Statistik I für WInf und WI

Prof. Dr. Wilhelm Stannat

Inhalt:

Das vorliegende Skript ist die Zusammenfassung des ersten Teils der Vorlesung Statistik I für WInf und WI im Wintersemester 2009/10. Die Lektüre des Skriptes ist kein gleichwertiger Ersatz für den Besuch der Vorlesung.

Korrekturen bitte per Email an: `stannat@mathematik.tu-darmstadt.de`

I. Deskriptive Statistik

1. Grundbegriffe

Die deskriptive oder auch beschreibende Statistik beschäftigt sich mit der Erhebung und Aufbereitung von Daten, die im Rahmen von Erhebungen, wie zum Beispiel Volkszählungen und Umfragen, oder bei Messungen gewonnen werden.
Erhoben werden **Merkmale** wie zum Beispiel Alter, Geschlecht, Einkommen, Temperatur oder Druck. Unterschieden werden Merkmale nach **qualitativen Merkmalen**, wie Geschlecht, Nationalität oder Beruf, und **quantitativen Merkmale**, die man ihrerseits nochmals in **diskrete** Merkmale, etwa Alter und Einkommen, und **stetige** Merkmale, etwa Temperatur und Geschwindigkeit unterteilt.
Die **Merkmalsausprägungen** sind die Gesamtheit der möglichen Werte eines Merkmals, also:

Beispiele

Geschlecht: männlich, weiblich

Alter: $0, 1, 2, 3, \ldots$

Temperatur: die reellen Zahlen $\mathbb{R}$ oder Teilmengen der reellen Zahlen

Als **Merkmalsträger** bezeichnet man die für die Erhebung der Daten relevanten Objekte. Das sind also zum Beispiel bei einer Umfrage die Menge der relevanten Personen. Die Gesamtheit der für eine statistische Erhebung relevanten Merkmalsträger heißt **Grundgesamtheit**.

Bei Erhebungen unterscheidet man zwischen einer **Vollerhebung**, bei der alle Merkmalsträger der Grundgesamtheit erfasst werden (etwa Volkszählung) und einer **Teilerhebung** oder **Stichprobenerhebung**, bei der nur eine zufällig gewonnene Teilmenge der Grundgesamtheit erfasst wird, wie es bei Umfragen der Fall ist.

Merkmalstypen, Skalierung, Klassierung

Wir haben bereits die Unterscheidung zwischen quantitativen und qualitativen Merkmalen angesprochen. Durch **Quantifizierung** kann ein qualitatives Merkmal in ein quantitatives umgewandelt werden, z.B.:

$$\begin{array}{ll} \texttt{grün} & = 23 \\ \texttt{blau} & = 14 \end{array} \quad \text{oder} \quad \begin{array}{ll} \texttt{Europa} & = 3 \\ \texttt{Asien} & = 1 \end{array}$$

Skalierung

Bei quantitativen Merkmalen spielt die Skalierung eine wichtige Rolle. Man unterscheidet folgende Skalen:

Nominalskala: die zugeordneten Zahlen dienen lediglich zur Unterscheidung der Merkmalsausprägungen

Beispiel Steuerklassen I, II, ..., V.

Ordinalskala, Rangskala: die Merkmalsausprägungen werden zueinander in einer Rangfolge in Beziehung gesetzt

Beispiel Schadstoffklassen 1, 2, 3, 4.

Kardinalskala: zusätzlich zur Rangfolge spielt auch noch der Abstand zwischen zwei Merkmalsausprägungen eine Rolle

Beispiele Temperatur, Einkommen.

Klassierung
Ein stetig verteiltes Merkmal kann durch die **Aufteilung** der Merkmalsausprägungen **in Teilintervalle (Klassen)** in ein diskretes Merkmal überführt werden.

Beispiel

$$\texttt{Körpergröße in cm} \longrightarrow \texttt{Klassen} \quad \begin{array}{ll} < 160\,\texttt{cm} & 180 \ldots 189\,\texttt{cm} \\ 160 \ldots 169\,\texttt{cm} & 190 \ldots 199\,\texttt{cm} \\ 170 \ldots 179\,\texttt{cm} & \geq 200\,\texttt{cm} \end{array}$$

Bei der Erhebung statistischer Daten unterscheidet man zwischen

- Befragung (z. B. Umfrage, Volkszählung)
- Beobachtung (z. B. Verkehrszählung, Messung,...)
- Experiment (Messung im "physikalischen" Experiment).

Bei der **Teilerhebung** statistischer Daten wird die **Stichprobenauswahl** entscheidend, d. h. von welchen Merkmalsträgern werden die Daten erhoben. Es gibt hierzu, neben **willkürlicher** Auswahl, Stichprobentechniken.

Beispiel Quotenauswahl

Bei der Auswahl achtet man darauf, dass bestimmte Merkmalsausprägungen in der Teilgesamtheit dieselbe relative Häufigkeit besitzen wie in der Grundgesamtheit. Man spricht dann von einer "repräsentativen" Auswahl, im Zusammenhang mit Umfragen etwa von einer repräsentativen Umfrage.

2. Auswertung eindimensionaler Datensätze

Die Gesamtheit der Daten aus der statistischen Erhebung bezeichnet man als **Urliste**. Wird nur ein Merkmal erhoben, so kann man die erhobenen Merkmalswerte als Folge aufschreiben:

$$x_1, x_2, x_3, \ldots, x_n$$

Auf diese Weise erhält man eine **Stichprobe der Länge** n. Alternativ spricht man auch von einer **Messreihe**, sowie statt von Merkmalswerten auch von **Messwerten** oder **Beobachtungen**.

Beispiel Jahreshöchsttemperaturen (in $^\circ C$) in Darmstadt in den Jahren 1996 - 2005

$$33.0 \quad 33.2 \quad 36.5 \quad 32.2 \quad 34.2 \quad 34.4 \quad 37.2 \quad 38.1 \quad 32.3 \quad 34.7$$

Absolute und relative Häufigkeiten

Es seien $a_1, a_2, \ldots, a_s$ die möglichen Merkmalsausprägungen. Die Anzahl der Merkmalswerte $x_1, \ldots x_n$, die mit a_j übereinstimmen, heißt **absolute Häufigkeit** von a_j und wird mit $h(a_j)$ bezeichnet $(j = 1, \ldots, s)$.

Der Anteil

$$f(a_j) := \frac{h(a_j)}{n} \qquad (j = 1, \ldots, s)$$

des Merkmalswertes a_j an der Gesamtzahl n der erhobenen Merkmalswerte heißt **relative Häufigkeit**. An den relativen Häufigkeiten kann man insbesondere sofort die Prozentanteile ablesen.

Offenbar gilt:

$$\sum_{j=1}^{s} h(a_j) = n \qquad \text{und} \qquad \sum_{j=1}^{s} f(a_j) = 1.$$

Graphische Darstellungen der Häufigkeitsverteilung

Die gängigen graphischen Darstellungen von Häufigkeitsverteilungen sind

- Tabellen
- Stabdiagramme und Histogramme
- Kreisdiagramme.

