Statistik

Vorlesungsmitschrift - Kurzfassung

(erstellt von C.Mongin, überarbeitet von M.Dittlinger)

Prof. Dr. rer. nat. B. Grabowski

LITERATUR

1. ***Dieses*** (vorlesungsbegleitende) ***Skript*** in den Teilen
   I - Deskriptive Statistik,
   II - Wahrscheinlichkeitsrechnung,
   III- Schließende Statistik,
   IV- Ausgewählte Gebiete
   V - Anhang (A-Tabellen, B-Klausuren, C-Lösungen zu den Übungsaufgaben)

2. J.Schwarze: " Grundlagen der Statistik I - Beschreibende Verfahren",
   " Grundlagen der Statistik II - Wahrscheinlichkeitsrechnung und induktive Statistik"
   " Aufgabensammlung zur Statistik"
   alle 3 Bände erschienen bei
   NWB-Studienbücher -Wirtschaftswissenschaften, Verlag Neue Wirtschaftsbriefe - Herne/Berlin, 1990

3. L. Papula : "Mathematik für Ingenieure", Band 3 (Kapitel Wahrscheinlichkeitsrechnung und Statistik)

4. Tinhofer : „Stochastik für Informatiker“

I. Deskriptive Statistik

Vorlesungsmitschrift - Kurzfassung

(erstellt von C.Mongin, überarbeitet von M.Dittlinger)

Prof. Dr. rer. nat. B. Grabowski

INHALTSVERZEICHNIS

I. Deskriptive Statistik

I. Deskriptive Statistik

1 Grundbegriffe

1.1 Gegenstand der Statistik - Grundgesamtheit und Stichprobe

***Statistik**:* = Wissenschaftliche Methoden zur Untersuchung und Beurteilung zufälliger Merkmale an irgendwelchen Objekten

Beispiel 1: Wir wollen wissen, wie sich die Saarländer bei der nächsten Landtagswahl verhalten werden.
Merkmal: Wahlverhalten
Objekt: 1 Saarländer

***Zufälliges Merkmal*:** Nicht eindeutig voraussagbar!

Das Interesse der Statistik richtet sich immer auf eine Menge (bzgl. des interessierenden Merkmals) „gleich-artiger" Objekte, die sogenannte Grundgesamtheit. (Menge der Saarländer, die wahlberechtigt sind)

Grundgesamtheit: = Gruppe von Objekten, die hinsichtlich eines Untersuchungsziels als gleichartig angesehen werden.

Merkmal	Objekt	Grundgesamtheit
Lebensdauer	Glühlampe	Glühlampen eines bestimmten Werkes
Zuverlässigkeit	Maschine	Maschinen eines bestimmten Typs
Anzahl verkaufter pro Tag	Tag	alle Tage eines Jahres
Wahlverhalten	Saarländer	alle wahlberechtigten Saarländer

Tabelle 1.1

Ziel statistischer Untersuchungen:

Aussagen über die Häufigkeit des Auftretens bestimmter Merkmalswerte in der Grundgesamtheit (z.B. wieviele wahlberechtigte Saarländer wählen CDU, SPD, ...).

Zur Beurteilung des Merkmals Wahlverhalten, kann man nicht alle Saarländer befragen, sondern nur einen Teil!

Man macht eine Stichprobe!

Stichprobe: = Teilmenge von Objekten der Grundgesamtheit. Die Auswahl erfolgt „zufällig".

Von der Verteilung der Merkmalswerte in der Stichprobe wird dann auf die Verteilung der Merkmalswerte in der Grundgesamtheit ***geschlossen***.

Grundfragen der Statistik:

- *Wie bestimmt man die Stichprobe?* (Stichprobenumfang)
- *Wie analysiert man die Stichprobe?* (Verfahren)
- *Wie schließt man von der Stichprobe auf die Grundgesamtheit?*
- *Wie groß ist der Fehler (Irrtumswahrscheinlichkeit), den man bei diesem Schluß macht?*

Teilgebiete der Statistik:

- ***Beschreibende (Deskriptive) Statistik****:*= Empirische und insbesondere grafische Methoden zur Analyse von Stichproben

- ***Schließende (Induktive) Statistik****:* = Mathematische Methoden des Schließens von der Stichprobe auf die Grundgesamtheit mit Angaben von Irrtums-wahrscheinlichkeiten

- ***Wahrscheinlichkeitsrechnung****:*= Mathematische Methoden zur Beschreibung des Zufalls. Mathematische Grundlage der Schließenden Statistik

DESKRIPTIVE STATISTIK ⟶ SCHLIEßENDE STATISTIK
WARSCHEINLICHKEITSRECHNUNG

Beispiel 2: 1. Untersuchung des Wahlverhaltens der Saarländer:
Was wäre, wenn nächsten Sonntag Landtagswahlen wären?

	Werten Sie die Wahlergebnisse der Stichprobe anhand nachfolgender Abbildung aus! Schließen Sie auf die Grundgesamtheit aller wahlberechtigten Saarländer!

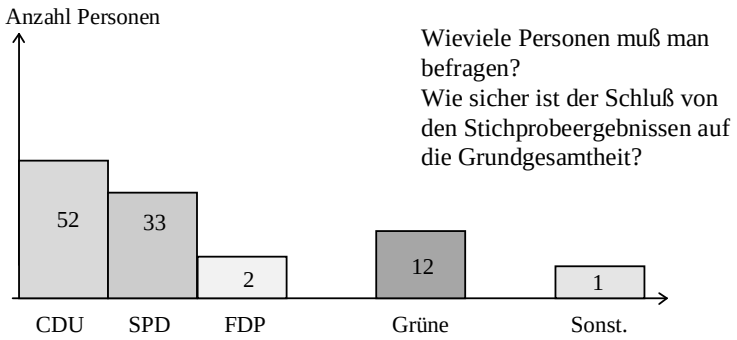


Abb. 1.1: Verdichtung durch Häufigkeitstabellen

Die Schließende Statistik liefert Aussagen der Form:
'Mit 99% iger Sicherheit wählen zwischen 50% und 54% aller wahlberechtigten Saarländer die CDU.'

2. Untersuchung der Wirksamkeit zweier Medikamente M1, M2

Werten Sie die folgende Tabelle aus! Welches Medikament ist in den gegebenen Stichproben besser?
Schließen Sie auf die Grundgesamtheit aller Patienten!

	verbessert	**nicht verbessert**
M1	98	22
M2	55	11

Tabelle 1.2

Mit welchem Maß kann man die Wirksamkeit der beiden Medikamente vergleichen?
Wieviele Personen muß man untersuchen?
Wie sicher sind diese Ergebnisse?

Die Schließende Statistik liefert Aussagen der Form:

'Mit 99%iger Sicherheit ist M2 wirksamer als M1.'

3. Untersuchung des Zusammenhangs zwischen Bremsgeschwindigkeit X und Bremsweg Y einer bestimmten Automarke:

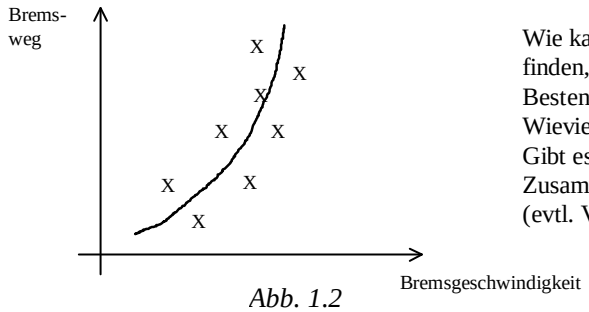


Abb. 1.2

Wie kann ich die Funktion finden, die die Daten 'am Besten' anpaßt?
Wieviele Autos muß ich testen?
Gibt es einen funktionalen Zusammenhang?
(evtl. Vorhersagen möglich?)

Die Schließende Statistik liefert Aussagen der folgenden Form:

' Wenn es einen funktionalen Zusammenhang y = f(x) gibt, dann liegt er mit 99% iger Sicherheit in einem Toleranzband $\{[f_u(x),f_o(x)], x \in R\}$.'

1.2 Merkmale, Merkmalsausprägungen, Skalen

Merkmale: X, Y, ... Merkmalsausprägungen: x, y, ...

werden am Objekt beobachtet:

X(Objekt) = x - beobachtete Merkmalsausprägung

X = Alter, Y = Geschlecht, Z = Beruf
Alter (Susanne) = 28,
Geschlecht (Susanne) = weiblich,
Beruf (Susanne) = Krankenschwester

Wir wollen „rechnen“ $\Rightarrow$ wir ordnen den Merkmalswerten Zahlen zu.

Skalierung:= Zuordnung von Merkmalsausprägungen zu Zahlenwerten

Wir unterscheiden 4 Skalentypen:

Nominalskala: Alle Skalenwerte stehen gleichberechtigt nebeneinander. Es gibt keine Ordnung, nur Unterschiede. Die Größe und die Ordnung der Zahlen spielen keine Rolle, d.h. sie dürfen nicht in statistische Auswertungen einbezogen werden. Alle Nominalskalen, die die in den Merkmalswerten enthaltenen Unterschiede und Gleichheiten abbilden, sind zueinander äquivalent.

Ordinalskala: Es gibt eine Ordnung und Unterschiede. Die Größe der Zahlen spielt keine Rolle. Alle Ordinalskalen, die die in den Merkmalswerten enthaltene Ordnungsrelation abbilden, sind zueinander äquivalent.

Intervallskala: Unterschiede, Ordnungen und Differenzen der Zahlen sind relevant. Differenzen sind bis auf die Maßeinheit eindeutig. Der Nullpunkt ist nicht eindeutig. Zahlenverhältnisse dürfen deshalb nicht in statistische Auswertungen einbezogen werden. Intervallskalen S2, die sich aus der Transformation S2 =aS1 + b aus einer gegebenen Intervallskala S1 ergeben, sind zu S1 äquivalent.

Proportionalitäts- (Verhältnis-) skala:
Unterschiede, Ordnungen, Differenzen und Verhältnisse sind relevant. Differenzen sind bis auf die Maßeinheit eindeutig, Verhältnisse sind eindeutig. Alle Proportionalitätdskalen S2, die sich aus der Transformation S2 =aS1 aus einer gegebenen Intervallskala S1 ergeben, sind zu S1 äquivalent.

Merkmal	Merkmalsausprägung	Skalen				Skalentyp
Geschlecht	weiblich	0	2	~~1~~		Nominal
	männlich	0	1	~~1~~		
Leistungen	sehr gut	1	20	2000	~~2~~	
	gut	2	17	100	~~6~~	
	befriedigend	3	15	10	~~3~~	Ordinal
	ausreichend	4	11	9	~~5~~	
	mangelhaft	5	8	7	~~8~~	
	ungenügend	6	6	3	~~4~~	
Temperatur		^{o}C, K, F alle Skalen, die durch Umrechnung S = a ^{o}C + b entstehen, sind erlaubt				Intervall
Lebensdauer		sek., min., Stunde, ... alle Skalen, die durch Umrechnung S = a * sek entstehen, sind erlaubt				Proportional

Tab. 1.3

Merkmale, deren Zahl möglicher Ausprägungen endlich oder abzählbar sind, werden in der Regel durch eine Ordinal- oder Nominalskala abgebildet.
Merkmale mit unendlich vielen verschiedenen Ausprägungen werden durch eine Intervall- oder Proportionalitätsskala abgebildet. Man nennt solche Merkmale auch *metrisch* skaliert.
Eine Ausnahme bilden das zufällige Merkmal: X=***Anzahl von*** ....(irgendetwas) und das zufällige Merkmal X=***Plazierung***.
Diese werden als ***proportionalitätsskaliert*** betrachtet (Spezialfall).

Def.: Unter einer *zufälligen Variable (Zufallsvariable, Zufallsgröße) X* verstehen wir eine Variable, die einem zufälligen Merkmal zugeordnet wird und deren Zahlenwerte den Merkmalsausprägungen entsprechen.

	Skalierung		
Merkmal		Zufallsvariable	X, Y, Z
Merkmalsausprägung	⟶	Zahlenwerte	x, y, z
Menge der Merkmalsausprägungen		Wertebereich	X, Y, Z

Beispiel 3: X - Geschlecht, $X = \{0,1\}$

Y - Leistung, $Y = \{1,2,3,4,5\}$

T - Lebensdauer, $Z = [0,\infty)$

Def.: Eine zufällige Variable X, deren Wertebereich nur endlich oder abzählbar viele Werte annehmen kann, heißt *diskrete Zufallsvariable* (X-Geschlecht, Y-Leistungen).
Eine zufällige Variable X, deren Wertebereich ein zusammenhängendes Intervall [a,b], a<b enthält, heißt *stetige Zufallsvariable* (Bsp. Temperatur).

X - diskret, falls $\mathbf{X = \{a_1, a_2, \ldots, a_k, \ldots\}}$

X - stetig, falls X ein kontinuierliches Intervall [a,b] enthält.

Beispiel 4: X-Geschlecht $\mathbf{X} = \{0,1\}$ diskret

L-Lebensdauer $\mathbf{X} = [0, \infty)$ stetig

1.3 Häufigkeitsverteilung diskreter Zufallsvariablen

***Bezeichnungen**:*

X - Zufallsvariable (Merkmal)

$\mathbf{X} = \{a_1, \ldots, a_k\}$ – Wertebereich von X (a_i - Merkmalsausprägung)

$x_1, \ldots, x_n$ – n Beobachtungen von X (Stichprobe vom Umfang n) Urliste

elementare statistische Tätigkeiten:

Auszählen, wieviele Beobachtungen auf einen bestimmten Merkmalswert fallen.

Beispiel 5: Untersuchung des Rauchverhaltens männlicher Personen zwischen 40 - 60 Jahren in der BRD

X - Rauchverhalten

$$\mathbf{X} = \begin{cases} 0 & \text{Nichtraucher} \\ 1 & \text{höchstens 10 Zigaretten pro Tag} \\ 2 & \text{mehr als 10 aber höchstens 20 pro Tag} \\ 3 & \text{mehr als 20 pro Tag} \end{cases}$$

Stichprobe von 100 männlichen Personen:

$x_1, x_2, \ldots, x_n$ - Stichprobe vom Umfang n = 100

0,1,3,0,0,...,1 - Urliste

a_i	$H_n(a_i)$ absolute Anzahl	$h_n(a_i)$ relative Anzahl	$H_{(i)}$ absol. Sum-menhäufigkeit	$h_{(i)}$ relative Summenhäufigkeit
0	40	0.4	40	0,.4
1	15	0.15	55	0.55
2	40	0.4	95	0.95
3	5	0.05	100	1.0
Σ	100	1.00		

Tab. 1.4: Statistische Häufigkeitstabelle

Def.: Die Aneinanderreihung von beobachteten Merkmalen heißt *Urliste*.

Def.: Die *Anzahl* $H_n(a_i)$ aller x_j mit $x_j = a_i$, j = 1,...,n heißt **absolute Häufigkeit von a_i**
Der *Anteil* $h_n(a_i)$ aller x_j mit $x_j = a_i$, j = 1,...,n heißt **relative Häufigkeit von a_i**

$H_n(a_i)$ = absolute Häufigkeit von a_i.

$h_n(a_i) = \frac{H_n(a_i)}{n}$ - relative Häufigkeit von a_i.

Die *Gesamtheit* aller $H_n(a_i)$, i=1,...,k (bzw. $h_n(a_i)$, i=1,...,k) heißt *absolute* bzw. r*elative Häufigkeitsverteilung* von X bzgl. der Stichprobe $x_1,...,x_n$.

