unter Beachtung der Vielfachheit des Auftretens im Datensatz) durch

$$x_1,\ x_2,\ \ldots,\ x_N,$$

so heißt der diesen Datensatz halbierende Wert

$$\hat{x} := x_{(\frac{N-1}{2}+1)} \quad \text{für ungerades } N \quad \text{bzw.} \quad \hat{x} := \frac{1}{2}(x_{\frac{N}{2}} + x_{(\frac{N}{2}+1)}) \quad \text{für gerades } N$$

der **Median** (Zentralwert) des Datensatzes.

[3] Vgl. damit die "Anekdote" in [Behrends] p.277 oben"!

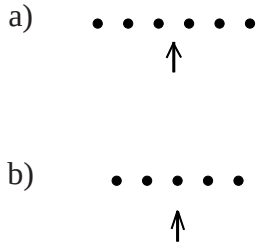

Abbildung 7: Lage des Medians bei gerader bzw. ungerader Anzahl von Daten (schematische Beispiele)

Der Median teilt die geordnete Datenliste in zwei Teillisten; deren Mediane heissen **unteres** bzw. **oberes Quartil** (oder 0,25-Qua**n**til bzw. 0,75-Qua**n**til).
Im Beispiel 1 (s.Tabelle 2) ist der Median gleich 8,8 (bzw. nach Klassenbildung, s. Tabelle 3, gleich 8,75); das untere Quartil ist 8,15 (in der unteren Teilliste ohne den Median), das obere Quartil gleich 9,65.

Mittelwert und Median kennzeichen die *Lage* der Verteilung; ebenso wichtig sind aber die **Streuungsmaße**. Als solches dienen die Differenz von oberem und unterem Quartil, öfter aber die Varianz und Streuung:

3.3.3 Definition

Ist $\overline{x}$ der Mittelwert des Datensatzes mit Werten $a_1, \ldots a_s$. Die mittlere quadratische Abweichung

$$\boxed{\mathrm{Var}(x) := \sum_{i=1}^{s} (a_i - \overline{x})^2 h(x = a_i)}$$

heißt (empirische) **Varianz** (Streuungsquadrat, Stichprobenvarianz) der Größe x (andere Bezeichnungen: $s^2(x)$ oder $\sigma^2(x)$; und

$$\boxed{s(x) := \sqrt{\mathrm{Var}(x)}}$$

wird **Standardabweichung** (Stichprobenstreuung) genannt.
Anmerkung: Sind $x_1, \ldots, x_N$ die Einzelwerte des Datensatzes unter Berücksichtigung der Häufigkeit ihres Auftretens (und $a_1, \ldots a_s$ die möglichen Ausprägungen), so gilt

$$\mathrm{Var}(x) = \sum_{i=1}^{s} (a_i - \overline{x})^2 \frac{[a_i]}{N} = \frac{1}{N} \sum_{j=1}^{N} (x_j - \overline{x})^2.$$

Die Varianz ist also der Durchschnitt der Quadrate der Differenzen zwischen Einzelwert und Mittelwert.
Beispiel: Nach Klassenbildung bei Beispiel 1 (Tabelle 3) erhält man

$\mathrm{Var}(x) \approx 0,16 \cdot (7,75-8,87)^2 + 0,24 \cdot (8,25-8,87)^2 + 0,2 \cdot (8,75-8,87)^2 + 0,12 \cdot (9,25-8,87)^2 + 0,16 \cdot (9,75-8,87)^2 + 0,12 \cdot (10,25-8,87)^2 \approx 0,67\ [\mathrm{mm}^2]$
und damit $s(x) = \sqrt{\mathrm{Var}(x)} \approx 0,82$.
Achtung: Aus Gründen, die wir erst später thematisieren, wird oft auch

$$\tilde{s}^2(x) = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \overline{x})^2$$

als (empirische) Varianz und $\tilde{s} = \sqrt{\tilde{s}^2}$ als empirische Streuung der Stichprobe bezeichnet.
In unserem Beispiel bedeutet das $\tilde{s}^2 = \frac{25}{24} \cdot s^2 \approx 0,69$ und $\tilde{s} = \sqrt{\tilde{s}^2} \approx 0,83$.