Beispiel Stimmenverteilung bei der Bundestagswahl 2005

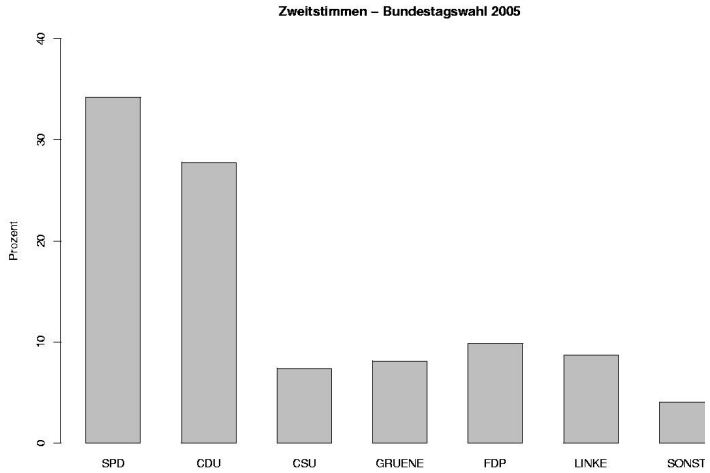
Das erhobene Merkmal ist in diesem Falle die mit der Zweitstimme gewählte Partei. Eine Beobachtungseinheit ist ein Stimmzettel. Die Gesamtheit der Merkmalswerte sind die zur Wahl stehenden Parteien, also SPD, CDU, CSU, usw. Um die Darstellung zu vereinfachen, sind die weniger häufig gewählten Parteien in der Klasse "Sonstige" zusammengefasst. Die Anzahl n der Merkmalswerte ist gleich der Anzahl der gültigen Zweitstimmen, in diesem Falle $n = 47\,287\,988$.

Häufigkeitstabelle

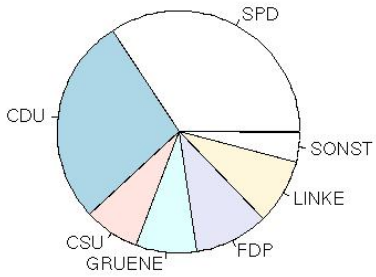
In der Häufigkeitstabelle werden die ermittelten absoluten und/oder relativen Häufigkeiten tabellarisch erfasst.

Partei	Zweitstimmen	Anteil in Prozent
SPD	16 194 665	34.2
CDU	13 136 740	27.8
CSU	3 494 309	7.4
Grüne	3 838 326	8.1
FDP	4 648 144	9.8
Die Linke	4 118 194	8.7
Sonstige	1 912 665	4.0

Stabdiagramm



Kreisdiagramm

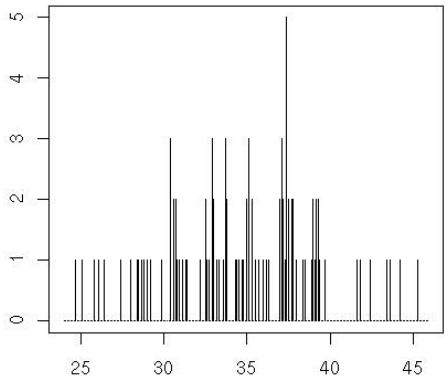


Bei stetigen oder quasistetigen Merkmalen ist die Aufstellung einer Häufigkeitstabelle oder eines Stabdiagramms sinnlos, denn die meisten Werte sind nur einfach oder gar nicht besetzt.

Beispiel

Jährliche Milchleistung von Kühen (in 100 Litern) (n=100).

37.4	37.8	29.0	35.1	30.9	28.5	38.4	34.7	36.3	30.4
39.1	37.3	45.3	32.2	27.4	37.0	25.1	30.7	37.1	37.7
26.4	39.7	33.0	32.5	24.7	35.1	33.2	42.4	37.4	37.2
37.5	44.2	39.2	39.4	43.6	28.0	30.6	38.5	31.4	29.9
34.5	34.3	35.0	35.5	32.6	33.7	37.7	35.3	37.0	37.8
32.5	32.9	38.0	36.0	35.3	31.3	39.3	34.4	37.2	39.0
41.8	32.7	33.6	43.4	30.4	25.8	28.7	31.1	33.0	39.0
37.1	36.2	28.4	37.1	37.4	30.8	41.6	33.8	35.0	37.4
33.7	33.8	30.4	37.4	39.3	30.7	30.6	35.1	33.7	32.9
35.7	32.9	39.2	37.5	26.1	29.2	34.8	33.3	28.8	38.9



Stabdiagramm

Ein Ausweg liefert hier die **Klassierung**. Bei der Wahl der Anzahl der Klassen ist allerdings zu beachten, dass

- bei zu großer Klassenanzahl viele Klassen unbesetzt bleiben,
- bei zu geringer Klassenanzahl Information verloren geht.

Als **Faustregel** gilt, dass die Anzahl der Klassen in etwa $\sqrt{n}$ entsprechen sollte, wobei n die Anzahl der Beobachtungen ist.
In obigem Beispiel erhalten wir bei der Wahl von 8 Klassen der Form

$$[a_1, a_2[,\ [a_2, a_3[,\ [a_3, a_4[,\ [a_4, a_5[,\ [a_5, a_6[,\ [a_6, a_7[,\ [a_7, a_8[,\ [a_8, a_9[$$

mit $a_1 = 24$, $a_2 = 27$, $a_3 = 29.6$, $a_n = 32$, $a_5 = 34.3$, $a_6 = 36.5$, $a_7 = 38.4$, $a_8 = 40.5$, $a_9 = 45.5$ die folgende Häufigkeitstabelle:

Milchleistung	[24, 27[	[27, 29.6[	[29.6, 32[	[32, 34.3[
Anzahl der Milchkühe	5	8	13	18
Milchleistung	[34.3, 36.5[	[36.5, 38.4[	[38.4, 40.5[	[40.5, 45.5[
Anzahl der Milchkühe	17	20	12	7

Im folgenden bezeichne K_j die Anzahl der Merkmalswerte in der Klasse $[a_j, a_{j+1}[$. K_j heißt **Klassenhäufigkeit** oder auch **Besetzungszahl**. Den zugehörigen relativen Anteil

$$k_j := \frac{K_j}{n}$$

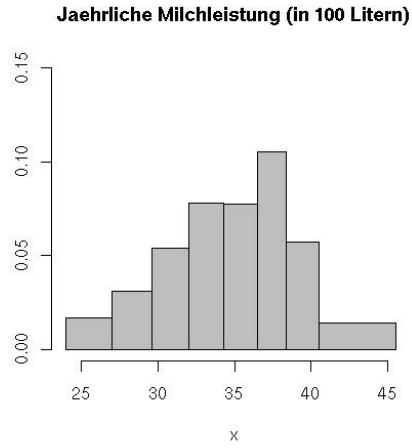
bezeichnet man als **relative Klassenhäufigkeit**.