Es gilt: **$0 \le H_n(a_i) \le n$** $\quad\quad \sum_{i=1}^{k} H_n(a_i) = n$

$0 \le h_n(a_i) \le 1$ $\quad\quad \sum_{i=1}^{k} h_n(a_i) = 1$

Typische grafische Darstellungen der Häufigkeitsverteilungen:

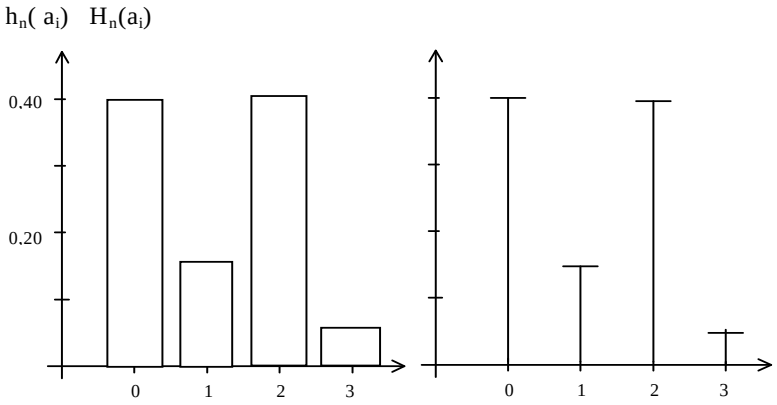


Abb. 1.3: Balkendiagramm *Abb. 1.4:* Stabdiagramm

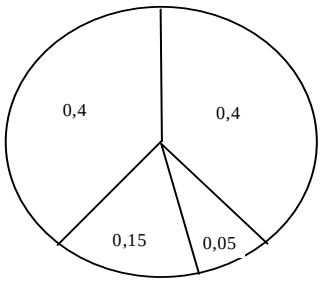


Abb. 1.6: Tortendiagramm

Def.: $H(i) = \sum_{j=1}^{i} H_n(a_j)$ i=1,...,k - **absolute Summenhäufigkeit**

$h(i) = \sum_{j=1}^{i} h_n(a_j)$ i=1,...,k - **relative Summenhäufigkeit**

H(i) – *Anzahl* der Objekte (Stichprobenwerte), die einen Merkmalswert $\leq a_i$ besitzen

h(i) – *Anteil* der Objekte (Stichprobenwerte), die einen Merkmalswert $\leq a_i$ besitzen

Mit Hilfe der relativen Summenhäufigkeit läßt sich die sogenannte empirische Verteilungsfunktion (Summenhäufigkeitsfunktion) definieren:

Def.: Empirische Verteilungsfunktion (Summenhäufigkeitsfunktion):

$$F_n(x) = \begin{cases} 0 & \text{für } x < a_1 \\ h(i) & \text{für } a_i \leq x < a_{i+1} \\ 1 & \text{für } x \geq a_k \end{cases} \quad , i = 1,...,k-1$$

$F_n(x)$ – *Anteil der Objekte* in der Stichprobe mit einem Merkmalswert $\leq x$

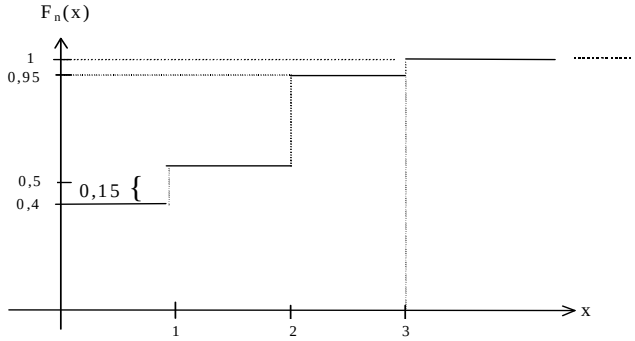


Abb.: 1.6: Empirische Verteilungsfunktion zu Bsp. 5

Der Anteil der Personen, die nicht mehr als 20 Zigaretten pro Tag rauchen, beträgt 95 %, die Anzahl ist 95.

Das Bild der empirischen Verteilungsfunktion ist eine Treppenfunktion. In jedem Merkmalswert a_i springt die Funktion um $h_n(a_i)$.

Aus der Summenhäufigkeit sind umgekehrt wieder die einzelnen Häufigkeiten berechenbar.

Es gilt: $H_n(a_i) = H(i) - H(i-1)$

$h_n(a_i) = h(i) - h(i-1)$ i=1,...,k

mit $h(0) = H(0) = 0$

1.4 Häufigkeitsverteilung klassifizierter Daten

Sei X ein stetiges Merkmal, d.h. der Wertebereich umfaßt unendlich viele Werte: $\exists$ [a,b,] $\subseteq$ X. I.a. werden dann bei einer Stichprobe $x_1,...,x_n$ vom Umfang n alle Werte voneinander verschieden sein, bzw. nur wenige Beobachtungen auf ein und denselben Merkmalswert fallen.
Um einen Eindruck von der Verteilung der Merkmalswerte zu bekommen, werden die Merkmalsausprägungen klassifiziert und die Häufigkeit nicht mehr einzelnen Merkmalsausprägungen sondern den Klassen zugeordnet.
Wir zerlegen also den Wertebereich X in k disjunkte Klassen:

$$X = K_1 \cup K_2 \cup ... \cup K_k, \qquad K_i \cap K_j = \varnothing \quad \text{für} \quad i \neq j.$$

und zählen aus, wieviele Beobachtungen in welche Klasse fallen.

Jede Klasse ist charakterisiert durch ihr Klassengrenzen:

a_i^o - obere Grenze der Klasse K_i.
a_i^u - untere Grenze der Klasse K_i.
$(a_i^o = a_{i+1}^u)$, i=1,...,k

ihre Klassenbreite:

$$\Delta K_i = (a_i^o - a_i^u), \qquad i=1,...,k.$$

Die Begriffe der absoluten und relativen Häufigkeit bleiben erhalten, beziehen sich aber nun auf die Klassen K_i und nicht mehr auf einen Merkmalswert a_i.

Def.: $H_n(K_i)$ – Anzahl aller x_j mit $x_j \in K_i$, j=1,...,n
– *Absolute Klassenhäufigkeit* der Klasse K_i.

$h_n(K_i) = \frac{H_n(K_i)}{n}$ – Anteil aller x_j mit $x_j \in K_i$, j=1,...,n
– R*elative Klassenhäufigkeit* der Klasse K_i.

Es gilt: $0 \le H_n(K_i) \le n, \qquad 0 \le h_n(K_i) \le 1$

$$\sum_{i=1}^{k} H_n(K_i) = n \qquad \sum_{i=1}^{k} h_n(K_i) = 1$$

Die Gesamtheit der $H_n(K_i)$, $h_n(K_i)$, i=1,...,k bezeichnet man als (Klassen-)*Häufigkeitsverteilung* von X in der Stichprobe.
Die typische tabellarische Darstellung entspricht der für diskrete Zufallsgrößen. Dabei werden wieder die Summenhäufigkeiten mit einbezogen.

Def.: $H(i) = \sum_{j=1}^{i} H_n(K_j)$ - *absolute Summenhäufigkeit*

$h(i) = \sum_{j=1}^{i} h_n(K_j)$ - *relative Summenhäufigkeit*

H(i) – Anzahl der Objekte (Stichprobenwerte), deren Merkmalswert $x_j \le a_i^o$ ist.

h(i) – Anteil der Objekte, die einen Merkmalswert $x_j \le a_i^o$ besitzen.

Die typische grafische Darstellungsform der Klassenhäufigkeitsverteilung ist das *Histogramm.*

Es ist oft üblich, die Häufigkeiten an der Stelle der Klassenmitten im Histogramm miteinander zu verbinden, wobei links und rechts eine zusätzliche Klasse mit der Klassenhäufigkeit 0 angenommen wird. Die Linie bildet das sogenannte *Häufigkeitspolygon*.

Oft wählt man bei der grafischen Darstellung der Häufigkeitsverteilung die Skalierung der y-Achse so, daß $\frac{h_n(K_i)}{\Delta K_i}$ die Höhe der Säulen darstellt.
Dann entspricht der Flächeninhalt jeder Säule ihrer zugehörigen relativen Häufigkeit $h_n(K_i)$ (siehe Abb. 1.7).

Def.: $h_n^*(K_i) = \frac{h_n(K_i)}{\Delta K_i}$ - *relative Häufigkeitsdichte*.

$h_n^*(K_i)$ ist die auf die Klassenbreite bezogene Klassenhäufigkeit.

$\Rightarrow$ der Anteil $h_n(K_i)$ der Daten $x_1,...,x_n$, die in K_i hineinfallen, ist gleich der Fläche über K_i.

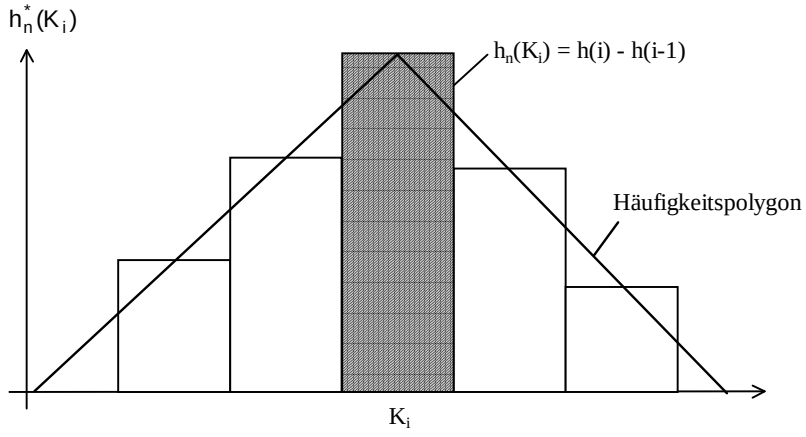


Abbildung 1.7

In der Regel werden bei großen Stichprobenanzahlen (n groß) die einzelnen Daten x_i, i=1,...,n, nicht gespeichert, sondern nur die Klassenhäufigkeitstabelle 'aufgehoben'. Dadurch gehen Informationen verloren, die die genaue Anzahl von Beobachtungsdaten x_i betreffen, die in einem beliebigen Intervall [a,b] liegen.
Um diese Anzahl aus der Klassen-Häufigkeitstabelle abzuschätzen, wird eine Modellannahme getroffen:

' Die Beobachtungen x_i sind in jeder Klasse gleichmäßig verteilt'.

Unter dieser Annahme kann man dann unter Verwendung der sogenannten empirischen Verteilungsfunktion die Anteile der ursprünglichen Stichprobendaten, die in einem Intervall [a,b] liegen, oder <a bzw. >b sind, abschätzen, siehe Abbildungen 1.8, 1.9.

Def.: Empirische Verteilungsfunktion (Summenhäufigkeitsfunktion) für klassifizierte Daten:

$$F_n(x) = \begin{cases} 0 & x < a_1^u \\ h(i-1) + b & x \in K_i \\ 1 & x > a_k^o \end{cases}$$

$$b = (x - a_i^u) h_n^*(K_i)$$

$$= \frac{(x - a_i^u)}{\Delta K_i} h_n(K_i)$$

Es gilt: $x \in K_i \Leftrightarrow h(i-1) < x \leq h(i)$

Diese Funktion gibt den Anteil der Beobachtungswerte an der Stichprobe an, die den Wert x nicht überschreiten ($x_j \leq x$). Siehe Abb. 1.8 !

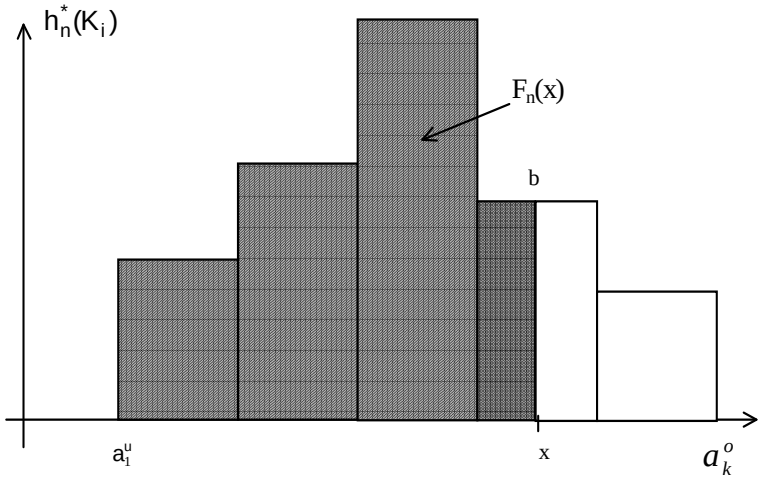


Abb. 1.8

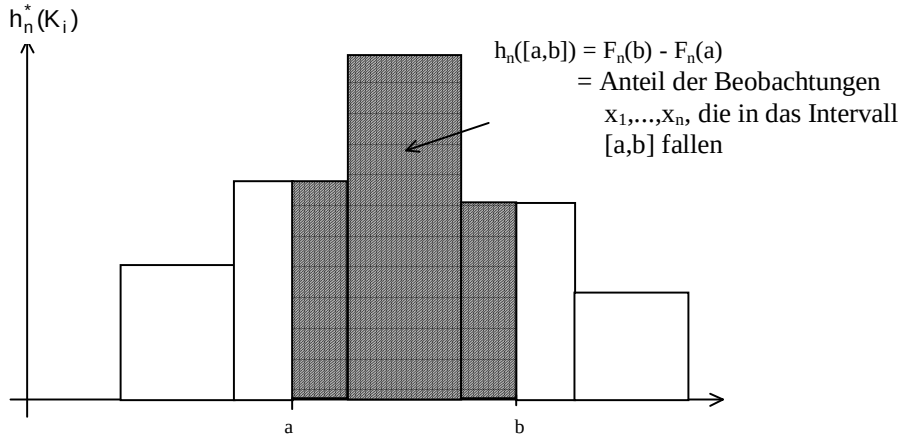


Abb. 1.9

Beispiel 6:

Es soll die Verteilung des Bruttomonatsverdienstes der Beschäftigten in einer Firma untersucht werden. Dazu werden 250 Beschäftigte befragt.

In Tabelle 1.5 sind die zugehörige Klasseneinteilung der Bruttomonatsverdienste und die Klassenhäufigkeit tabellarisch und grafisch dargestellt.