3.3.4 Hilfssatz

* Es seien $x_1, \ldots, x_N \in \mathbb{R}$ und $t \in \mathbb{R}$. Dann gilt:

(i) $t_0 = \overline{x} \Leftrightarrow t = t_0$ minimiert $\sum_{i=1}^{N} (x_i - t)^2$.

(ii) t_1 ist der Median von $x_1, \ldots, x_N \Leftrightarrow t = t_1$ minimiert $\sum_{i=1}^{N} |x_i - t|$.

Beweisskizze:
(i) 1. Beweis-Möglichkeit: mittels Ableitung; vgl. Übungsaufgabe U2.
2. Möglichkeit (s.[Behrends] Satz 9.3.3, p. 275 f.):
Man betrachte die folgenden Vektoren aus $\mathbb{R}^N$: $\begin{array}{rcl} A & = & (\overline{x}, \ldots, \overline{x}) \\ B_t & = & (t, \ldots, t) \\ C & = & (x_1, \ldots, x_N) \end{array}$.
Mit dem kanonischen Skalarprodukt erhält man:

$$\begin{aligned} (B_t - A) \cdot (A - C) &= \sum_{i=1}^{N} (t - \overline{x})(\overline{x} - x_i) \\ &= (t - \overline{x}) \sum_{i} (\overline{x} - x_i) \\ &= (t - \overline{x})(N\overline{x} - \sum_{i} x_i) = 0 \end{aligned}$$

d.h. $(B_t - A) \perp (A - C)$. Nach dem Satz des Pythagoras folgt

*Vgl. [Behrends] p.274

$$||B_t - C||^2 = ||B_t - A||^2 + ||A - C||^2.$$

Da $||A-C||^2$ konstant bleibt, wenn $B_t = t(1,\ldots,1)$ alle möglichen Vektoren durchläuft, so ist $||B_t - C||$ minimal für $B_t = A$, also für $t = \overline{x}$. □

Anmerkung: Geometrisch gesehen ist A beste Approximation an C durch Elemente des 1-dim Unterraumes $W = \{t(1,\ldots,1)|t \in \mathbb{R}\}$.

(ii) s.[Behrends] l.c.

3.4 Korrelation und Regression

Um Zusammenhänge zwischen zwei quantitativen Merkmalen zu beschreiben (z.B. zwischen Gewicht und Größe von Personen), betrachtet man Stichproben (x_i, y_i) (mit $i = 1,\ldots,N$) einer Grundgesamtheit. Es sei $\overline{x}$ Mittelwert von $x_1,\ldots,x_N$ und $\overline{y}$ derjenige von $y_1,\ldots,y_N$. Wir definieren $p_i := (x_i - \overline{x})(y_i - \overline{y})$ und überlegen: Wenn die x_i dann groß (bzw. klein) sind, wenn es die y_i sind, so sind die p_i tendenziell positiv; bei gegensätzlichem Verhalten der Größen x und y sind die p_i meist negativ. Bei Unabhängigkeit von x und y gibt es keine bevorzugte Tendenz. So kommen wir (durch Skalierung) zu

3.4.1 Definition

Sind nicht alle x_i gleich $\overline{x}$ und nicht alle y_i gleich $\overline{y}$, so versteht man unter dem **Korrelationskoeffizient**en der (quantitativen) Stichprobe $(x_1,y_1),\ldots,(x_N,y_N)$ die Zahl

$$\boxed{r_{xy} := \frac{\sum\limits_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{N}(x_i-\overline{x})^2 \cdot \sum\limits_{i=1}^{N}(y_i-\overline{y})^2}}.}$$

Eigenschaften:
Setzt man $\tilde{x} := (x_1 - \overline{x},\cdots,x_N - \overline{x})$ und $\tilde{y} := (y_1 - \overline{y},\cdots,y_N - \overline{y})$, so erhält man

$$r_{xy} = \frac{\tilde{x}\cdot\tilde{y}}{||\tilde{x}||\cdot||\tilde{y}||} = \cos\sphericalangle(\tilde{x},\tilde{y}) \in [-1,+1].$$

Man kann zeigen, dass im Falle $r_{xy} \approx 1$ die Werte x_i die gleiche 'Tendenz' haben wie die y_i, im Falle $r_{xy} \approx -1$ entgegengesetzte Tendenz, im Falle $r_{x,y} = 0$ eine Art Unabhängigkeit.
Dass r_{xy} symmetrisch in x und y ist, zeigt, dass man im Falle $|r_{xy}| = 1$ aber nicht von einem kausalem Zusammenhang ' x steuert y' ausgehen kann.