Zur graphischen Darstellung klassierter Daten eignen sich **Histogramme**. Hierbei wird über jedem der Teilintervalle $[a_j, a_{j+1}[$ ein Rechteck mit der Fläche k_j errichtet. Die Höhe d_j des Rechtecks errechnet sich also gemäß der folgenden Gleichung

$$d_j(a_{j+1} - a_j) = k_j\,.$$

Man beachte, dass bei **gleicher Klassenbreite** nicht nur die Fläche, sondern **auch die Höhe** der Rechtecke proportional zur relativen Klassenhäufigkeit k_j ist.

Histogramm zu obigem Beispiel



Kumulierte Häufigkeitsverteilung

Die Funktion

$$H(x) := \sum_{a_j \leq x} h(a_j) \qquad \text{für } x \in \mathbb{R}$$

heißt **absolute kumulierte Häufigkeitsverteilung**. Sie zählt zu gegebenem $x \in \mathbb{R}$ die Anzahl der Beobachtungswerte die kleiner gleich x sind. Die Funktion

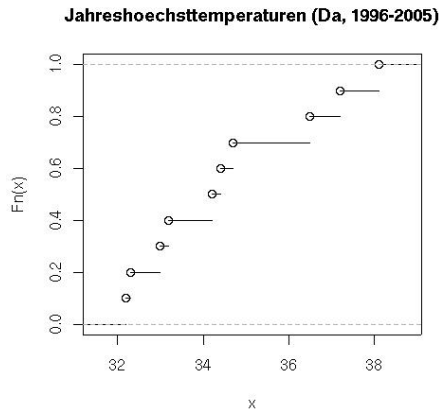
$$F(x) := \frac{1}{n} H(x) = \sum_{a_j \leq x} f(a_j)\,, \qquad x \in \mathbb{R}$$

heißt **relative kumulierte Häufigkeitsverteilung** oder **empirische Verteilungsfunktion**.

Eigenschaften der empirischen Verteilungsfunktion

- F ist eine monoton wachsende Treppenfunktion
- $0 \leq F \leq 1$
- F besitzt Sprünge an den Merkmalsausprägungen a_j

Als Beispiel für den typischen Verlauf einer empirischen Verteilungsfunktion im folgenden die Verteilungsfunktion zu den Jahreshöchsttemperaturen in Darmstadt aus den Jahren 1996-2005.



Lagemaße

Modalwert x_{Mod}

Diejenigen Ausprägungen a_j mit der größten Häufigkeit werden als **Modalwerte** bezeichnet. Die Verwendung des Modalwertes zur Beschreibung von Datensätzen sollte auf den Fall unimodaler Verteilungen beschränkt bleiben.

Median x_{Med}

Der **Median** oder auch **Zentralwert** ist derjenige Wert x_{Med}, für den mindestens 50 % aller Merkmalswerte kleiner gleich x_{Med} und mindestens 50 % aller Merkmalswerte größer gleich x_{Med} sind.
Zur Bestimmung des Medians ordnet man die Werte x_1, ..., x_n zunächst der Größe nach an,

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

und erhält auf diese Weise die sogannte **geordnete Urliste**. Dann definiert man

$$x_{Med} := \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls } n \text{ gerade} \end{cases} \tag{1.1}$$

Arithmetisches Mittel (Durchschnittswert)

Der bekannteste Lageparameter ist das arithmetische Mittel

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{s} a_j f(a_j)\,.$$

Beispiel Preise für Normal-Benzin an 20 örtlichen Tankstellen der Größe nach geordnet:

129.4	129.9	129.9	130.4	131.4
131.4	132.9	132.9	132.9	133.9
134.4	134.4	134.9	134.9	134.9
134.9	135.4	135.4	135.9	136.4

In diesem Beispiel ist $x_{Mod} = 134.9$, $x_{Med} = 134.15$, $\overline{x} = 133.325$. Würde eine Tankstelle als besondere Werbemaßnahme den Benzinpreis von 132.9 auf 125.9 senken, so würde dies den Durchschnittswert $\overline{x}$ von 133.325 auf 132.975 senken. Einen Einfluss auf den Median (oder auf den Modalwert) hätte die Senkung dagegen nicht.

Lagemaße, die nicht empfindlich auf Extremwerte oder Ausreißer reagieren heißen **robust**. Der Median ist also ein robustes Lagemaß.

Bemerkung

(i) Median und arithmetisches Mittel stimmen i.a. nicht mit einer der möglichen Merkmalsausprägungen überein.

Prominentes Beispiel: Durchschnittliche Anzahl der Kinder pro Familie.

(ii) **Äquivarianz unter linearer Transformation** Transformiert man die Daten gemäß einer affin linearen Transformation der Form

$$y_i = a + bx_i\,,$$

so gilt für das arithmetische Mittel

$$\overline{y} = a + b\overline{x}$$

und ebenso

$$y_{Mod} = a + bx_{Mod}\,, \quad y_{Med} = a + bx_{Med}\,.$$

(iii) **Optimalitätseigenschaften** Das arithmetische Mittel $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ **minimiert** die **Summe der quadratischen Abstände**, d.h. es gilt

$$\sum_{i=1}^n (x_i - \overline{x})^2 < \sum_{i=1}^n (x_i - r)^2 \text{ für alle } r \in \mathbb{R}, r \neq \overline{x}\,.$$

Beweis

$$\begin{aligned}\sum_{i=1}^n (x_i - r)^2 - \sum_{i=1}^n (x_i - \overline{x})^2 &= \sum_{i=1}^n \underbrace{(x_i - r)^2 - (x_i - \overline{x})^2}_{-2x_i r + r^2 + 2x_i\overline{x} - \overline{x}^2} \\ &= -2n\overline{x}r + nr^2 + 2n\overline{x}^2 - n\overline{x}^2 \\ &= n(r - \overline{x})^2 > 0 \text{ für } r \neq \overline{x}\,.\end{aligned}$$

Auch Median und Modalwert erfüllen entsprechende Optimalitätskriterien.

- Der Median x_{Med} minimiert die Summe der Abstände, d.h. es gilt

$$\sum_{i=1}^n |x_i - x_{Med}| < \sum_{i=1}^n |x_i - r| \text{ für alle } r \in \mathbb{R}\,, r \neq x_{Med}\,.$$

- Der Modalwert minimiert die Summe

$$\sum_{i=1}^n 1_{\{x_i \neq r\}} \text{ mit } 1_{\{x_i \neq r\}} = \begin{cases} 1 & \text{falls } x_i \neq r \\ 0 & \text{falls } x_i = r\,. \end{cases}$$

Weitere Lagemaße

Annahme: $x_1, \ldots, x_n > 0$

Geometrisches Mittel $\bar{x}_{geom}$

$$\bar{x}_{geom} := (x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$

Findet Verwendung im Zusammenhang mit Wachstums- und Zinsmodellen. Sind etwa $x_1, \ldots, x_n$ die beobachteten Wachstumsfaktoren eines Portfolios mit Anfangsbestand K_0, so ist

$$K_n = K_0 \cdot x_1 \cdot \ldots \cdot x_n$$

der Bestand am Ende der Periode n. Schreibt man

$$K_n = K_0 \left(\underbrace{(x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}}}_{=\bar{x}_{geom}} \right)^n = K_0 \cdot \bar{x}_{geom}^n$$

so lässt sich $\bar{x}_{geom}$ als **mittlerer Wachstumsfaktor** über die n Perioden $1, \ldots, n$ interpretieren.