Klasse Nr. i	Bruttomonatsverdienst in DM	Klassenbreite in DM Δx_i	Beschäftigte	
			Anzahl $H_n(K_i)$	Anzahl $h_n(K_i)$
1	500 bis 800	300	6	0,024
2	über 800 bis 1100	300	13	0,052
3	über 1100 bis 1400	300	22	0,088
4	über 1400 bis 1700	300	32	0,128
5	über 1700 bis 2000	300	40	0,160
6	über 2000 bis 2300	300	42	0,168
7	über 2300 bis 2600	300	39	0,156
8	über 2600 bis 2900	300	31	0,124
9	über 2900 bis 3200	300	20	0,080
10	über 3200 bis 3500	300	5	0,020
$\sum$			250	1,000

Tabelle 1.5

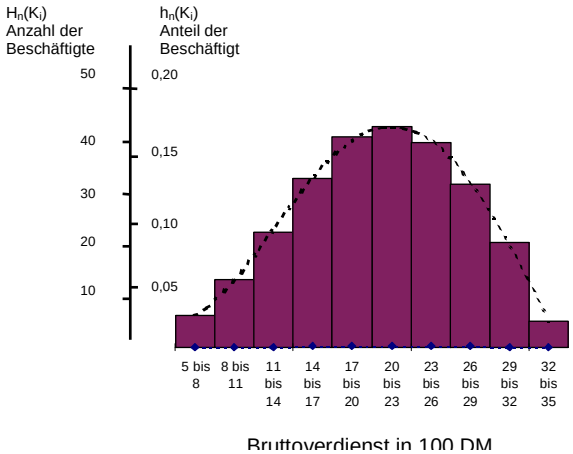


Abb. 1.10: Histogramm mit Klassen konstanter Breite sowie das entsprechende Häufigkeitspolygon (gestrichelter Linienzug)

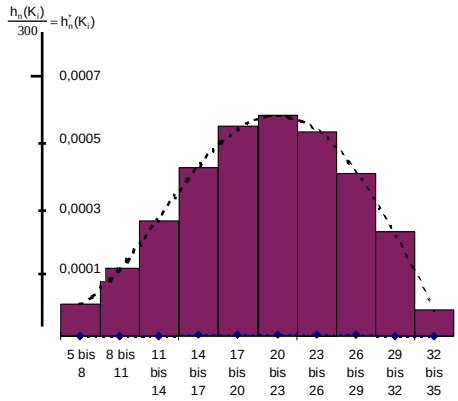


Abb. 1.11: Histogramm der relativen Häufigkeitsdichte

Tabelle 1.6 zeigt die zugehörigen Summenhäufigkeiten.

Klasse Nr. i	**Bruttomonatsverdienst in DM**	**Beschäftigte**	
		Kumulierte Anzahl H(i)	**Kumulierter Anteil h(i)**
1	bis 800	6	0,024
2	bis 1100	19	0,076
3	bis 1400	41	0,164
4	bis 1700	73	0,292
5	bis 2000	113	0,452
6	bis 2300	155	0,620
7	bis 2600	194	0,776
8	bis 2900	225	0,900
9	bis 3200	245	0,980
10	bis 3500	250	1,000

Tabelle 1.6: Absolute und relative Summenhäufigkeit

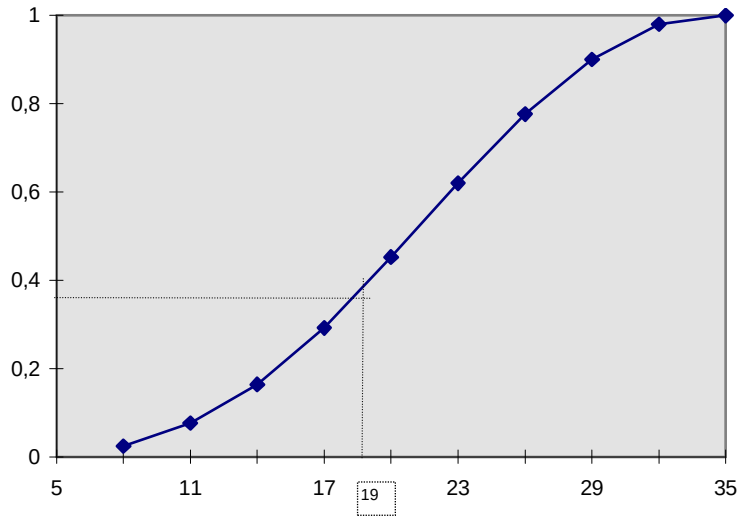
Hieraus können wir sofort entnehmen,

- wieviele Beschäftigte haben einen Bruttoverdienst von bis zu 2000 DM ?
  113 bzw. 45,2%
- wieviele Beschäftigte haben einen Bruttoverdienst zwischen 2000 DM und 2600 DM ?
  194 - 113 = 81 bzw. 77,6 - 45,2% = 32,4%

Nicht sofort entnehmen kann man Angaben über „Zwischenwerte“, z.B.
- wieviele Beschäftigte haben einen Bruttolohn von höchstens 1900 DM.

Diese Informationen sind der Klasseneinteilung verloren gegangen. Unter der Modellannahme, daß die Beobachtungen in einer Klasse gleichmäßig über die gesamte Klassenbreite streuen, kann man Angaben über solche „Zwischenwerte“ unter Benutzung der empirischen Verteilungsfunktion berechnen.

Abb. 1.12:



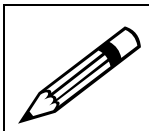
Interessiert etwa der Anteil der Empfänger, die ein Bruttomonatsverdienst von höchstens x = 1900 DM beziehen, so findet man mit Hilfe der Summenhäufigkeitsfunktionen diesen gesuchten Anteil zu 0,40.

$$F_n(1900) = 0{,}292 + 0{,}16 \cdot \left(\frac{1900 - 1700}{300} \right) = 0{,}4$$

Anteil der Empfänger, die einen Bruttolohn zwischen 1700 und 1900 DM bekommen:

$$F_n(1900) - F_n(1700) = 0{,}4 - 0{,}292 = 0{,}108$$

$\Rightarrow$ 10,8% oder ca. 27 Beschäftigte.



- Wieviel % der Beschäftigten haben einen Bruttoverdienst bis 1100 DM?

  $F_n(1100) \cdot 100\% = 0{,}076 \cdot 100\% = 7{,}6\%$

- Wie groß sind Anteil und Anzahl der Beschäftigten mit einem Bruttoverdienst zwischen 1700 und 2300 DM?

  $F_n(2300) - F_n(1700) = 0{,}620 - 0{,}292 = 0{,}328 \Rightarrow 32{,}8\%$

  $\text{Anzahl} = 0{,}328 \cdot 250 = 82$

1.4.1 Wahl der Klassen

Wie wählt man nun die Klassen?
Wieviele Klassen mit welcher Breite nimmt man?

Das Beispiel 7 illustriert die Abhängigkeit des „Informationsgewinns“ über die Verteilung der Länge von 40 Schrauben von der gewählten Klassenbreite.

Beispiel 7: In der Tabelle 1.7 sind keine Klassen gebildet sondern die Merkmalswerte nur ausgezählt worden. Aus dem Stabdiagramm erkennt man keine Häufungen. Die Informationen müssen „verdichtet“ werden.
In den Tabellen 1.8 und 1.9 sind Klasseneinteilungen mit der Klassenbreite 5 bzw. 9 mm gebildet worden. Die entsprechenden Histogramme zeigen eine symmetrische Verteilung und eine Häufung im mittleren Bereich.
Tabelle 1.10 zeigt desweiteren eine Klasseneinteilung der Klassenbreite 30 mm in nur 2 Klassen. Hier erkennt man wiederum keine Häufungen. Es gehen zu viele Informationen verloren.
Wählen wir also zu kleine Klassenbreiten (zu viele Klassen) (Tabelle 1.7), so erkennt man keine Häufungen. Dasselbe geschieht bei der Wahl zu großer Klassenbreiten (zu wenig Klassen) (Tabelle 1.10).

Häufigkeitsverteilungen der Längen von 40 Schrauben:

Urliste: Länge von 40 Schrauben in mm

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	178	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

Länge i	Anzahl $H_n(K_i)$
119	1
125	1
126	1
128	1
132	1
135	3
136	1
138	2
140	2
142	2
144	2
145	2
146	2
147	2
148	1
149	1
150	2
152	1
153	1
154	1
156	1
157	1
158	1
161	1
163	1
164	1
165	1
168	1
173	1
176	1

Tabelle 1.7

Klassenbreite: 5

Länge (mm)	Anzahl	Häufigkeit
118-122	1	1
123-127	2	2
128-132	2	2
133-137	4	4
138-142	6	6
143-147	8	8
148-152	5	5
153-157	4	4
158-162	2	2
163-167	3	3
168-172	1	1
173-177	2	2
		Summe 40

Tab. 1.8

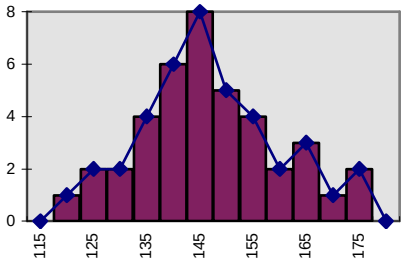


Abb. 1.13

Klassenbreite: 9

Länge (mm)	Tab. 1.10	Häufigkeit
118-126	3	3
127-135	5	5
136-144	9	9
145-153	12	12
154-162	5	5
163-171	4	4
172-180	2	2
		Summe 40

Tabelle 1.9

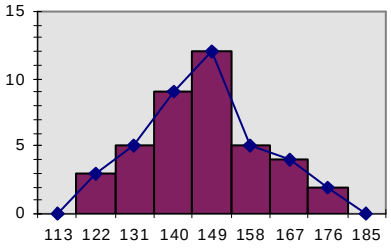


Abb. 1.14

Klassenbreite: 30

Länge	Häufigkeit
118-147	23
148-177	17

Tab. 1.10

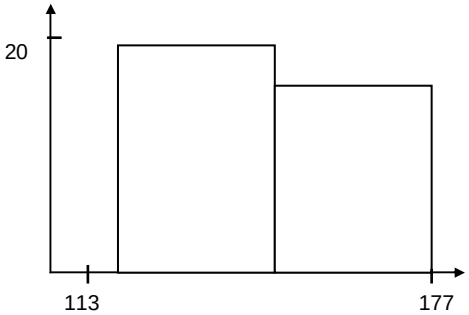


Abb. 1.15

Vorschriften zur Klassenbildung (heuristische Vorschriften):

$n \geq 25$!! (Mache keine statistische Analyse mit zu wenigen Daten!!)

- **Wahl der Klassenzahl k:**

  1. „Faustregel“:
     $k = \sqrt{n}$ $5 \leq k \leq 20$ ($\Rightarrow$ n > 25 !)

  2. DIN-Normen für spezielle Anwendungen, z.B. DIN 55302

n	k
bis 100	>10
bis 1000	>13
bis 10000	>16

  3. Informationstheoretisches Kriterium:
     $k = 1 + 3{,}32 \log(n)$, $5 \leq k \leq 20$ $k \leq 5 \log(n)$

  In jedem Fall wird k auf einen ganzzahligen Wert auf- oder abgerundet.

- **Wahl der Klassengrenzen/ Reduktionslage:**

Vorschriften, die besagen, wie mit Werten zu verfahren ist, die auf eine Klassengrenze fallen:

1. Die Klassengrenzen sollen so gewählt werden, daß kein Wert auf eine Klassengrenze fällt. (z.B., indem man sie zwischen zwei Werte legt (½-Genauigkeit!)).

   Das wird z.B. erreicht, indem man den unteren Wert a_1^u der ersten Klasse K_1 gemäß $a_1^u = x_{min} - \varepsilon$ festlegt, wobei ε die halbe Meßgenauigkeit ist und die Klassenbreite B auf die Meßgenauigkeit aufrundet.

   a_1^u wird als ***Reduktionslage*** bezeichnet.

2. Die Werte, die auf eine Klassengrenze fallen, werden zur Hälfte der einen und zur Hälfte der anderen zugeordnet. Bei ungeraden Anzahlen erhält die obere Klasse den „überschüssigen“ Wert.

usw.

- **Wahl der Klassenbreite:**

Man wählt in der Regel eine konstante Klassenbreite $B = \Delta K_i$ für i=1,...,k.

Sei x_{min} – kleinster Wert und x_{max} – größter Wert der Stichprobe.

In Anbetracht, daß x_{min} und x_{max} nicht auf eine Klassengrenze fallen sollen, muß ein Bereich von x_{min} - a bis x_{max} + b in k Klassen zerlegt werden (a,b – vorgegebene Größe, z.B. (½-Maßeinheit)):

Klassenbreite = B = $\frac{(x_{max} + b - x_{min} + a)}{k}$

Man **rundet B** auf die Meßgenauigkeit **auf**!

Vorschrift * :

Klassenzahl: $k = \sqrt{n}$, $5 \le k \le 20$, k wird auf einen ganzzahligen Wert gerundet

Reduktionslage: (= a_1^u) a_1^u = x_{min} - ε, wobei ε die halbe Meßgenauigkeit ist

Klassenbreite: B = $\frac{(x_{max} - x_{min} + 1)}{k}$, (a=b=ε), B wird auf die Meßgenauigkeit ***auf***gerundet.

Zu Beispiel 7:
Wir wollen eine Klasseneinteilung der in der Urliste Beispiel 7 aufgelisteten Längen der 40 Schrauben bilden:

1. Variante: Klassenzahl: Vorschrift 3: $k = 1 + 3{,}32 \log(40) = 1 + 3{,}32 \cdot 1{,}602 = 6{,}32$

$\Rightarrow$ k = 7 (wir runden hier auf).

Reduktionslage: x_{min} = 119, x_{max} = 176

Wir zerlegen den Bereich von 118 – 177 in 7 Klassen (a=b=1).

a_1^u =x_{min}-1=118

Klassenbreite: $B = \left(\frac{177 - 118)}{7}\right) = 8{,}4 \Rightarrow$ B = 9

Wir erhalten die in Tabelle 1.9 gegebene Klasseneinteilung.

Es ist noch zu überprüfen, ob ein Wert auf eine Klassengrenze fällt. Das ist hier nicht der Fall.

2. Variante: Klasseneinteilung nach Vorschrift* :

Klassenzahl: $\sqrt{40} = 6{,}32 \Rightarrow$ k = 6 (wir runden hier nach der üblichen Vorgehensweise).

Reduktionslage: x_{min} = 119, x_{max} = 176

Wir wählen: a_1^u =x_{min}-1/2 und zerlegen den Bereich von 118,5 – 176,5 in 6 Klassen (a=b=ε=1/2).

Klassenbreite: $B = \left(\frac{176 - 119 + 1}{6} \right) = 9{,}67 \quad \Rightarrow$ B = 10

	Bilden Sie die Klassenhäufigkeitstabelle der 40 Schraubenlängen nach Vorschrift*.

1.5 Übungsaufgaben

1. Sind folgende Merkmale diskret oder stetig?

   a) Die durch eine wahlberechtigte Person der BRD gewählte Partei bei der Bundestagswahl.
   b) Kraftstoffverbrauch eines Personenkraftwagens auf 100 km.
   c) Zahl der pro Stunde in einem Geschäft eintreffenden Kunden.
   d) Lebensdauer einer bestimmten Leuchtstoffröhrenart.
   e) Plazierung beim 100-Meter-Lauf (von Studenten der HTW)

   Geben Sie jeweils das Merkmal, die Objektmenge und den Wertebereich des Merkmals an. Welche Art Skalierung liegt hier vor?

2. An einem Bankschalter werden die Kundenankünfte (Anzahl der pro 10-Minuten-Zeitintervall ankommenden Kunden) beobachtet. Für 40 derartige Zeitintervalle erhält man folgende Ergebnisse:

   0, 0, 1, 3, 4, 1, 2, 2, 1, 1,
   1, 2, 3, 0, 2, 0, 1, 3, 1, 2,
   2, 0, 1, 1, 6, 1, 0, 2, 3, 1,
   1, 4, 2, 3, 2, 0, 3, 0, 1, 2,

Ermitteln Sie absolute und relative Häufigkeiten der Kundenankünfte und stellen Sie Häufigkeitsverteilung und Summenhäufigkeitsfunktion grafisch dar!