Eine genauere Analyse der Korrelation erhält man durch die Suche nach einer Geraden, die sich der "Punktwolke" $\{(x_i, y_i) | i = 1, \ldots, N\}$ möglichst gut anpasst.

3.4.2 Definition

Eine Gerade mit Gleichung $y = a + bx$ heißt **Regressionsgerade**, falls (der quadratische Abstand der Abszissen)

$$\sum_{i=1}^{N} \big(y_i - (a + bx_i)\big)^2$$

unter allen möglichen Werten von a und b minimal ist.

3.4.3 Satz

Gilt $s_x > 0$ für die Stichprobenstreuung s_x von (x_i), so sind die Koeffizienten der (eindeutig bestimmten) Regressionsgeraden mit Gleichung $y = a + bx$ durch

$$b = \frac{r_{xy} s_y}{s_x} \quad \text{und} \quad a = \overline{y} - b\overline{x}$$

gegeben.

Beweisidee:
Man sucht das Minimum von $\Phi(a, b) = \sum_i [y_i - (a + bx_i)]^2$ auf $\mathbb{R}^2$ durch Null-Setzten der partiellen Ableitungen (und Prüfen, ob man so ein Minimum erhalten hat).

Bei positiver Steigung der Regressionsgeraden geht man oft davon aus, dass eine Zunahme des x-Merkmals "in der Regel" eine Zunahme des y-Merkmals impliziert. Dies ist nicht immer gerechtfertigt: S. das

3.4.4 Simpson-Paradoxon

. . . s. [Behrends] p.281.

Literaturauswahl zur beschreibenden Statistik

EHRHARD BEHRENDS: Elemenmtare Stochastik. Springer Spektrum-Verlag, Wiesbaden 2013, Kapitel 9.

GÜNTHER BOURIER: Beschreibende Statistik: Praxisorientierte Einführung - Mit Aufgaben und Lösungen. Gabler Verlag, 2011^9 (im OPAC online)

MARCO BURKSCHAT, ERHARD CRAMER & UDO KAMPS: Beschreibende Statistik: Grundlegende Methoden der Datenanalyse (EMIL@A-stat). Springer Spektrum 2012^2 (im OPAC online)

HERBERT EGGS: Stochastik I. Diesterweg/Salle/Sauerländer, 1984. (vergriffen; in der Universitätsbibliothek der FU vorhanden)

LUDWIG FAHRMEIR, RITA KÜNSTLER, IRIS PIGEOT & GERHARD TUTZ: Statistik: Der Weg zur Datenanalyse. Springer-Lehrbuch, 2009^7 (elektronischer Fernzugriff über den OPAC)

MARTIN HENGST: Einführung in die Mathematische Statistik und ihre Anwendung. BI, 1967 (noch gebraucht zu kaufen; in der FB-Bibliothek vorhanden)

ERWIN KREYSZIG: Statistische Methoden und ihre Anwendungen. Vandenhoek & Ruprecht, Göttingen 1965, 1979^7 (in der FB-Bibliothek vorhanden)

WOLF-GERT MATTHÄUS & JÖRG SCHULZE: Statistik mit Excel: Beschreibende Statistik für jedermann. Vieweg+Teubner Verlag 2005^2 (im OPAC online)

LEOPOLD SCHMETTERER: Einführung in die mathematische Statistik. Wien und New York, Springer 1966^2 (2.Kapitel) (in der Universitätsbibliothek der FU vorhanden)

PETER M.SCHULZE: Beschreibende Statistik. Oldenbourg Wissenschaftsverlag 2007 (in der Universitätsbibliothek der FU vorhanden)
Interdisziplinäre Gruppe Theoretische Biologie am IZMB, Univ.Bonn. Vorlesung Kap.2. www.theobio.uni-bonn.de/studies/files/vorlesung2.pdf

---

Version des Skripts vom 23. Januar 2013