Beziehung zum arithmetischen Mittel

Logarithmiert man die Messwerte $y_i := \ln x_i$ so folgt

$$\ln \bar{x}_{geom} = \frac{1}{n} \ln(x_1 \cdot \ldots \cdot x_n) = \frac{1}{n} \sum_{i=1}^{n} \ln x_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$

d.h., $\ln \bar{x}_{geom}$ stimmt mit dem arithmetischen Mittel der logarithmierten Messwerte $y_i = \ln x_i$ überein.

Harmonisches Mittel $\bar{x}_{harm}$

$$\bar{x}_{harm} := \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}}$$

Typische Anwendung: Ermittlung von Gesamtdurchschnittswerten aus Durchschnitten über einzelne Teilbereiche.

Beispiel Der ICE von Frankfurt nach Berlin fährt

- 150 km mit durchschnittlich 100 km pro Stunde
- 450 km mit durchschnittlich 200 km pro Stunde

Es sei x_i die Durchschnittsgeschwindigkeit bei Kilometer i, $i = 1, \ldots 600$. Dann beträgt die Durchschnittsgeschwindigkeit über die gesamte Strecke

$$\frac{1}{\frac{1}{600} \left(\frac{150}{100} + \frac{450}{200} \right)} = 160 \left[\frac{km}{h} \right] .$$

Quantile und Box-Plots

Lagemaße alleine reichen zur Beschreibung der Daten einer Urliste nicht aus. Vergleicht man etwa eine Einkommenserhebung in zwei Ländern, so können die Durchschnittseinkommen gleich sein, jedoch in einem Land größere Einkommensunterschiede bestehen als im anderen Land. Daher benötigt man zusätzliche Kennzahlen, um die Lage der Daten möglichst effizient erfassen zu können. Eine wichtige Methode sind **Box-Plots**, die mit Hilfe von Quantilen definiert werden.

Definition Es sei $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ eine geordnete Urliste und $p \in]0,1]$. Jeder Wert x_p mit der Eigenschaft

$$\frac{1}{n}(\text{ Anzahl der Messwerte } \leq x_p) \geq p$$

und

$$\frac{1}{n}(\text{ Anzahl der Messwerte } \geq x_p) \geq 1-p.$$

heißt p-**Quantil**.

Damit folgt

$$x_p = x_{([np]+1)} \text{ falls } np \text{ nicht ganzzahlig}$$
$$x_p \in [x_{(np)}, x_{(np+1)}] \text{ falls } np \text{ ganzahlig.}$$

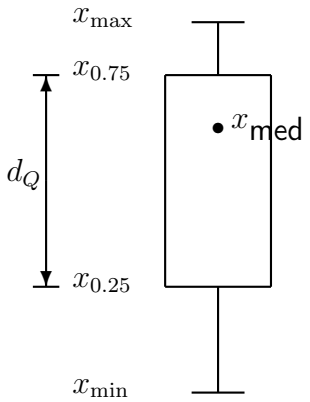
Der Median x_{Med} ist also insbesondere ein $\frac{1}{2}$-Quantil.

Spezialfälle

$x_{0.25} =$ Unteres Quartil $\qquad$ $x_{0.75} =$ Oberes Quartil

Die Distanz $d_Q = x_{0.75} - x_{0.25}$ heißt **Quartilsabstand**.

Aufbau eines zugehörigen **Box-Plots**



Modifikationen

Die Länge der Linien (engl. "whiskers", Barthaare) ober- bzw. unterhalb der Box können variieren. Eine gängige Variation besteht darin, die untere von

$$\max\{x_{0.25} - 1.5 * d_Q, x_{\min}\} \text{ bis } x_{0.25}$$

und die obere von

$$x_{0.75} \text{ bis } \min\{x_{0.75} + 1.5 * d_Q, x_{\max}\}$$

zu führen. Messwerte, die darunter bzw. darüber liegen, können gegebenenfalls als Ausreißer durch einzelne Punkte explizit kenntlich gemacht werden.

Streumaße

Neben der absoluten Lage der Messdaten ist auch ihre Streuung von großer Bedeutung. Die bekannteste Maßzahl für die Streuung einer Messreihe ist die **empirische Varianz** oder auch **mittlere quadratische Abweichung**:

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{j=1}^{s}(a_j - \overline{x})^2 f(a_j). \tag{1.2}$$

Sie ist also definiert als das arithmetische Mittel der quadratischen Abstände der einzelnen Messwerte zu ihrem Mittelwert. Die Wurzel hieraus

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

heißt **Standardabweichung**.

Der Zusammenhang zwischen der Standardabweichung s und der Streuung der Messwerte kann folgendermaßen präzisiert werden:

Für $k \geq 1$ liegen mindestens $100 \cdot \left(1 - \frac{1}{k^2}\right)$ Prozent der Messwerte $x_1, \ldots, x_n$ im Intervall $[\bar{x} - ks, \bar{x} + ks]$. Insbesondere:
im Intervall

- $[\overline{x} - \sqrt{2}s, \overline{x} + \sqrt{2}s]$ liegen mindestens 50 % der Daten
- $[\overline{x} - 2s, \overline{x} + 2s]$ liegen mindestens 75 % der Daten
- $[\overline{x} - 3s, \overline{x} + 3s]$ liegen mindestens 90 % der Daten.

Begründung der Abschätzung: Es reicht zu zeigen, dass

$$H := \text{Anzahl der } x_i \text{ mit } |x_i - \bar{x}| > k \cdot s$$

kleiner gleich $\frac{n}{k^2}$ ist. Zur Abschätzung von H beachte man, dass

$$H = \sum_{i=1}^{n} 1_{\{|x_i - \bar{x}| > k \cdot s\}} \text{ mit } 1_{\{|x_i - \bar{x}| > k \cdot s\}} = \begin{cases} 1 & \text{falls } |x_i - \bar{x}| > k \cdot s \\ 0 & \text{falls } |x_i - \bar{x}| \leq k \cdot s\,. \end{cases}$$

Offensichtlich gilt nun aber

$$\sum_{i=1}^{n} 1_{\{|x_i - \bar{x}| > k \cdot s\}} \leq \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{k \cdot s}\right)^2 = \frac{1}{k^2 \cdot s^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{= n \cdot s^2} = \frac{n}{k^2}\,.$$

Diese Abschätzung ist allgemein gültig und daher in vielen Fällen sehr ungenau. Wir werden später im Zusammenhang mit einem wahrscheinlichkeitstheoretischen Resultat sehen: Ist das Merkmal in etwa normalverteilt, so gilt:

im Intervall

- $[\overline{x} - s, \overline{x} + s]$ liegen etwa 68 % der Daten
- $[\overline{x} - 2s, \overline{x} + 2s]$ liegen etwa 95 % der Daten
- $[\overline{x} - 3s, \overline{x} + 3s]$ liegen etwa 99 % der Daten.