3. 1000 Motoren eines bestimmten Typs weisen folgende Lebensdauerverteilung auf:

Lebensdauer in Jahren	Anzahl der Motoren
bis 2	33
über 2 bis 4	276
über 4 bis 6	404
über 6 bis 8	237
über 8 bis 10	50

Tab.: Lebensdauerverteilung

   Stellen Sie diese Häufigkeitsverteilung und die dazugehörende Summenhäufigkeitsfunktion grafisch dar und bestimmen Sie den Anteil der Motoren mit einer Lebensdauer von mehr als 5 Jahren!

4. Werten Sie nachstehende Urliste aus!
   Wählen Sie eine geeignete Klassenzahl, den unteren Wert der 1. Klasse und die Klassenbreite nach Vorschrift*!
   Stellen Sie die absoluten, relativen Klassenhäufigkeiten, sowie die relativen (Klassen-) Summenhäufigkeiten tabellarisch und grafisch (Histogramm, Summenhäufigkeitsfunktion) dar!

Tab.: Lebensdauer x_i von 30 Glühbirnen in Stunden

i	x_i
1	375,3
2	392,5
3	467,9
4	503,1
5	591,2
6	657,8
7	738,5
8	749,6
9	752,0
10	765,8
11	772,5
12	799,4
13	799,6
14	803,9
15	808,7
16	810,5
17	812,1
18	848,6
19	867,0
20	904,3
21	918,3
22	935,4
23	951,1
24	964,9
25	968,8
26	1006,5
27	1014,7
28	1189,0
29	1215,6
30	1407,2

2 Stichprobenkennwerte

Häufigkeitsverteilungen stellen eine Zusammenfassung („Verdichtung“) der Urdatenliste dar. Häufig ist man jedoch noch daran interessiert, diese Häufigkeitsverteilungen in noch knapperer Form zu charakterisieren.

Dies ist durch die Berechnung statistischer Maßzahlen möglich.
Diese Maßzahlen charakterisieren in knapper Form die Gestalt der Häufigkeitsverteilungen.

Wir unterscheiden zwischen Lagemaßen (Maßzahlen, die die Lage auf der x-Achse beschreiben) und Streuungsmaßen (Maßzahlen, die die Variabilität der Daten beschreiben).

$\tilde{x}_{0,5}$ sei der Wert, der die Stichprobe (Wertemenge) teilt: d.h. 50 % der Daten liegen links und 50 % liegen rechts von $\tilde{x}_{0,5}$. Dann ist $\tilde{x}_{0,5}$ ein Lageparameter.

Die Abstände: $\left|\tilde{x}_{0,5} - x_{min}\right|$ und $\left|x_{max} - \tilde{x}_{0,5}\right|$ sagen etwas über die Streuung um $\tilde{x}_{0,5}$ aus.

Wir erkennen daran, daß z.B. die Verteilung rechtsschief sein muß, falls $\left|\tilde{x}_{0,5} - x_{min}\right| < \left|x_{max} - \tilde{x}_{0,5}\right|$ gilt.

Diese Abstände sind Streuungsparameter!

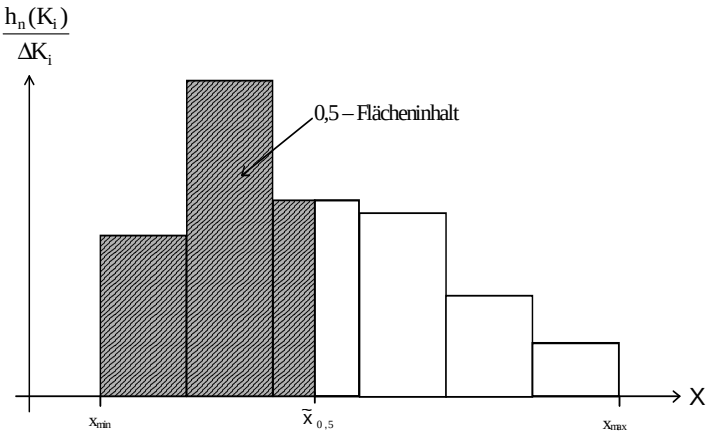


Abb. 2.1

2.1 Lageparameter (Mittelwert, Median, Quantile, Modelwert):

2.1.1 Das arithmetische Mittel

Sei X ein zufälliges Merkmal und sei $x_1,...,x_n$ eine Stichprobe von X vom Umfang n.

Def.: Unter dem *arithmetischen Mittelwert* der Stichprobe verstehen wir den durchschnittlichen Stichprobenwert

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

Beispiel 8: Legt ein Angestellter den Weg zwischen Wohnung und Arbeitsstätte an 5 Tagen in 12, 10, 16, 12 und 17 Minuten zurück, dann beträgt die durchschnittliche Zeit, die er für den Weg benötigt:

$$\bar{x} = \frac{(12+10+16+12+17)}{5} = 13{,}4 \text{ Minuten}$$

Besitzen Beobachtungswerte den gleichen numerischen Wert a_i, so empfiehlt es sich, sie zusammenzufassen:

$$\bar{x} = \frac{(2 \cdot 12 + 10 + 16 + 17)}{5}$$

Ist X ein diskretes Merkmal und $X = \{a_1, \ldots, a_k\}$, so gilt offenbar:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} a_i H_n(a_i) = \sum_{i=1}^{k} a_i h_n(a_i)$$

Ist X ein stetiges Merkmal und liegen die Daten in klassifizierter Form und als Urliste vor, so gilt offenbar:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i H_n(K_i).$$

Dabei ist $\bar{x}_i$ der Mittelwert der Klasse i:

$$\bar{x}_i = \frac{1}{H_n(K_i)} \sum_{x_j \in K_i} x_j .$$

Liegt die Urliste nicht mehr vor, sondern nur noch die Klasseneinteilung, so verwendet man in der o.g. Formel anstelle der Klassenmittelwerte die Klassenmitten x_i' als Approximation und erhält als Approximation für das arithmetische Mittel:

$$\bar{x}' = \frac{1}{n}\sum_{i=1}^{k} x_i' \cdot H_n(K_i)$$

Beispiel 9: Wie groß ist die mittlere Lebensdauer von 1000 Maschinen, deren Lebensdauerverteilung durch folgende Tabelle gegeben ist:

i	K_i	x_i'	$H_n(K_i)$
1	bis 2 Jahre	1	33
2	(2-4] Jahre	3	276
3	(4-6] Jahre	5	404
4	(6-8] Jahre	7	237
5	(8-10] Jahre	9	50

$$\bar{x}' = \frac{(33 + 276 \cdot 3 + 404 \cdot 5 + 237 \cdot 7 + 50 \cdot 9)}{1000} = 4{,}99 \text{ Jahre}$$

Die mittlere Lebensdauer beträgt etwa 5 Jahre.

Nachteile des arithmetischen Mittels:

Das arithmetische Mittel ist nicht immer das aussagekräftigste Mittel.
Betrachtet man zum Beispiel eine Gruppe von 10 Personen, von denen 9 ein Jahreseinkommen von je 40000 DM und eine ein Jahreseinkommen von 400000 DM bezieht, so ist das arithmetische Mittel

$$\bar{x} = 9 \cdot 40000 + 1 \cdot 400000 = 76000 \text{ DM}$$

zur Charakterisierung dieser Personengruppe wenig sinnvoll.

Gibt es also sogenannte „Ausreißer", d.h. Werte, die weit weg vom Zentrum der Verteilung entfernt liegen oder hat die Verteilung kein eindeutiges Zentrum (z.B. ein Gipfel), so ist das arithmetische Mittel wenig aussagekräftig.

Passender ist der im folgenden Abschnitt definierte Median.

2.1.2 Der Median

Der Median (Zentralwert) $x_{0,5}$ ist die Merkmalsausprägung desjenigen Elements, das in der der Größe nach geordneten Beobachtungsreihe in der Mitte steht.
Damit man die Stichprobe ordnen kann, muß das Merkmal mindestens ordinal skaliert sein!
Sei $x_1,...,x_n$ die Stichprobe und ihre der Größe nach geordnete Reihenfolge sei

$$x_{[1]} < x_{[2]} < \ldots < x_{[n]}.$$

Def.: Der *Median* der Stichprobe ist definiert durch:

$$x_{0,5} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{für n ungerade} \\ \dfrac{x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n+2}{2}\right]}}{2} & \text{für n gerade} \end{cases}$$

Für das oben angeführte Beispiel eines Angestellten, der an 5 Tagen die von seiner Wohnung zur Arbeitbenötigte Zeit festhält, ergibt sich die geordnete Beobachtungsreihe 10 12 12 16 17 und folglich der Median

$$x_{0,5} = x_{[3]} = 12 \text{ Minuten} .$$

Vorteil des Medians: Er ist relativ unempfindlich gegenüber Ausreißern.

Für gruppierte Daten (d.h., wenn nur noch die Klassenhäufigkeitstabelle, nicht aber die Urliste vorliegt) erhält man eine Approximation $\tilde{x}_{0,5}$ für den Median $x_{0,5}$ unter Verwendung der empirischen Verteilungsfunktion $F_n(x)$ auf folgende Weise:

Der Median liegt in der Klasse i, in der die Summenhäufigkeitsfunktion den Wert 0,5 erreicht: D.h., für die Klasse i, die den Median enthält, muß gelten:

$$h_{i-1} < 0{,}5 \leq h_i$$

Wir bestimmen also zunächst die Klasse i, die den Median enthält. Der Median ergibt sich dann durch die Umstellung der Formel:

$$F_n(x_{0,5}) = 0{,}5 = h(i-1) + h_n(K_i) \cdot \frac{(\tilde{x}_{0,5} - a_i^u)}{(a_i^o - a_i^u)} \qquad \text{nach } \tilde{x}_{0,5}$$

als

$$\tilde{x}_{0,5} = a_i^u + \frac{(0{,}5 - h(i-1))}{h_n(K_i)} \cdot (a_i^o - a_i^u) .$$

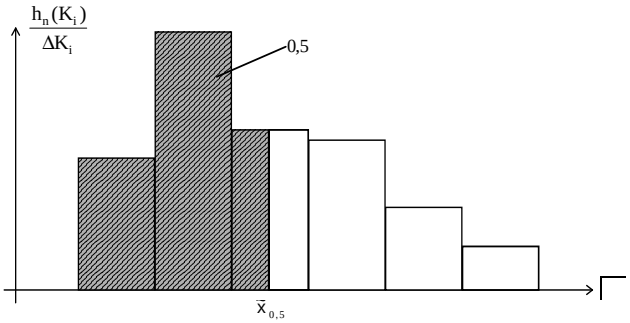


Abb. 2.2

Für das Beispiel der Bruttomonatsverdienste von 250 Beschäftigten eines Betriebes ergibt sich aus Tabelle 1.6: die Klasse, in welcher der Median liegt, ist i = 6 und

$$\begin{aligned}\tilde{x}_{0,5} &= 2000 + \frac{(0{,}5 - 0{,}452)}{0{,}168} \cdot 300 \\ &= 2085{,}71\text{DM} \\ &\approx 2100\ \text{DM}\end{aligned}$$

D.h., 50% der Beschäftigten der Firma verdienen mehr und 50% weniger als 2100DM.

Unter Benutzung der grafischen Darstellung der empirischen Verteilungsfunktion ergibt sich eine grafische Näherungslösung für den Median, wie unten stehenden Skizze zeigt.

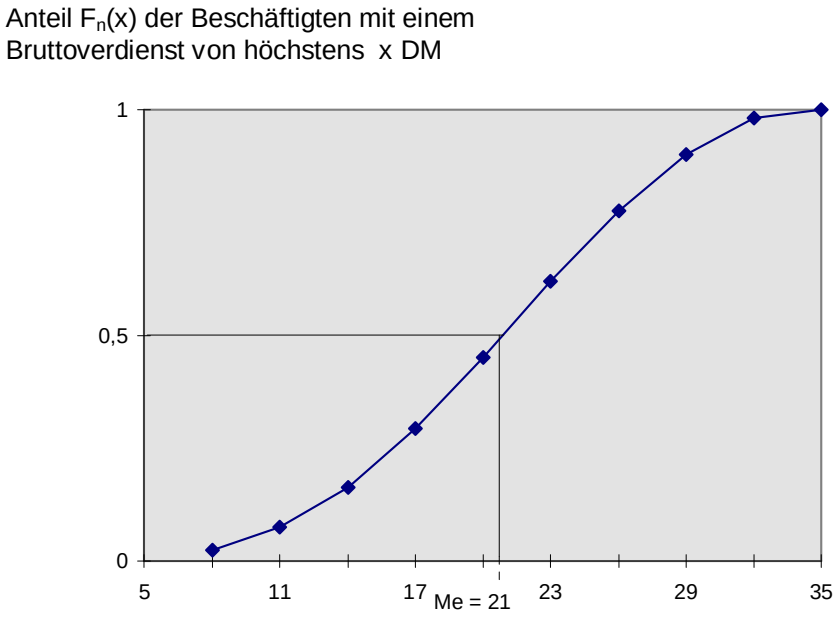


Abb. 2.3: Grafische Bestimmung des Medians bei klassifizierten Daten

2.1.3 Das α-Quantil einer Beobachtungsreihe

Unter dem α-Quantil einer Beobachtungsreihe $x_1,...,x_n$ versteht man den Wert $x_\alpha, (0 < \alpha < 1)$, den $\alpha \cdot 100\%$ der (bzw. $n \cdot \alpha$) Beobachtungswerte nicht überschreiten.
Sei $x_{[1]} < x_{[2]} < ... < x_{[n]}$ die geordnete Stichprobe.

Def.: Der Wert

$$x_\alpha = \begin{cases} \dfrac{(x_{[k]} + x_{[k+1]})}{2} & \text{falls } k = n \cdot \alpha \text{ eine ganze Zahl ist} \\ x_{[k]} & \text{falls } n \cdot \alpha \text{ keine ganze Zahl ist und k die auf } n \cdot \alpha \text{ folgende Zahl ist} \end{cases}$$

heißt *α-Quantil* der Beobachtungsreihe $x_1,...,x_n$.

Bei Vorliegen gruppierter Daten erhält man eine Approximation $\tilde{x}_\alpha$ für das α-Quantil analog zum Median:

Man bestimme zunächst die Klasse i in der das α-Quantil liegt, d.h. für die gilt:

$$h(i-1) < \alpha \le h(i)$$

$\tilde{x}_\alpha$ ergibt sich dann aus der Formel:

$$\tilde{x}_\alpha = a_i^u + \frac{(\alpha - h(i-1))}{h_n(K_i)} \cdot (a_i^o - a_i^u) .$$

Die grafische Bestimmung von $\tilde{x}_\alpha$ aus der empirischen Verteilungsfunktion zeigt nachstehende Skizze:

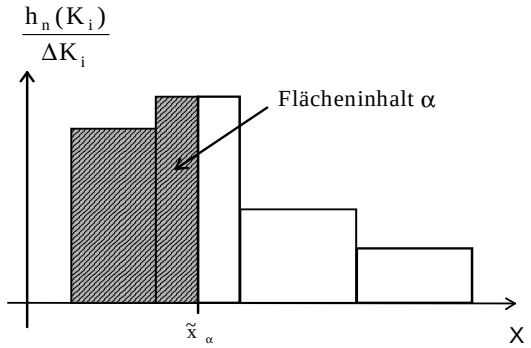


Abb. 2.3

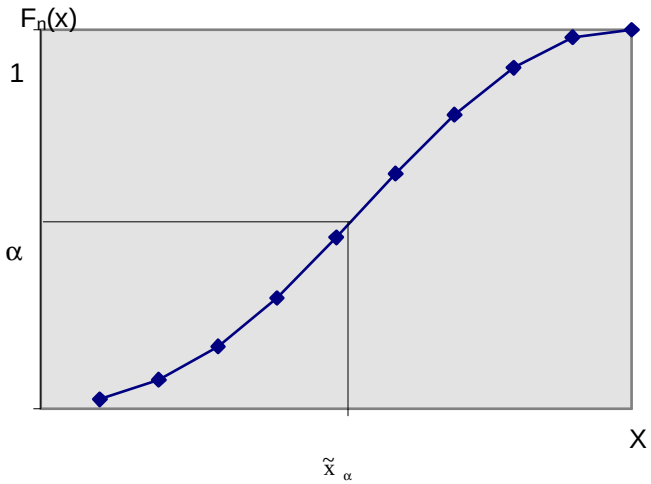


Abb. 2.4

Der Median ist damit ein spezielles, nämlich genau das 0,5-Quantil der Beobachtungsreihe.