Diese Abschätzung ist also deutlich besser!

Bemerkung

In der induktiven Statistik verwendet man statt (1.2) die modifizierte Form

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

Sie heißt **Stichprobenvarianz** und ist in vielen Statistikprogrammpaketen voreingestellt. Für großen Stichprobenumfang n ist der Unterschied zwischen den beiden Normalisierungsfaktoren $\frac{1}{n}$ und $\frac{1}{n-1}$ vernachlässigbar.
Die Normierung mit $\frac{1}{n-1}$ statt mit $\frac{1}{n}$ liegt darin begründet, dass die Beziehung $\sum_{i=1}^{n} x_i - \overline{x} = 0$ eine der Abweichungen $x_i - \overline{x}$ bereits durch die übrigen $n-1$ eindeutig festlegt. Die Anzahl der Freiheitsgrade in der Summe $\sum_{i=1}^{n} (x_i - \overline{x})^2$ beträgt also $n-1$ und nicht n.

Eigenschaften der empirischen Varianz

(i) **Transformationsregel** Werden die Daten gemäß

$$y_i = a + bx_i$$

linear transformiert, so folgt für die empirische Varianz $s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_1 - \overline{y})^2$ der transformierten Daten

$$s_y^2 = b^2 s_x^2 \,.$$

Beweis

$$s_y^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(y_i - \overline{y})^2}_{(a+bx_i)-(a+b\overline{x})} = b^2 \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \square$$

Insbesondere folgt für die Standardabweichungen:

$$s_y = |b| \, s_x.$$

(ii) **Verschiebungssatz**

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

denn

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(x_i - \overline{x})^2}_{=x_i^2 - 2x_i\overline{x} + \overline{x}^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - 2\frac{1}{n} \sum_{i=1}^{n} x_i \overline{x} + \overline{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 \,.$$

Konzentrationsmaße

Als Ausgangspunkt betrachten wir folgende aus [2] entnommene Statistik zu monatlichen Umsätzen der Möbelbranche in 1000 Euro in den drei Städten G, M und V:

Einrichtungshäuser	G	M	V
1	40	180	60
2	40	5	50
3	40	5	40
4	40	5	30
5	40	5	20

In der Stadt G ist der Umsatz unter den 5 Möbelhäusern also ausgeglichen, während in der Stadt M ein Möbelhaus quasi eine Monopolstellung besitzt. Zur Quantifizierung solcher Konzentrationen gibt es Konzentrationsmaße. Zur Diskussion solcher Maße betrachten wir folgende Ausgangsposition:

Gegeben sei ein kardinalskaliertes Merkmal mit nichtnegativen Merkmalsausprägungen. Weiterhin sei $x_1 \leq x_2 \leq ... \leq x_n$ eine bereits geordnete Stichprobe der Länge n mit positiver Merkmalssumme $\sum_{i=1}^{n} x_i > 0$.

Lorenzkurve

Es sei

$$v_k := \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i} \qquad k = 0, 1, 2, \ldots, n$$

der Anteil der k kleinsten Merkmalsträger an der gesamten Merkmalssumme. Trägt man die Punkte

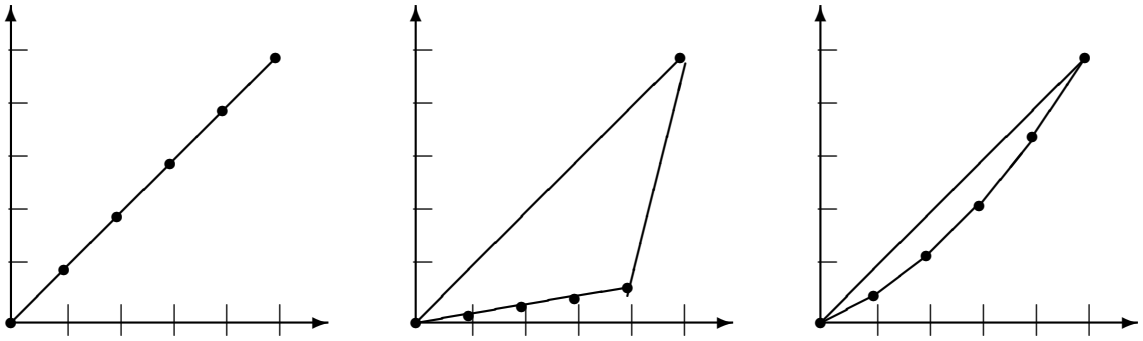
$$\left(\frac{k}{n}, v_k\right), k = 0, 1, 2, \ldots, n$$

in das Einheitsquadrat ein und verbindet sie durch einen Streckenzug, so erhält man die zugehörige **Lorenzkurve**.

In obigem Beispiel erhält man:

	Stadt G	Stadt M	Stadt V
k	v_k	v_k	v_k
1	0.2	0.025	0.10
2	0.4	0.050	0.25
3	0.6	0.075	0.45
4	0.8	0.100	0.70
5	1.0	1.0	1.0

Man erhält als zugehörige **Lorenzkurven**



Stadt G Stadt M Stadt V

Eigenschaften der Lorenzkurve

- Die Lorenzkurve ist immer monoton wachsend und konvex (d.h. nach unten gewölbt).
- Die Stärke der Wölbung, also ihre Abweichung von der Winkelhalbierenden, ist ein Maß für Konzentration. Verläuft die Kurve auf der Winkelhalbierenden, so liegt ein ausgewogener Markt vor.

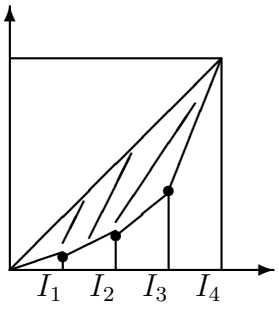
Der **Gini-Koeffizient** G ist definiert durch

$$\begin{aligned} G &= \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und horizontaler Achse}} \\ &= 2 \cdot \text{ Fläche zwischen Diagonale und Lorenzkurve} \end{aligned}$$

Für die Berechnung des Gini-Koeffizienten gilt die folgende Formel:

$$G = \frac{2}{n} \frac{\sum_{i=1}^{n} i x_i}{\sum_{i=1}^{n} x_i} - \frac{n+1}{n}.$$

Beweis



Die Fläche der I_i beträgt gerade

$$I_i = \frac{1}{n} v_{i-1} + \frac{1}{2n}(v_i - v_{i-1})$$

also summiert sich die Gesamtfläche der I_i zu

$$\frac{1}{n}\sum_{i=1}^{n} v_{i-1} + \frac{1}{2n}\underbrace{\sum_{i=1}^{n}(v_i - v_{i-1})}_{=v_n - v_0 = 1} = \frac{1}{n}\sum_{i=1}^{n-1} v_i + \frac{1}{2n}\,.$$

Beachtet man noch, dass

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n-1} v_i &= \frac{1}{n}\frac{1}{\sum_{j=1}^{n} x_j}\left(\sum_{i=1}^{n-1}\sum_{k=1}^{i} x_k\right)\\ &= \frac{1}{n}\frac{1}{\sum_{j=1}^{n} x_j}\sum_{k=1}^{n}(n-k)x_k = 1 - \frac{1}{n}\frac{\sum_{k=1}^{n} k x_k}{\sum_{j=1}^{n} x_j}\end{aligned}$$

so erhält man nach Einsetzen in die obere Gleichung

$$G = 2\left(\frac{1}{2} - \left(1 - \frac{1}{n}\frac{\sum_{j=1}^{n} j x_j}{\sum_{j=1}^{n} x_j} + \frac{1}{2n}\right)\right) = \frac{2}{n}\frac{\sum_{j=1}^{n} j x_j}{\sum_{j=1}^{n} x_j} - \frac{n+1}{n}\,. \quad \square$$

3. Auswertung zwei- und mehrdimensionaler Messreihen

Zweidimensionale Messreihen

Werden bei einer Erhebung zwei Merkmale X und Y zugleich erhoben, so besteht die Urliste aus **Wertepaaren**

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Typische Fragestellungen im Zusammenhang zweier Merkmale sind die nach Abhängigkeiten/Unabhängigkeiten zwischen den beiden erhobenen Merkmalen. Zur Darstellung der zweidimensionalen Daten gibt es zunächst zwei Möglichkeiten:

- **Kontingenztabelle:** geeignet für nominalskalierte Merkmale
- **Streuungsdiagramm:** geeignet für kardinalskalierte Merkmale

(A) Kontingenztabelle

Bei diesem Verfahren werden die absoluten Häufigkeiten der möglichen Paare von Ausprägungen des Merkmals x und des Merkmals y tabellarisch aufgelistet:

	Ausprägungen von Y		
Ausprägungen von X	b_1	$\ldots$	b_l
a_1	h_{11}	$\ldots$	h_{1l}
$\vdots$	$\vdots$		$\vdots$
a_k	h_{k1}	$\ldots$	h_{kl}

Hierbei steht $h_{ij} = h(a_i, b_j)$ für die absolute Häufigkeit der Wertepaare (a_i, b_j).

Beispiel (entnommen aus [1])

Zur Untersuchung von Abhängigkeiten zwischen Berufsgruppen und sportlicher Betätigung werden 1000 Personen befragt. Es entstand dabei folgende **Kontingenztabelle:**

	sportl. Bet.		
	nie	gelegentlich	regelmäßig
Arbeiter	240	120	70
Angestellter	160	90	90
Beamter	30	30	30
Landwirt	37	7	6
sonst. freier Beruf	40	32	18

Die Einträge in der Kontigenztabelle heißen **gemeinsame Häufigkeiten**. Statt der absoluten, lassen sich hier natürlich auch die relativen Häufigkeiten betrachten:

$$f_{ij} = f(a_i, b_j) = \frac{h_{ij}}{n}.$$

Fragt man nach der absoluten Häufigkeit einer Merkmalsausprägung a_i (bzw.b_j) so hat man die gemeinsamen Häufigkeiten h_{ij} der entsprechenden Zeile (bzw. der entsprechenden Spalte) aufzusummieren:

$$h(a_i) = h_{i\cdot} := \sum_{j=1}^{l} h_{ij}$$

$$h(b_j) = h_{\cdot j} := \sum_{i=1}^{k} h_{ij}$$

Diese Häufigkeiten werden auch als **Randhäufigkeiten** bezeichnet.

In obigem Beispiel

	sportl. Bet.			
	nie	gelegentlich	regelmäßig	Randhäufigkeiten
Arbeiter	s.o.	s.o.	s.o.	430
Angesteller				340
Beamter				90
Landwirt				50
sonst. freier Beruf				90
Randhäufigkeiten	507	279	214	1000

Um nun die beiden Merkmale auf Abhängigkeit/Unabhängigkeit hin zu untersuchen, bildet man die **bedingten relativen Häufigkeiten**

$$f_X(a_i|b_j) := \frac{h_{ij}}{h_{\cdot j}} \text{ der Ausprägung } a_i \text{ gegeben die Ausprägung } b_j$$

und

$$f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i\cdot}} \text{ der Ausprägung } b_j \text{ gegeben die Ausprägung } a_i\,.$$

Die bedingte relative Häufigkeit $f_X(a_i|b_j)$ gibt also die relative Häufigkeit der Ausprägung a_i an unter allen Merkmalsträgern, die bzgl. des anderen Merkmals die Ausprägung b_j besitzen. Sind die bedingten relativen Häufigkeiten

$$f_X(a_1|b_j), f_X(a_2|b_j), ..., f_X(a_k|b_j)$$

der Ausprägung $a_1, ..., a_k$ des ersten Merkmals unabhängig von b_j (also gleich für $j = 1, \ldots, l$), so beeinflussen sich die Merkmale nicht und man sagt, dass sie **unabhängig** sind.

Dieser Fall tritt genau dann ein, wenn auch die umgekehrten bedingten relativen Häufigkeiten

$$f_Y(b_1|a_i), f_Y(b_2|a_i), ..., f_Y(b_l|a_i)$$

unabhängig sind von a_i für $i = 1, \ldots, k$.

Im Falle der Unabhängigkeit gilt insbesondere

$$f_X(a_i|b_{j1}) = f_X(a_i|b_{j2})$$

und damit

$$h_{ij_1} \cdot h_{\cdot j_2} = h_{ij_2} \cdot h_{\cdot j_1}$$

Summation über $j_1 = 1, \ldots, l$ ergibt

$$h_{i\cdot} h_{\cdot j_2} = h_{ij_2} \cdot n$$

also

$$h_{ij_2} = \frac{h_{i\cdot} h_{\cdot j_2}}{n}$$

und somit - da j_2 beliebig:

$$h_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}\,. \tag{1.3}$$

Die **gemeinsamen Häufigkeiten** sind in diesem Falle über (1.3) also bereits durch die **Randhäufigkeiten** bestimmt.

Für die bedingten relativen Häufigkeiten folgt hieraus insbesondere

$$f_X(a_i|b_j) = \frac{h_{ij}}{h_{\cdot j}} = \frac{h_{i\cdot}}{n} \qquad \text{bzw. } f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n}\,,$$

sie sind also unabhängig von der Ausprägung des jeweils anderen Merkmals.

Der Kontingenzkoeffizient

Um die Abhängigkeit zwischen zwei Merkmalen X und Y quantitativ erfassen zu können, bildet man die folgende, als **Chi-Quadrat Koeffizient**, bezeichnete Größe:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}.$$

Hierbei ist $\tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$.