Wir bezeichnen das
0,25-Quantil als *unteres Quartil*.
0,75- Quantil als *oberes Quartil*.

2.1.4 Der Modalwert

Bei bestimmten Daten wie der Geschlechtszugehörigkeit, Beruf etc. kann man keinen Mittelwert, Median usw. bilden, da die Daten nur nominal skaliert sind.
Ein Lagemaß, das auch für solche Merkmale definiert ist, ist der „Modus“ oder Modalwert.

Def.: Der *Modalwert* x_{mod} ist der häufigste Wert einer Verteilung.
Ist X diskret mit $X = \{a_1, a_2, ...\}$ so gilt

$h_n(x_{mod}) \geq h_n(a_j)$ für alle Merkmalsausprägungen $a_1,...,a_k$.

Bei gruppierten Daten ist der Modalwert die Klassenmitte der am dichtesten besetzen Klasse; wir schreiben in diesem Fall auch x_{mod}' statt x_{mod} .

Bemerkung: In Fällen, bei denen mehrere Ausprägungen diese Bedingung erfüllen, d.h. der Modalwert nicht eindeutig ist, ist es nicht sinnvoll, ihn als Lagemaß zu benutzen.

Im Beispiel 8 ist der Modalwert $x_{mod} = 12$.

2.1.5 Bemerkungen zu den Lagemaßzahlen

Bei mehrgipfligen und insbesondere u-förmigen Verteilungen sind im Gegensatz zu eingipfligen Verteilungen die Lagemaße oft nicht charakteristisch für die Häufigkeitsverteilung:

Beispiel 10: Die Lebensdauer der Menschen im Mittelalter entsprach einer u-förmigen Häufigkeitsverteilung:

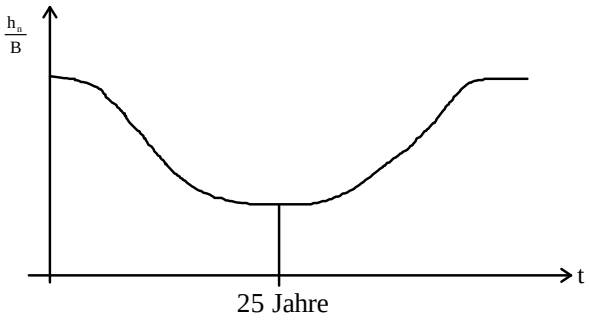


Abb. 2.5

Die Säuglingssterblichkeit war sehr hoch, aber danach hatte man gute Chancen, ein recht hohes Alter zu erreichen.
Die Aussage: „Im Mittelalter wurden die Menschen im Durchschnitt 25 Jahre alt“ impliziert eine falsche Vorstellung von der damaligen Lebenserwartung.

Die Lagemaße dienen insbesondere zur näheren Charakterisierung eingipfliger Verteilungen
Seien $\bar{x}$ - das arithmetische Mittel

$x_{0,5}$ - der Median

x_{mod} - der Modalwert

Def.: Eine Häufigkeitsverteilung heißt

rechtsschief oder *linksschief*, falls	$\bar{x} > x_{0,5} > x_{mod}$
linksschief oder *rechtssteil,* falls	$\bar{x} < x_{0,5} < x_{mod}$
symmetrisch, falls	$\bar{x} = x_{0,5} = x_{mod}$

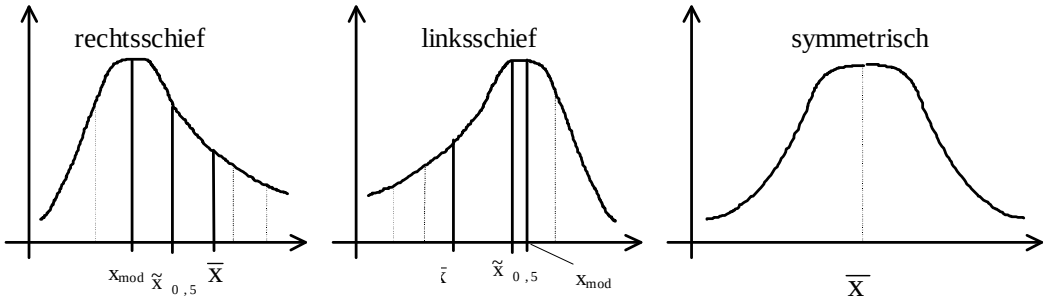


Abb. 2.6

2.2 Streuungsmaße

Lagemaße geben i.a. wenig Auskunft über die Gesamtstruktur einer Häufigkeitsverteilung. Sie beschreiben nur das „Zentrum“ einer Verteilung, geben aber keine Auskunft darüber, wieweit ein konkreter Merkmalswert von einem solchen Zentrum abweichen kann.

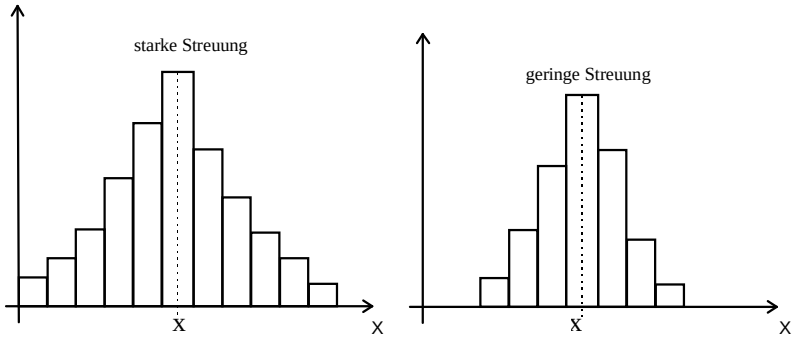


Abb. 2.7

Maße, die die Abweichung von einem Zentrum einer Häufigkeitsverteilung beschreiben, heißen Streuungsmaße oder Dispersionsmaße.
Die gemeinsame Angabe von „Lage- und Streuungsmaßen“ ergeben eine recht präzise Vorstellung von der gesamten Verteilung.

2.2.1 Spannweite (Range)

Seien $x_1,...,x_n$ eine Stichprobe, x_{min} der kleinste und x_{max} der größte Wert der Stichprobe.

$R := x_{max} - x_{min}$

Es ist nur sinnvoll, die Spannweiten verschiedener Beobachtungsreihen zu vergleichen, wenn sie gleichen Stichprobenumfang besitzen!
Die Spannweite ist kein sehr gutes Streuungsmaß, denn es gehen nur zwei Werte der Verteilung ein.
$\Rightarrow$ Sie ist sehr empfindlich gegenüber Ausreißern!

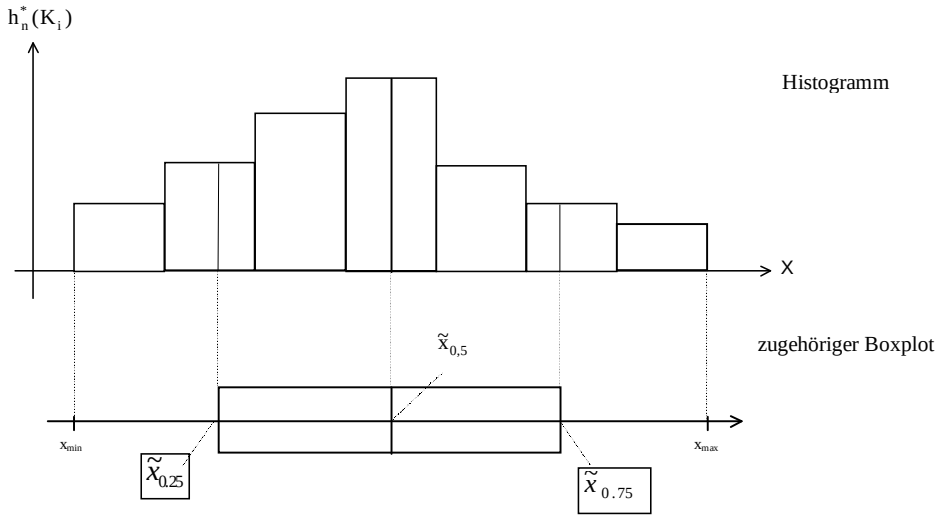
2.2.2 Quartilabstand

$$S_Q = x_{0,75} - x_{0,25}$$

(bzw. bei Vorliegen gruppierter Daten (Klasseneinteilung) $S_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$)

S_Q ist nicht so empfindlich gegenüber Ausreißern.
S_Q gibt den Bereich an, in dem etwa die Hälfte (50 %) aller Beobachtungen liegen.

Oft stellt man eine Häufigkeitsverteilung vereinfacht als sogenannten Boxplot dar:



(Die Breite des Kästchens hat keine Bedeutung!)

Abb. 2.8

Man zeichnet lediglich die Lage des minimalen und maximalen Wertes der Stichprobe, sowie den Median und die Quartile auf.
Die Abstände der Werte ergeben einen Eindruck von Lage, Schiefe und Streuung der Verteilung.
Man verwendet Box-Plots auch zum 'schnellen' Vergleich mehrerer Häufigkeitsverteilungen, siehe Abbildung 2.9.

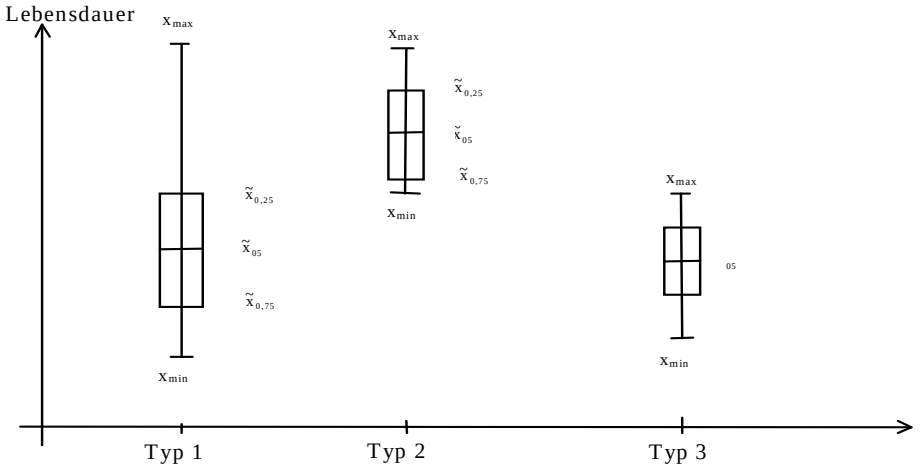


Abb. 2.9: Lebensdauerverteilungen verschiedener Maschinentypen im Vergleich

2.2.3 Zentilabstand

$$S_{Q10} = x_{20} - x_{10}\,,\ S_{Q10} = x_{30} - x_{20}\,,\ S_{Q10} = x_{40} - x_{30}\,,....$$

2.2.4 Streuung

Die Streuung ist eine Maßzahl für die Variabilität, die Daten auf metrischem Niveau voraussetzt. Sie nimmt Bezug auf den Mittelwert.
Die Streuung ist der „mittlere“ quadratische Abstand der Daten von ihrem Mittelwert:

Def.: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right)$

heißt Streuung (Stichprobenvarianz) der Stichprobe $x_1,...,x_n$.

Besonderheit: Wir nehmen den quadratischen Abstand!

Bemerkung: Die Begründung dafür, daß in der Formel für s^2 durch (n-1) und nicht durch n dividiert wird, liefert die schließende Statistik (siehe Kapitel III).

Ist X diskret mit $X = \{a_1, \ldots, a_k\}$ so gilt:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(a_i - \bar{x})^2 \cdot H_n(a_i) = \frac{n}{n-1}\sum_{i=1}^{k}(a_i - \bar{x})^2 \cdot h_n(a_i)$$

Für gruppierte Daten, bei denen die Klassenmittelwerte $\bar{x}_j$ ($\bar{x}_j$ - Mittelwert aller Daten, die in Klasse K_j fallen) bekannt sind, wählt man als Approximation von s^2 die Größe:

$$s_0^2 = \frac{1}{n-1}\sum_{j=1}^{k}(\bar{x}_j - \bar{x})^2 \cdot H_n(K_j)$$

ISind die Klassenmittelwerte nicht bekannt, so ersetzt man sie durch die Klassenmitten $\bar{x}_j'$, sowie $\bar{x}$ durch $\bar{x}'$ erhält eine Approximation der Streuung gemäß

$$s_*^2 = \frac{1}{n-1}\sum_{j=1}^{k}(x_j' - \bar{x}')^2 \cdot H_n(K_j)$$

Untersuchungen haben gezeigt:
Ist die Klassenzahl klein im Vergleich zum Stichprobenumfang n, so wird die Streuung s^2 durch s_*^2 über und durch s_0^2 unterschätzt!

Sheppardsche Korrekturformel:

Haben alle Klassen die gleiche Klassenbreite B, so läßt sich der Wert s_*^2 korrigieren:

$$s_{**}^2 = s_*^2 - \frac{B^2}{12}.$$

Beispiel 11: Lebensdauer von 30 Glühbirnen

Klasse K_j	$H_n(K_j)$	$\bar{x}_j$	x_j'
[300, 700)	6	497,767	500
[700, 800)	7	765,200	750
[800, 900)	6	825,133	850
[900, 1000)	6	940,467	950
[1000, 1500)	5	1186,660	1250

$\bar{x} = 826{,}53$, k = 5

$$s_0^2 = \frac{1}{29}\sum_{j=1}^{5}(\bar{x}_j - 826{,}53)^2 \cdot H_n(K_i) = \underline{45781{,}986}, \qquad s_*^2 = \underline{57664{,}792}$$

a) Verifizieren Sie die obigen Ergebnisse für $\bar{x}$, s_0^2 und s_*^2 !
b) Berechnen Sie s^2 anhand der Tabelle der Aufgabe 4 in der Übungsaufgaben-Serie 1.5!
c) Berechnen Sie s_{**}^2 und $\bar{x}'$!

	d) Vergleichen Sie s^2, s_0^2, s_*^2 und s_{**}^2, sowie $\bar{x}'$ und $\bar{x}$ miteinander !

2.2.5 Standardabweichung

Def.: $s = \sqrt{s^2}$ – Standardabweichung.

Die Standardabweichung ist eines der gebräuchlichsten statistischen Streuungsmaße. Sie ist die Wurzel aus der Streuung. Sie hat die gleiche Dimension wie der Mittelwert.

Bemerkung:
Mit Hilfe des arithmetischen Mittels $\bar{x}$ und der Standardabweichung s einer Stichprobe läßt sich für Zufallsgrößen X, deren Verteilung der Beobachtungen (Realisierungen) von X in der Grundgesamtheit eine bestimmte Form hat, der Bereich abschätzen, in dem ***sämtliche*** Beobachtungen von X liegen.
Sind die Beobachtungen von X zum Beispiel alle gleichhäufig und begrenzt auf endliches Intervall (wir sprechen von einer Gleichverteilung), so sind die Intervallgrenzen gerade durch $\bar{x} \pm s\sqrt{3}$ abschätzbar.
Sind die Beobachtungen von X alle symmetrisch und nicht gleichmäßig um ein Mittelwert verteilt (wir sprechen von einer Normal- bzw. Gaußverteilung), so ist der Wertebereich von X durch $\bar{x} \pm 3s$ abschätzbar.
Die Begründung für diese Sachverhalte liefert die Wahrscheinlichkeitsrechnung (siehe Kapitel II).

2.2.6 Der Variationskoeffizient

Die Streuung und Standardabweichung benutzen wir als Bezugspunkt des arithmetischen Mittels und messen die mittlere Abweichung vom Mittelwert.

Für einen Mittelwert $\bar{x} = 10000$ ist die Standardabweichung s = 10 eine geringe Abweichung, für einen Mittelwert $\bar{x} = 1$ stellt s = 10 eine große Abweichung dar.

Um die Größe der Standardabweichung relativ zum Mittelwert beurteilen zu können, verwendet man den Variationskoeffizienten:

Def.: $v = \frac{s}{\bar{x}}$ – *Variationskoeffizient*

v ist ein vom Mittelwert bereinigtes Streuungsmaß. v eignet sich besonders zum Vergleich der Streuungen verschiedener Meßreihen.

Für unsere beiden Beispiele oben erhalten wir:

$$v_1 = \frac{10}{10000} = 0{,}001 \text{ und } v_2 = \frac{10}{1} = 10 .$$

a) Beweisen Sire die Formeln (3) und (4) !
b) Zeigen Sie, daß (1) aus (3) und (2) aus (4) folgen!

2.2.7 Eigenschaften des arithmetischen Mittels und der Stichprobenstreuung - Hinweise zur praktischen Berechnung

Eigenschaften:

1. Die Summe der Abweichungen einer Menge von Zahlen von ihrem arithmetischen Mittel ist gleich Null:

$$\sum_{j=1}^{n}(x_j - \bar{x}) = 0$$

2. Die Summe der quadratischen Abweichungen der Zahlen von ihrem arithmetischen Mittel ist kleiner als die Summe der quadratischen Abweichungen von einer beliebigen anderen Zahl:

$$\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2 \le \sum_{j=1}^{n}\left(x_j - a\right)^2 \qquad \text{für alle } a \in \mathbb{R}$$

3. Werden die Einzelwerte einer lineare Transformation unterzogen:

$$x_j^* = a \cdot x_j + b\,,$$

so unterliegt der Mittelwert der gleichen Transformation:

es gilt: $\bar{x}^* = \frac{1}{n}\sum_{j=1}^{n} x_j^* = a \cdot \bar{x} + b\,.$

Für die Streuung s^{*2} der transformierten Daten erhält man: $s^{*2} = a^2 \cdot s^2\,.$

Beispiel: Eine Autovermietung berechnet für ihre Wagen eine feste Tagesgebühr von $b = 40\ \text{DM}$ und einen Kilometersatz von $a = 0{,}40\ \text{DM/km}$; ferner sei bekannt, daß die Wagen täglich im Durchschnitt $\bar{x} = 250\ \text{km}$ zurücklegen bei einer Standardabweichung von s = 5 km. Die durchschnittlichen täglichen Einnahmen pro Wagen ergeben sich dann als

$$\bar{x}^* = a \cdot \bar{x} + b = 0{,}40 \cdot 250 + 40 = 140\ \text{DM}$$

bei einer Standardabweichung von

s*=a s =0,40 •5=2 DM.

4. Standardisierung: Die linearеTransformation $x_i^* = \frac{x_i - \bar{x}}{s}$ nennen wir Standardisierung.

Es gilt: $\bar{x}^* = 0$ und $s^{*2} = 1.$

Zur Berechnung von $\bar{x}$ und s:

5. Es berechnet sich leichter

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

Bei Verwendung gruppierter Daten, wo nur die Klassenmitten x_j' bekannt sind:

$$\bar{x}' = \frac{1}{n} \cdot \sum_{j=1}^{k} \left(x_j' \cdot H_n(K_j)\right)$$

$$s' = \sqrt{\frac{1}{n-1} \cdot \sum_{j=1}^{k} \left(x_j'^2 \cdot H_n(K_j) - n \cdot \bar{x}'^2\right)}$$

6. Man hat schon Werte $\bar{x}$ und s für $x_1,...,x_n$ ermittelt und kommt noch ein Wert x_{n+1} hinzu, so sind folgende Formeln nützlich:

$$\bar{x}_{n+1} = \frac{1}{n+1}\left(x_{n+1} + n \cdot \bar{x}\right) \qquad (1)$$

$$s_{n+1} = \sqrt{(n+1)\left(\bar{x}_{n+1} - \bar{x}\right)^2 + \frac{n-1}{n} \cdot s^2} \qquad (2)$$

7. Hat man für zwei Stichproben vom Umfang n_1 und n_2 Mittelwerte $\bar{x}_1$, $\bar{x}_2$ und Streuungen ${s_1}^2$, ${s_2}^2$ gegeben, so berechnen sich der Mittelwert und die Streuung der gemeinsamen Stichprobe vom Umfang $n = n_1 + n_2$ gemäß:

$$\bar{x} = \frac{\left(n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2\right)}{\left(n_1 + n_2\right)} \qquad (3)$$

$$s^2 = \frac{(n_1 - 1) \cdot {s_1}^2 + (n_2 - 1) \cdot {s_2}^2}{n_1 + n_2 - 1} + \frac{n_1\left(\bar{x}_1 - \bar{x}\right)^2 + n_2\left(\bar{x}_2 - \bar{x}\right)^2}{n_1 + n_2 - 1} \qquad (4)$$

Beispiel.: Ein Unternehmen stellt in zwei Betriebsteilen Reifen gleicher Sorte her. Es ist die durch schnittliche Lebensdauer von 100 Reifen zu bestimmen, wobei 20 im ersten und 80 im zweiten Betriebsteil untersucht werden. Für die beiden Betriebsteile ergab sich:

$$\bar{x}_1 = 4 \text{ Jahre} \quad \text{bzw.} \quad \bar{x}_2 = 5 \text{ Jahre}.$$

Die mittlere Lebensdauer der Gesamt-Stichprobe von 100 Reifen beträgt dann:

$$\bar{x} = \frac{(4 \cdot 20 + 5 \cdot 80)}{100} = 4{,}8 \text{ Jahre}.$$

Berechnen Sie die Streuung!

2.2.8 Anwendbarkeit der Maßzahlen für die verschiedenen Skalentypen

Maße	**Skala**			
	Nominalskala	**Ordinalskala**	**Intervallskala**	**Verhältnis-skala**
Lagemaße				
Modus	X	X	X	X
Median α-Quantile		X	X	X
Arithm. Mittel		(X)	X	X
Streuungsmaße				
Spannweite Quantilabstände		X	X	X
Mittlere absolute Abweichung vom Median			X	X
Streuung Standardabweichung			X	X
Variationskoeffizient				X

Stab-, Blakendiagramm — Klasseneinteilung Histogramm

Tab. 2.2: Anwendbarkeit der statischen Maßzahlen für verschiedene Skalentypen

a) Begründen Sie, warum der Variationskoeffizient nicht für intervallskalerte Merkmale anwendbar ist!

b) Begründen Sie, was man bei der Mittelwertberechnung bei ordinalskalierte Merkmale beachten muß! (Unter welchen Bedingungen können Mittelwerte verwendet werden?

2.3 Übungsaufgaben

1. Bestimmen Sie zu Aufgabe 4 der 1. Serie jeweils unter Verwendung der 30 Stichprobendaten und unter Verwendung der Klasseneinteilung
   a) den Anteil der Glühlampen, deren Lebensdauer zwischen 400 und 600 Stunden beträgt!
   b) Wieviele (Anzahl) besitzen eine Lebensdauer > 600 Stunden
   c) Wie hoch ist der Anteil der Lampen, deren Lebensdauer 400 Stunden nicht überschreitet?
   d) Welche Lebensdauer wird von 20 % der Lampen überschritten?

2. Berechnen Sie zu Aufgabe 4 der 1. Serie:
   a) den Mittelwert aus der Stichprobe („fein“) und mit der Klasseneinteilung. Vergleichen Sie beide Werte!
   b) die Streuung s^2 aus der Stichprobe und die Sheppard'sche Korrekturformel s_{**}^2 mit Hilfe der Klasseneinteillung! Vergleichen Sie beide Werte!
   c) den minimalen und maximalen Beobachtungswert, sowie den Median, das 0.25 und 0.75 %-Quantil aus der Stichprobe und unter Verwendung der Klasseneinteilung. Zeichnen Sie für beide Fälle (Stichprobe und Klasseneinteilung) den zugehörigen Boxplot! Ist die Verteilung links- oder rechtssteil?
   d) den Modalwert unter Verwendung der Klasseneinteilung!

3. Berechnen Sie für Aufgabe 2 der 1. Serie Mittelwert, Modus und Median der Kundenankünfte!

4. Zeigen Sie:
   a) Für die Beobachtungen einer diskreten Variablen X mit den verschiedenen Realisierungen $a_1, a_2,...,a_k$ gilt:
   $$\sum_{i=1}^{k} (a_i - \bar{x}) \cdot H_n(a_i) = 0 \ !$$
   b) Bei der Transformation (Standardisierung) von Daten $x_1,...,x_n$ gemäß $x_i^* = \frac{x_i - \bar{x}}{s}$ gilt für den Mittelwert und Streuung der transformierten Daten:
   $$\bar{x}^* = 0 \text{ und } s_*^2 = 1$$

5. In zwei Betriebsteilen eines Unternehmens werden Kühlaggregate gleicher Sorte hergestellt. Zur Untersuchung der Qualität der Produktion wurden Lebensdaueranalysen mit je einer Stichprobe vom Umfang m_1 im 1. Und m_2 im 2. Betriebsteil gemacht. Für die mittlere Lebensdauer und die Streuung ergab sich:

   $\bar{x}_1 = 30\ J, \quad s_1^2 = 25\ J^2, \quad m_1 = 20;$
   $\bar{x}_2 = 25\ J, \quad s_2^2 = 16\ J^2, \quad m_2 = 50;$

Wie groß sind Mittelwert, Streuung und Variationskoeffizient der gemeinsamen Stichprobe vom Umfang $n_1 + n_2$?

6. Lebensdauer zwischen 2 Motoren:

Typ 1

Jahre	$H_n(K_i)$
0 - 2	0
2 - 4	10
4 - 6	80
6 - 8	200
8 - 10	100
10 - 12	10
>12	0

Typ 2

Jahre	$H_n(K_i)$
0 - 2	0
2 - 4	50
4 - 6	200
6 - 8	50
> 8	0

a) Vergleichen Sie die Lebensdauern beider Typen auf geeignete Weise!
b) Berechnen Sie den Median, Modalwert, $x_{0,25}$, $x_{0,75}$, $\bar{x}$ von beiden Typen!
c) Wieviel % der Motoren vom Typ ½ haben eine Lebensdauer zwischen 3 und 7 Jahren?

7. Bei Piloten kommt es darauf an, daß sie möglichst schnell auf optische Signale reagieren.

Um die Reaktionszeit zu verbessern, wurde eine Trainingsmethode entwickelt.

Eine Untersuchung an 10 Personen ergab folgende Reaktionszeiten vor Anwendung und nach Anwendung der Trainingsmethode (in Sekunden):

Person	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Vorher	14	9	57	15	19	38	43	29	16	17
Nachher	10	11	20	10	15	30	32	30	16	17

Untersuchen Sie mit einer geeigneten Methode der deskriptiven Statistik, ob das Training eine Verbesserung der Reaktionszeit (zumindest in der Stichprobe) bewirkt!
Fertigen Sie gegebenenfalls Zeichnungen an und begründen Sie Ihre Aussagen!

3 Zweidimensionale Verteilungen - Zusammenhangsmaße

Beispiel 12: **Frage:** Ist die Reaktionszeit in einem bestimmten Experiment vom Geschlecht unabhängig?

Gg der Objekte: Gesunde Personen zwischen 30 und 40 Jahren

Merkmale: T = Reaktionszeit (ms),
S = Geschlecht (0 – männlich, 1– weiblich)
Z = (T, S) – 2-dimensionales Merkmal, T stetig, S-diskret, nominal

Stichprobe: Testen 200 Probanden,
$z_i = (t_i, s_i)$ – Merkmalsausprägung von Proband i , i = 1,...,200

Erfassung der Ergebnisse in einer 2-dimensionalen Häufigkeitstabelle:

T stetig ⇒ Reaktionszeit wird in Klassen eingeteilt:
K_1 = [100, 130) ms, K_2 = [130, 160) ms, K_3 = [160, 190) ms

T \ S	0 m	1 w	∑
[100, 130)	20	40	60
[130, 160)	50	60	110
[160, 190)	10	20	30
∑	80	120	200

Absolute Zell-Häufigkeiten

T \ S	0 m	1 w	∑
[100, 130)	0,1	0,2	0,3
[130, 160)	0,25	0,3	0,55
[160, 190)	0,05	0,1	0,15
∑	0,4	0,6	1

Relative Zell-Häufigkeiten

Tab.: 3.1: Zweidimensionale Häufigkeitstabellen von Reaktionszeit T und Geschlecht S

Verteilungen der Reaktionszeiten in den Gruppen m und w bezogen auf die (verschiedenen) Beobachtungszahlen aus den Gruppen:

T :	K_1	K_2	K_3
m :	20/80	50/80	10/80
w :	40/120	60/120	20/120
∑	60/200	110/200	30/200

T :	K_1	K_2	K_3
m :	0,25	0,625	0,125
w :	0,33	0,5	0,17
∑	0,3	0,55	0,15

Diese Verteilungen sind in etwa gleich.
Die Reaktionszeit kann man als unabhängig vom Geschlecht beobachten.

Ziel: Statistische Maßzahlen zur Messung der Abhängigkeit/Unabhängigkeit zwischen 2 Merkmalen entwickeln.

3.1 Zweidimensionale Häufigkeitsverteilungen

Sei Z = (X, Y), $X \in X$ und $Y \in Y$ ein 2-dimensionales Merkmal

1.Fall: X, Y diskret: $X = \{a_1,...,a_k\}$, $Y = \{b_1,...,b_j\}$

Wertebereich von Z: $Z = X \times Y = \{c_{ij} = (a_i, b_j) / a_i \in X, \ b_j \in Y\}$

Wir beobachten Z an n Objekten:

$z_i = (x_i, y_i)$, $i = 1,...,n$ – Stichprobe vom Umfang n

Wie im eindimensionalen Fall besteht zunächst die elementare statistische Tätigkeit darin, die Häufigkeitsverteilung der Stichprobe von Z zu ermitteln, d.h. auszuzählen, wieviele Beobachtungspaare (x_v, y_v), v = 1,...,n jeweils auf eine Merkmalsausprägung $c_{ij} = (a_i, b_j)$ von Z fallen.