χ^2 ist genau dann 0, wenn die Merkmale unabhängig sind, also wenn $h_{ij} = \tilde{h}_{ij}$ gilt. Je kleiner also der χ^2-Koeffizient, umso stärker spricht dies für die Unabhängigkeit der beiden Merkmale X und Y. Allerdings hängt die Größenordnung des χ^2-Koeffizienten von der Dimension der Kontingenztafel ab. Daher geht man vom χ^2-Koeffizienten über zum **Kontingenzkoeffizienten**

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

Der Kontingenzkoeffizient K nimmt Werte an zwischen 0 und

$$K_{max} = \sqrt{\frac{M-1}{M}}, \text{ wobei } M = \min\{k, l\}\,.$$

Durch Normierung mit K_{max} erhält man hieraus schließlich den **normierten Kontingenzkoeffizienten**

$$K_* = \frac{K}{K_{max}}.$$

Beispiel (obiges Beispiel zum Zusammenhang zwischen Berufstätigkeit und sportlicher Betätigung)

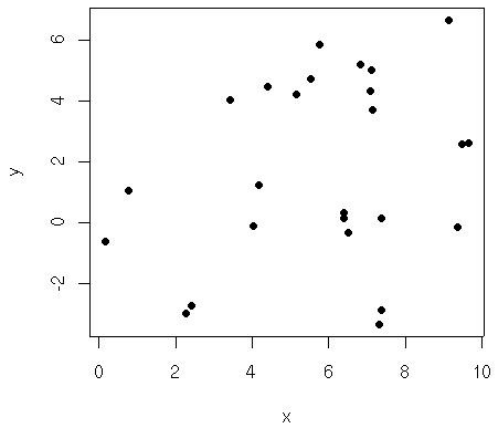
In diesem Falle ist $\chi^2 = 38.55412$ und wegen $n = 1000$ folgt für den Kontingenzkoeffizienten $K = 0.192673$ sowie wegen $k = 5$, $l = 3$, also $M = \min\{k, l\} = 3$, folgt für den normierten Kontingenzkoeffizienten $K_* = 0.2359753$.

(B) Streuungsdiagramm

Bei kardinalskalierten Merkmalen kann man die Wertepaare

$$(x_1, y_1), \ldots, (x_n, y_n)$$

der Urliste als Punkte der Ebene auffassen und somit ein zugehöriges **Streuungsdiagramm** erstellen:



Beispiel

In einem Krankenhaus wurden von 5 Neugeborenen Körperlänge X und Kopfumfang Y (in cm) gemessen. Es ergab sich folgende nach Köperlänge geordnete Messreihe:

(48.6, 35.1), (49.5, 34.1), (50.7, 36.8), (51.1, 35.7), (52.4, 37.4)

Zu den jeweiligen Messwerten bildet man zunächst die beiden Mittelwerte

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Im Beispiel $\overline{x} = \frac{1}{5}$ 252.3 = 50.46, $\overline{y} = \frac{1}{5}$ 179.1 = 35.82.

Liegt bei einem Wertepaar (x_i, y_i) der erste Wert um den Durchschnitt $x_i \sim \overline{x}$, aber der zweite Wert y_i deutlich über oder unter dem Durchschnitt $\overline{y}$, so spricht dies eher

für die Unkorreliertheit der beiden Merkmale Körperlänge X und Kopfumfang Y. Liegen jedoch bei diesem Wertepaar bei beiden Merkmalen deutliche Abweichungen vom Durchschnitt vor, so spricht dies für Korrelation. Folglich liefert das Produkt

$$(x_i - \overline{x})(y_i - \overline{y})$$

einen brauchbaren Ansatz für ein Korrelationsmaß.

Aufsummieren über die gesamte Stichprobe und Normierung ergibt die **empirische Kovarianz**

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

Nach Normierung mit den jeweiligen Standardabweichungen

$$s_X = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{\frac{1}{2}} \text{ und } s_Y = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{\frac{1}{2}}$$

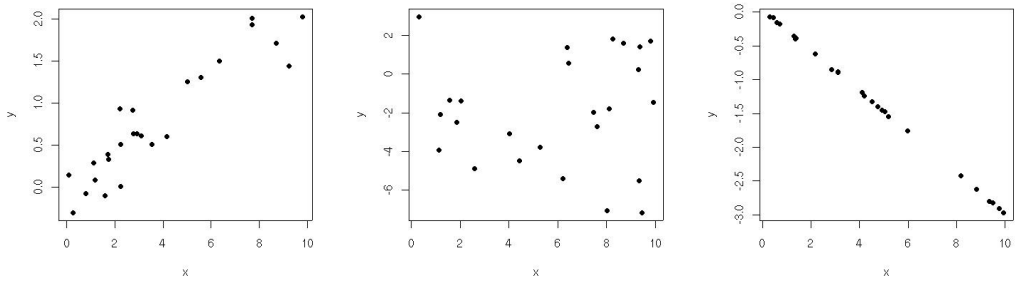
erhält man den **empirischen Korrelationskoeffizienten**

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

Eigenschaften

- $-1 \leq r_{XY} \leq 1$
- $r_{XY} = -1$ (bzw. $r_{XY} = +1$) genau dann wenn die Wertepaare (x_i, y_i) auf einer Geraden mit negativer (bzw. positiver) Steigung liegen.
- $r_{XY} = 0$ spricht für die Unkorreliertheit der Merkmale X und Y. In diesem Falle sind die Wertepaare (x_i, y_i) "regellos" verteilt.
- Die Merkmale X und Y heißen
  - **positiv korreliert**, falls $r_{XY} > 0$
  - **negativ korreliert**, falls $r_{XY} < 0$.

$r_{XY} = 0.827$ $\qquad$ $r_{XY} = 0.046$ $\qquad$ $r_{XY} = -0.999$



- eine rechentechnisch günstigere Darstellung für den Korrelationskoeffizienten ist

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2)}}\,.$$

Regressionsrechnung

Liegen die Wertepaare der n Beobachtungen (x_i, y_i) annähernd auf einer Geraden, so kann man von einem **linearen Zusammenhang** der Form

$$y = a + bx \tag{1.4}$$

sprechen. Die Koeffizienten a und b wählt man dabei so, dass sich die zugehörige Gerade der gegebenen Punktwolke am besten anpasst. "Beste Anpassung" bedeutet dabei, dass die Summe der quadratischen Abstände

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \,,$$

zwischen Messwert y_i und entsprechendem Punkt $a + bx_i$ auf der Geraden $y = a + bx$, minimal wird. ("Prinzip der kleinsten Quadrate" nach C.F. Gauß).