Bezeichnungen:

(a_i, b_j) - *Zelle*

H_{ij} := Anzahl aller Paare (x_v, y_v) mit $x_v = a_i$ und $y_v = b_j$
– *absolute „Zellhäufigkeit"*

h_{ij} := $\frac{H_{ij}}{n}$ - Anteil aller Paare (x_v, y_v) mit $x_v = a_i$ und $y_v = b_j$
– *relative „Zellhäufigkeit"*

2. Fall: X stetig oder Y stetig

Wir machen eine Klasseneinteilung von X bzw. Y

K_i^x i = 1,...,k – Klasseneinteilung von X
K_j^y j = 1,...,l – Klasseneinteilung von Y

$\left(K_i^x, K_j^y\right)$ – *Zelle*

H_{ij} – Anzahl aller (x_v, y_v) mit $x_v \in K_i^x$ und $y \in K_j^y$ – **absolute Zellhäufigkeit**

h_{ij} – Anteil aller (x_v, y_v) mit $x_v \in K_i^x$ und $y \in K_j^y$ – **relative Zellhäufigkeit**

Def.: Die Gesamtheit aller absoluten bzw. relativen Häufigkeiten

$\left(\left(H_{ij} \quad, i = 1,...,k;\ j = 1,...,l\right)\right)$ bzw. $\left(\left(h_{ij} \quad, i = 1,...,k;\ j = 1,...,l\right)\right)$

heißt zweidimensionale *absolute* bzw. *relative Häufigkeitsverteilung von Z*.

Kontingenztabelle:

X \ Y	K_1^Y b_1	K_2^Y b_2 ..... K_j^Y b_j	K_l^Y b_l	Zeilensumme Σ
K_1^X a_1	H_{11}	H_{12} ..... H_{1j}	H_{1l}	$H_{1\bullet}$
K_2^X a_2	H_{21}	H_{22} ..... H_{2j}	H_{2l}	$H_{2\bullet}$
K_i^Y a_i	H_{i1}	H_{i2} ..... H_{ij}	H_{il}	$H_{i\bullet}$
K_k^X a_k	H_{k1}	H_{k2} ..... H_{kj}	H_{kl}	$H_{k\bullet}$
Σ Spaltensumme	$H_{\bullet 1}$	$H_{\bullet 2}$ ..... $H_{\bullet j}$	$H_{\bullet l}$	n

Tab. 3.2

Bezeichnungen:

$H_{i\bullet} = \sum_{j=1}^{l} H_{ij}$ – i-te Zeilensumme

$H_{\bullet j} = \sum_{i=1}^{k} H_{ij}$ – j-te Spaltensumme

$H_{\bullet j}$, $H_{i\bullet}$ – Randsummen

$H_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{l} H_{ij} = n$ – Stichprobenumfang

Def.: Die Gesamtheit der Randsummen bilden die Häufigkeitsverteilungen von X bzw. Y:

$\left(\left(H_{i\bullet} \quad , i = 1,\ldots,k\right)\right)$ - Häufigkeitsverteilung von X

$\left(\left(H_{\bullet j} \quad , j = 1,\ldots,l\right)\right)$ - Häufigkeitsverteilung von Y

} *Randverteilung*

Bemerkung: Betrachten wir h_{ij} statt H_{ij}, so erhalten wir die Randsummen $h_{i\bullet}$, $h_{\bullet j}$. Diese bilden die relative Häufigkeitsverteilung von X bzw. Y.

Beispiel 13: Auf einer Hühnerfarm seien zufällig 1000 Eier ausgewählt, vermessen und gewogen worden. Die Merkmale sind X = Länge (cm) und Y = Breite (cm)

Frage: Gibt es einen Zusammenhang zwischen Länge und Breite?

Objekte: Eier dieser Farm
Merkmal: Z = (X, Y) = 2-dimensionales Merkmal

X, Y stetig $\Rightarrow$ für beide Merkmale wird eine Klasseneinteilung gemacht.

Häufigkeitstabelle der 1000 Beobachtungen:

Y / X	$2 < Y \leq 3$	$3 < Y \leq 4$	$4 < Y \leq 5$	$5 < Y \leq 6$	Summe
$4 < X \leq 5$	10	66	0	0	76
$5 < X \leq 6$	5	250	157	0	412
$6 < X \leq 7$	0	47	380	29	456
$7 < X \leq 8$	0	0	15	41	56
Summe	15	363	552	70	1000

Zweidimensionale Häufigkeitsverteilung für (X_g, Y_g)

Übliche grafische Darstellung zweidimensionaler Verteilungen:

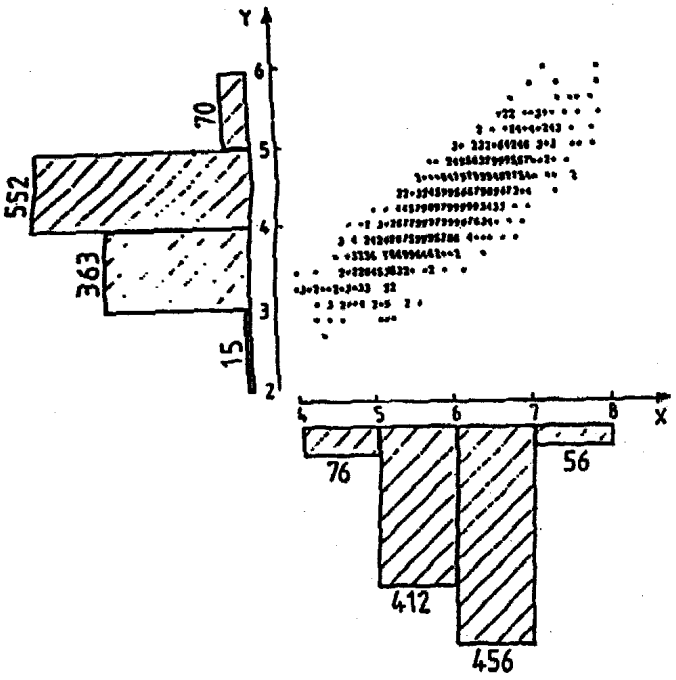
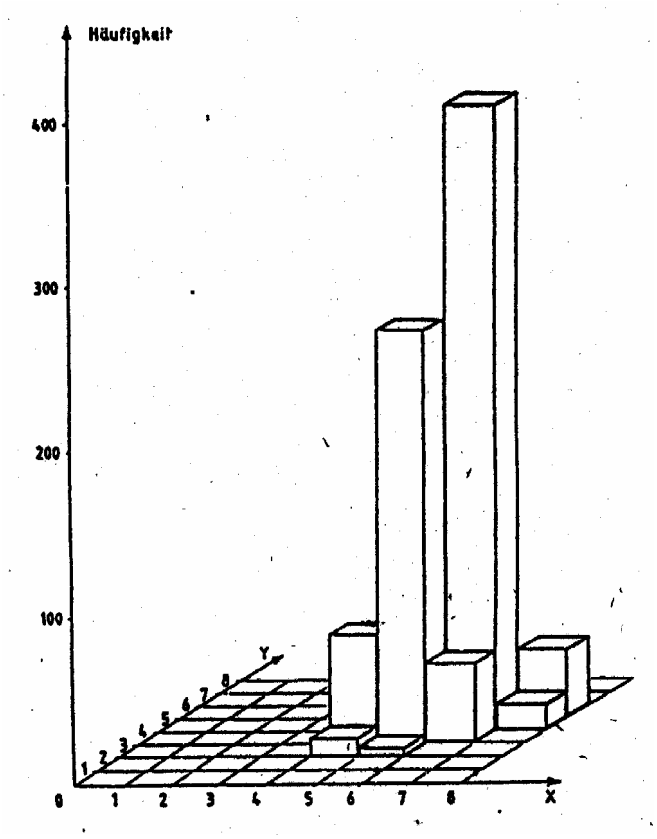


Abb.: Zweidimensionaler Plot



3.2 Messung der Unabhängigkeit zweier Merkmale / der χ^2-Kontingenzkoeffizient

Im Beispiel 12 hatten wir die Verteilung der Reaktionszeit jeweils in den Gruppen S = männlich und S = weiblich betrachtet.

Bezeichnung:

$h_n = (X = a_i) := \frac{H_{i\bullet}}{n}$ – relative Häufigkeit von „X = a_i“

$h_n(X = a_i / Y = b_j) := \frac{H_{ij}}{H_{\bullet j}}$ – bedingte relative Häufigkeit von „X = a_i“ unter der Bedingung, daß „Y = b_j“ ist

Def.: Die Gesamtheit der bedingten relativen Häufigkeiten $\left(\left(\frac{H_{ij}}{H_{\bullet j}} \quad , i = 1,...,k\right)\right)$ heißt *bedingte relative Häufigkeitsverteilung* von X unter der Bedingung „Y = b_j“ (bzw. $Y \in K_j^Y$).

Für unser Beispiel 12:

(0.25, 0.625, 0.125) – bedingte Verteilung der Reaktionszeit unter der Bedingung, daß das Geschlecht männlich ist

(0.33, 0.5, 0.17) – bedingte Verteilung der Reaktionszeit unter der Bedingung, daß das Geschlecht weiblich ist

Bei Unabhängigkeit von X und Y:

Verteilung von X hängt nicht von Y ab,

bzw. bedingte Verteilung von X ≈ unbedingte Verteilung von X

bzw. $h_n\left(X = a_i \,/\, Y = b_j\right) \approx h_n\left(X = a_i\right)$ für alle i = 1,...,k ; j = 1,...,l

bzw. $\frac{H_{ij}}{H_{\bullet j}} \approx \frac{H_{i\bullet}}{n}$ für alle i = 1,...,k ; j = 1,...,l

bzw. $H_{ij} \approx \frac{H_{i\bullet} \cdot H_{\bullet j}}{n}$ für alle i = 1,...,k ; j = 1,...,l

Bezeichnung:

$H_{ij}^{(E)} := \frac{H_{i\bullet} \cdot H_{\bullet j}}{n}$ erwartete Zell-Häufigkeit bei Unabhängigkeit

$H_{ij}^{(B)} := H_{ij}$ beobachtete Zell-Häufigkeit

Methode:

Wir stellen $H_{ij}^{(E)}$ und $H_{ij}^{(B)}$ gegenüber,

falls $H_{ij}^{(E)} \approx H_{ij}^{(B)}$ $\forall i \forall j$ $\Rightarrow$ X und Y unabhängig

Zu Beispiel 12:

S / T	m		w		Σ
[100, 130)	$\frac{20}{24}$	(4)	$\frac{40}{36}$	(4)	60
[130, 160)	$\frac{50}{44}$	(6)	$\frac{60}{66}$	(6)	110
[160, 190)	$\frac{10}{12}$	(2)	$\frac{20}{18}$	(2)	30
Σ	80		120		200

$$H_{11}^{(E)} = \frac{H_{1\bullet} \cdot H_{\bullet 1}}{n} = \frac{60 \cdot 80}{200} = 24$$

$$H_{12}^{(E)} = \frac{H_{1\bullet} \cdot H_{\bullet 2}}{n} = \frac{60 \cdot 120}{200} = 36$$

$$H_{21}^{(E)} = \frac{H_{2\bullet} \cdot H_{\bullet 1}}{n} = \frac{110 \cdot 80}{200} = 44 \qquad \text{usw.}$$

Frage: Wie sind die Differenzen ( ) einzuschätzen (groß oder klein)?

⇒ Statistische Maßzahl muß her!

Def.:

$$\chi^2 = CHI^2 = \sum_{j=1}^{l} \sum_{i=1}^{k} \frac{\left(H_{ij}^{(E)} - H_{ij}^{(B)}\right)^2}{H_{ij}^{(E)}}$$

heißt *χ^2-Kontingenzkoeffizient*

Bei Unabhängigkeit gilt: $\chi^2 \approx 0$

Im Beispiel 12:

$$X^2 = \frac{4^2}{24} + \frac{4^2}{36} + \frac{6^2}{44} + \frac{6^2}{66} + \frac{2^2}{12} + \frac{2^2}{18} = \underline{\underline{3{,}03}}$$

Ist das „groß“ oder „klein“? ⇒ Normiertes Maß muß her!

Def.:

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

Eigenschaften: $C = 0$ bei Unabhängigkeit

aber: $C < 1$ $\Rightarrow$ C eignet sich nicht als Maß

Def.: $C_{korr} = C \cdot \sqrt{\frac{q^*}{q^* - 1}}$ mit $q^* = \min(l, k)$ heißt *Korrigierter Kontingenzkoeffizient*

Es gilt: $0 \le C_{korr} \le 1$

$C_{korr} \approx 0$ falls Unabhängigkeit

Standard:

$0 \le C_{korr} \le 0{,}2$ $\Rightarrow$ unabhängig

$0{,}2 < C_{korr} \le 0{,}5$ $\Rightarrow$ schwach abhängig

$0{,}5 < C_{korr} \le 1$ $\Rightarrow$ abhängig

Zum Beispiel 12:

$$C = \sqrt{\frac{3{,}03}{203{,}03}} \cdot \sqrt{2}$$

$= 0{,}173$ $\Rightarrow$ T unabhängig von S

Beantworten Sie unter Benutzung nachfolgender Häufigkeitstabelle die Frage, ob das Wahlverhalten der Saarländer abhängig vom Geschlecht ist?
(Verwenden Sie als Maßzahl den korrigierten Kontingenzkoeffizienten!)

Häufigkeitsverteilung des Wahlverhaltens und des Geschlechts von 100 Saarländern:

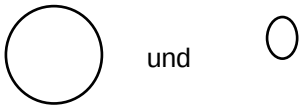
Wahlverhalten/ Geschlecht	**männlich**	**weiblich**	
CDU/CSU	25	10	35
SPD	15	10	25
FDP	5	5	10
B90/Grüne	10	10	20
Sonstige	5	5	10
	60	40	100

3.3 Korrelationsmaße

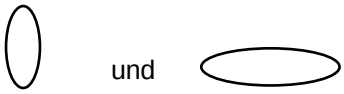
3.3.1 Der Pearsonsche Korrelationskoeffizient

Beispiel 14: X – Länge von Eiern
Y – Breite von Eiern

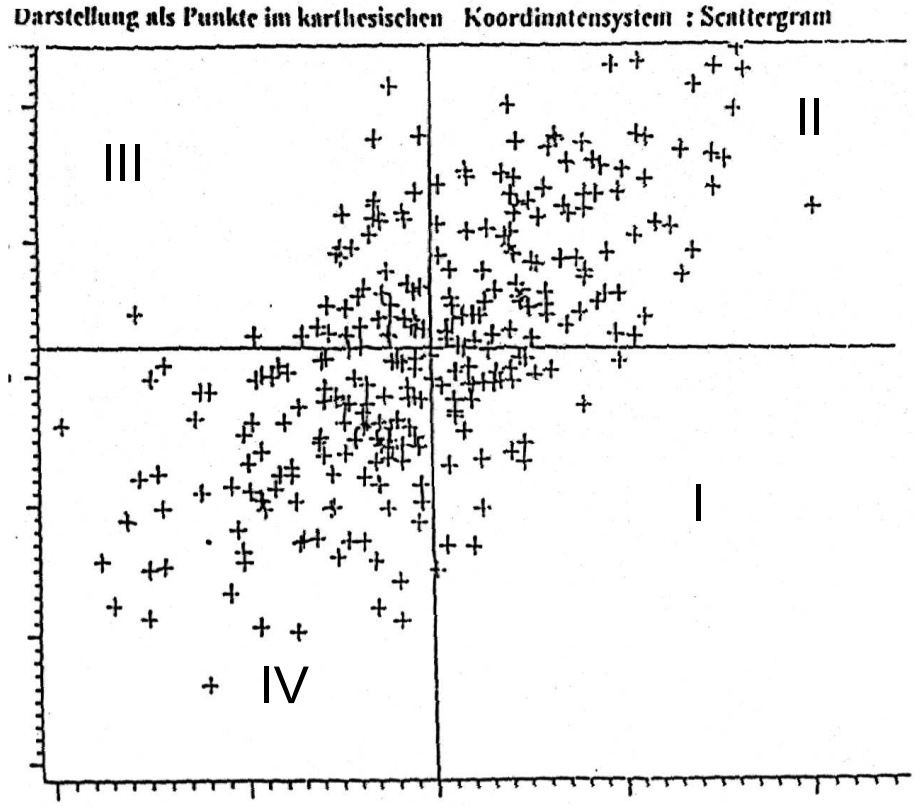
Frage: Hängen Länge und Breite zusammen?
Sind die Eier mehr so:



oder mehr so:



Eine Stichprobe von 1000 Eiern lieferte folgende Beobachtungen (x_i, y_i), $i = 1,...,1000$:



$(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$ $\Rightarrow$ (x_i, y_i) liegt im I. oder III. Quadranten

$(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$ $\Rightarrow$ (x_i, y_i) liegt im II. oder IV. Quadranten

Statistische Maßzahlen für diesen Sachverhalt:

Seien:

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

– arithmetische Mittel der Stichprobe bzgl. x_i und y_i

Def.: $$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$ – heißt Stichprobenkovarianz von X und Y bzgl. der Stichprobe (x_i, y_i), i = 1,...,n

Bemerkung:

Liegen die Daten hauptsächlich im I. und III. Quadranten $\Rightarrow S_{xy} < 0$

II. und IV. Quadranten $\Rightarrow S_{xy} > 0$

X, Y, unabhängig $\Rightarrow$ gleich viele im I., II., III., IV. Quadranten $\Rightarrow S_{xy} = 0$

Problem:

Beurteilung der Größe von S_{xy}. In unserem Beispiel ergibt sich $S_{x\,y} = 1{,}2$.