Diejenige Gerade, die sich der Punktwolke dabei am besten anpasst, heißt **Ausgleichsgerade** oder **Regressionsgerade**. Ihre Koeffizienten sind bestimmt durch

$$\hat{b} = \frac{s_{XY}}{s_X^2} \,, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \,. \tag{1.5}$$

Beispiel In obigem Beispiel ist

$$s_{XY} = \frac{1}{4}(9043.6 - 9037.386) \sim 1.55$$

und damit $r_{XY} \sim 0.8$ (d. h. Körpergröße und Kopfumfang sind (erwartungsgemäß) stark positiv korreliert). Die Koeffizienten der zugehörigen **Regressionsgeraden** sind gegeben durch

$$\hat{b} \sim 0.72 \text{ und } \hat{a} \sim -0.51$$

also hat die Regressionsgerade die Form

$$y = -0.51 + 0.72x \,.$$

Mit Hilfe der Regressionsgeraden können wir nun zum Beispiel einen Vorhersagewert ("Prognose") für den Kopfumfang eines Neugeborenen bei einer Körperlänge von 50 cm bestimmen: $y(50) = 35.49$.

Zu gegebenem Wertepaar (x_i, y_i) heißt die Differenz

$$u_i := y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i)$$

zwischen beobachtetem Wert y_i und dem durch die Regressionsgerade erklärten entsprechenden Wert $\hat{y}_i = \hat{a} + \hat{b}x_i$ **Residuum**. Den Quotienten

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{XY}^2$$

bezeichnet man als **Bestimmtheitsmaß**. Er ist ein Maß für die Güte der Approximation der Messwerte y_i durch die berechnete Ausgleichsgerade und stimmt mit dem Quadrat des Korrelationskoeffizienten überein.

Zur **Optimalität** der **Regressionsgeraden**

Satz Es sei $s_X^2 \neq 0$ und $\hat{a}$, $\hat{b}$ wie in (1.5). Dann gilt:

$$Q(a,b) > Q(\hat{a},\hat{b}) \qquad \text{für alle } (a,b) \neq (\hat{a},\hat{b})\,.$$

Beweis:

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

ist ein Polynom vom Grad 2 mit Gradient

$$\begin{aligned} \text{grad } Q(a,b) &= \left(\frac{\partial Q}{\partial a}(a,b), \frac{\partial Q}{\partial b}(a,b) \right) \\ &= -2 \left(\sum_{i=1}^{n} [y_i - (a+bx_i)], \sum_{i=1}^{n} x_i [y_i - (a+bx_i)] \right) \end{aligned}$$

und Hesse-Matrix

$$H_Q(a,b) = \begin{bmatrix} \frac{\partial^2 Q}{\partial a^2}(a,b) & \frac{\partial^2 Q}{\partial a \partial b}(a,b) \\ \frac{\partial^2 Q}{\partial a \partial b}(a,b) & \frac{\partial^2 Q}{\partial b^2}(a,b) \end{bmatrix} = 2 \begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

Also

$$\det H_Q(a,b) = 4 \left(n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2 \right) = 4n^2 s_X^2 > 0\,,$$

damit ist H_Q positiv definit und somit Q gleichmäßig strikt konvex.

Folglich besitzt Q genau ein eindeutig bestimmtes Minimum und dies wird an der "Nullstelle" (bzw. der kritischen Stelle) des Gradienten angenommen:

$$\begin{aligned} \text{grad } Q(a,b) = 0 &\Leftrightarrow \frac{\partial Q}{\partial a}(a,b) = 0 \text{ und } \frac{\partial Q}{\partial b}(a,b) = 0 \\ &\Leftrightarrow \overline{y} = a + b\overline{x} \text{ und} \\ &\qquad 0 = \sum_{i=1}^{n} x_i (y_i - (a+bx_i)) = \sum_{i=1}^{n} x_i (y_i - bx_i - (\overline{y} - b\overline{x})) \\ &\qquad = \sum_{i=1}^{n} x_i y_i - b \sum_{i=1}^{n} x_i^2 - n\overline{x}\overline{y} + nb\overline{x}^2 \\ &\Leftrightarrow a = \overline{y} - b\overline{x} \text{ und} \\ &\qquad b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{s_{XY}}{s_X^2} \quad \square \end{aligned}$$

Bemerkung (Nichtlineare Regression)

Bei vielen zweidimensionalen Messreihen ist von vorneherein klar, dass kein linearer Zusammenhang zwischen den beobachteten Messwerten erwartet werden kann, sondern ein funktionaler Zusammenhang der Form

$$y = f(x)$$

für eine geeignete **nichtlineare** Funktion f, z.B.

$$y = ae^{bx} \text{ für } b \in \mathbb{R}, a > 0.$$

Gesucht sind wieder diejenigen Parameter a und b, für die sich der zugehörige Funktionsgraph der gegebenen Punktwolke am besten anpasst. Häufig kann man durch geeignete Transformation der Daten das Problem auf einen linearen Zusammenhang zurückführen, wie etwa im Beispiel $y = ae^{bx}$

$$\log y = \log a + bx$$

und zu bestimmen ist die Regressionsgerade zu den transformierten Beobachtungswerten

$$(x_1, \log y_1), (x_2, \log y_2), ..., (x_n, \log y_n)\,.$$

Ausblick auf mehrdimensionale Messreihen

Bei einer statistischen Erhebung können natürlich mehr als zwei Merkmale zugleich erhoben werden. Als Urliste enstehen Tupel (d.h. geordnete Mengen) von Messwerten

$$(x_{11}, \ldots, x_{1m})\,, (x_{21}, \ldots, x_{2m})\,, \ldots (x_{n1}, \ldots, x_{nm})\,,$$

die man in einer **Datenmatrix** zusammenfasst:

$$\begin{bmatrix} x_{11} & \ldots & x_{1m} \\ x_{21} & \ldots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nm} \end{bmatrix}$$

Die graphische Darstellung der Urliste als Streuungsdiagramm ist für $m \geq 4$ nicht mehr möglich. Zur Aufklärung von Abhängigkeiten zwischen den erhobenen Merkmalen könnte man zwar für jedes Paar von Merkmalen das zweidimensionale Streuungsdiagramm bzw. die zweidimensionale Kontingenztabelle aufstellen. Da aber die Anzahl der Merkmalspaare mit der Anzahl m der erhobenen Merkmale sehr schnell anwächst, ist dieser Ansatz sehr aufwändig. Effizientere Methoden sind Gegenstand weiterführender Veranstaltungen in der Statistik.

Literatur

[1] G. Bamberg, F. Baur, M. Krapp, Statistik, 13. Auflage, R. Oldenbourg Verlag, 2007.

[2] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, Statistik, 6. Auflage, Springer Verlag, 2007.

Weitere Literatur

[3] J. Bleymüller, G. Gehlert, H. Gülicher, Statistik für Wirtschaftswissenschaftler, 14. Auflage, Verlag Vahlen, 2004.

[4] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, A. Caputo, S. Lang, Arbeitsbuch Statistik, 4. Auflage, Springer Verlag, 2004.

[5] J. Schira, Statistische Methoden der VWL und BWL: Theorie und Praxis, 2. Auflage, Pearson Studium, 2005.