Ist S = 1,2 groß oder klein?

$\Rightarrow$ S normieren!!

Bemerkung:

$s_x{}^2 := S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ – Streuung (Stichprobenvarianz) von X

$s_y{}^2 := S_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ – Streuung (Stichprobenvarianz) von Y

Def.: Das Maß

$$r = \frac{S_{xy}}{\sqrt{s_x{}^2 \cdot s_y{}^2}}$$

heißt *Pearsonscher Korrelationskoeffizient* von X und Y bzgl. der Stichprobe $\left(x_i, y_i\right),\ i = 1,\ldots,n$.

Satz: Es gilt:

1. $-1 \le r \le 1$ (r ist normiert)
2. $r = \begin{cases} 1 & \text{falls } y_i = a \cdot x_i + b,\quad a > 0 \\ -1 & \text{falls } y_i = a \cdot x_i + b,\quad a < 0 \end{cases}$

Standard: Bezeichnungen

$r < -0{,}2$ $\Rightarrow$ Daten sind ***negativ korreliert*** (große x $\Leftrightarrow$ kleine y-Werte)
(I., III. Quadrant)

$r > +0{,}2$ $\Rightarrow$ Daten sind ***positiv korreliert*** (große x $\Leftrightarrow$ große y-Werte)
(II., IV Quadrant)

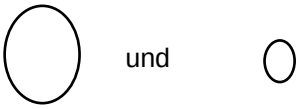
„Feinabstufung“:

1. $0 \le |r| \le 0{,}2$ ***unkorreliert***

2.	$0{,}2 < \lvert r \rvert \leq 0{,}5$	***schwach korreliert***
3.	$0{,}5 < \lvert r \rvert \leq 0{,}8$	***korreliert***
4.	$0{,}8 < \lvert r \rvert \leq 1$	***hoch korreliert***

Für unser Beispiel gilt: r = 0,83

⇒ Länge und Breite sind hoch positiv korreliert (siehe Scattergramm: Dick Eier sind lang, dünne Eier sind kurz

(die Eier sind  und

und nicht  und )

Typische „Punktwolke“ für Korrelationskoeffizienten r:

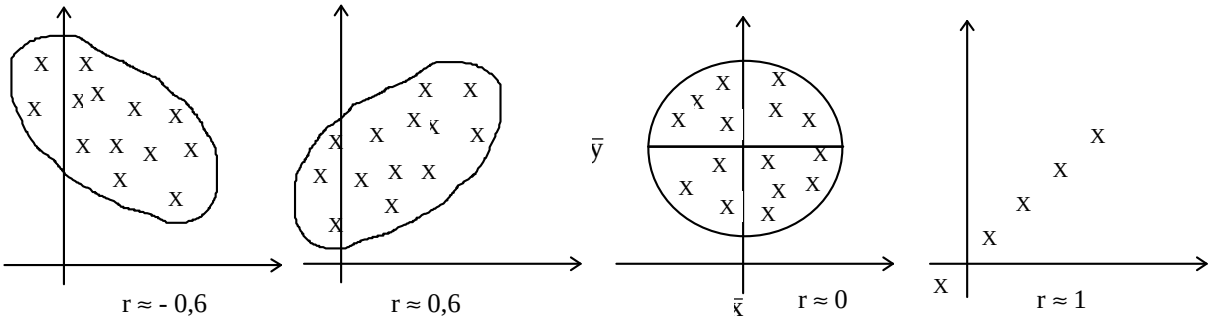


Abb. 3.3

Bemerkungen:

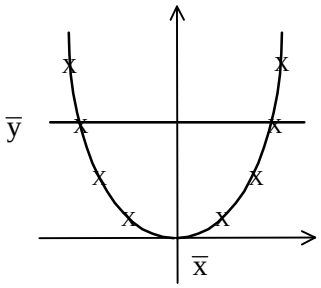
1.) Voraussetzung: X, Y stetig (mindestens intervallskaliert)

2.) Sei $Y = X^2$

→ Für alle Beobachtungen (x_i, y_i), i = 1,...,n gilt:

$y_i = x_i^2$ (Die Punkte (x_i, y_i) liegen auf einer Parabel.)

In diesem Fall ist r≈0 !!



$\Rightarrow S_{xy} \approx 0$

$\Rightarrow r \approx 0$

$\Rightarrow$ r ≈ 0 bedeutet nicht „Unabhängigkeit“ sondern nur „Unkorreliertheit“ (d.h. X und Y weichen „total“ vom linearen Zusammenhang ab, aber sie können daraus auf eine andere Weise funktional zusammenhängen)
Der Pearson'sche Korrelationskoeffizient r mißt nur den Grad des linearen Zusammenhangs.

3.3.2 Der Spearman'sche Korrelationskoeffizient für ordinal skalierte Daten

Wenn mindestens ein Merkmal ordinal skaliert ist, ist r nicht anwendbar. Solche Fragestellungen treten oft auf. Wir wenden in diesen Fällen den sogenannten *Korrelationskoeffizienten* ρ von Spearman an.

Beispiele:

a) Gibt es einen Zusammenhang zwischen Mathematik- und Englischleistungen?
b) A-, B- Note beim Eiskunstlauf (Gibt es einen Zusammenhang?)
   $\Rightarrow$ X, Y ordinal skaliert
c) Zusammenhang zwischen Körpergröße und Bildungsgrad?

Der Spearman'sche Korrelationskoeffizient ρ:

Seien X und Y zwei Zufallsgrößen, mindestens ordinal skaliert.
Der Spearman'sche Korrelationskoeffizient erfordert die Zuordnung von sogenannten Rangplätzen $R[x_i]$ zu Stichprobenwerten x_i, i = 1,...,n.

Rangplatzbestimmung:

Sei $x_1,...,x_n$ eine Stichprobe vom Umfang n und $x_{[1]} \le x_{[2]} \le \cdots \le x_{[n]}$ die geordnete Stichprobe.
Unter dem Rangplatz $R[x_i]$ von x_i versteht man den Platz von x_i in der geordneten Stichprobe.

Die Ordnung der Stichprobendaten x_i ergibt sich aus der Ordnung der Merkmalsausprägungen, denen die x_i zugeordnet sind. Dabei ordnet man die Merkmalsausprägungen so, daß der in der Ordnung an erster Stelle stehende Merkmalswert der "beste", der an zweiter Stelle stehende Merkmalswert der "zweitbeste" usw. sind.
Haben mehrere Daten x_i den gleichen Rangplatz (d.h. den gleichen Wert), so wird ihnen der mittlere Rangplatz zugewiesen:

Beispiel 15: Merkmal: Qualität

Ausprägung (Qualitätsklasse)	A	B	C	D	E	F	G
durch Sklaierung zugeordnete Daten: x_i	2	3	4	5	6	7	100

x_i	2	3	7	3	5	3	100
Platz in geordneter Stichprobe	1	2	6	3	5	4	7
$R[x_i]$	1	3	6	3	5	3	7

$R[3] = \frac{2+3+4}{3}$ – mittlerer Rangplatz des Werts 3 in der geordneten Stichprobe

Sei (x_i, y_i) , i = 1,...,n eine Stichprobe und seien $R[x_i]$, $R[y_i]$ die den Stichprobenwerten x_i und y_i zugeordneten Rangplatzzahlen.

Der Spearman'sche Korrelationskoeffizient ist als Pearson'scher Korrelationskoeffizient der Rangplätze definiert.

Def.: Der Koeffizient

$$\rho := r_{R_{[x]},R_{[y]}} = \frac{S_{R_{[x]},R_{[y]}}}{\sqrt{s^2_{R_{[x]}} \cdot s^2_{R_{[y]}}}}$$

heißt *Spearman'scher Korrelationskoeffizient* (ρ := Pearson'scher Koeffizient für die Rangplatzzahlen) ⇒ „Rangplatzkorrelationskoeffizient"

Zur Erleichterung der Brechnung des Spearman'sche Korrelationskoeffizienten kann man folgenden Satz anwenden:

Satz: Es gilt: $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}$, $d_i = R_{[x_i]} - R_{[y_i]}$

Berechnung des Spearman'schen Korellationskoeffizienten:

1.) Man ordnet jedem x_i und y_i einen Rangplatz $R[x_i]$, $R[y_i]$ zu.

2.) Man bildet die Rangplatzdifferenzen $|d_i| := |R[x_i] - R[y_i]|$, d_i^2 und $\sum d_i^2$.

3.) Man berechnet den Spearman'schen Korrelationskoeffizienten ρ.

x	x_1	x_i	x_n
y	y_1	y_i	y_n

R[x] **R[y]**	$R[x_1]$ $R[y_1]$	.....	$R[x_i]$ $R[y_i]$	..... $R[x_n]$ $R[y_n]$	
$\|d_i\|$	$\|d_1\|$	.....	$\|d_i\|$	..... $\|d_n\|$	$\sum$
d_i^2	d_1^2		d_i^2	d_n^2	$\sum$ **d_i^2**

Tab. 3.6

Auswertung von ρ: $-1 \le \rho \le 1$

(wie bei r !)

$\lvert\rho\rvert \le 0{,}2$	$\Rightarrow$ unkorreliert	negative und positive Korrelation
$\lvert\rho\rvert \le 0{,}5$	$\Rightarrow$ schwach korreliert	
$\lvert\rho\rvert \le 0{,}8$	$\Rightarrow$ korreliert	
$1 \ge \lvert\rho\rvert > 0{,}8$	$\Rightarrow$hoch korreliert	

Beispiel 16: Noten: Gibt es einen Zusammenhang zwischen den Leistungen in Mathe und in Englisch?

Person	1	2	3	4	5	6	
Mathe	1 1	5 4	4 3	6 4	2 2	3 3	Noten
Englisch	6 10	5 14	1 19	3 16	4 14	2 19	Punkte
$R_{[Mathe]}$	1	5,5	3,5	5,5	2	3,5	
$R_{[Englisch]}$	6	4,5	1,5	3	4,5	1,5	
| di |	5	1	2	2,5	2,5	2	$\sum$
d_i^2	25	1	4	6,25	6,25	4	**46,5**

$$\rho = 1 - \frac{6 \cdot 46{,}5}{6 \cdot 35} = 1 - \frac{46{,}5}{35} = \underline{\underline{-0{,}32}}$$

Mathe- und Englischleistungen sind schwach negativ korrelliert!

Beispiel 17: Eiskunstlauf: Gibt es einen Zusammenhang zwischen A und B-Note?

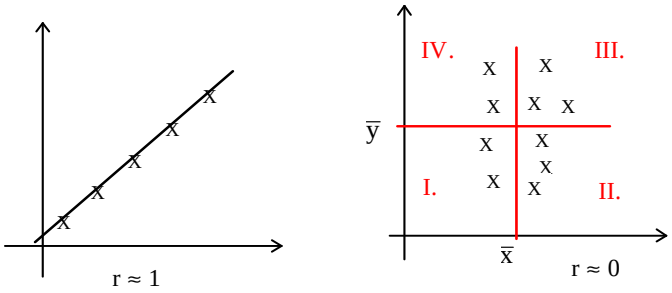
Läufer	1	2	3	4	5	6	7	8	9	
A-Note	5,3	5,6	5,0	5,3	4,9	4,6	5,3	5,0	5,2	
B-Note	5,4	5,4	5,1	5,2	5,0	4,5	5,5	4,8	5,1	
Rangplatz A	3	1	6,5	3	8	9	3	6,5	5	
Rangplatz B	2,5	2,5	5,5	4	7	9	1	8	5,5	
$\|d_i\|$	0,5	1,5	1	1	1	0	2	1,5	0,5	$\sum$
d_i^2	0,25	2,25	1	1	1	0	4	2,25	0,25	12

$$\Rightarrow \rho = 1 - \frac{6 \cdot 12}{9 \cdot (81-1)} = 1 - \frac{6 \cdot 12}{9 \cdot 80} = 0{,}9$$

$\Rightarrow$ A und B sind hochkorreliert!

3.4 Regressionsanalyse (Ausgleichsrechnung)

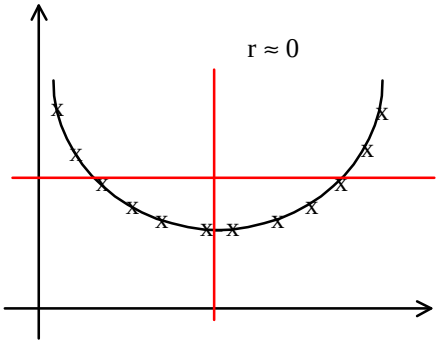
3.4.1 Das Regressionsproblem



r mißt nur den linearen Zusammenhang:

Konsequenz: Sei

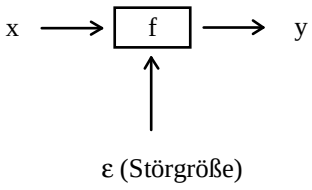
$$y = f(x), \quad \text{und sei f keine Gerade, sei z.B.}$$
$$y = ax^2 + bx + c \qquad (x_i, y_i), \quad i = 1, \ldots, n$$



⇒ Ich erkenne mit r keinen Zusammenhang, obwohl es einen gibt (weil f nicht linear ist)

Aufgabenstellung der Regressionsanalyse:

Schätzung einer Funktion f, die den Zusammenhang zwischen x und y beschreibt und nur fehlerbehaftet beobachtet werden kann:
Modellannahme: $y = f(x) + \varepsilon$



Ziel: y vorhersagen können anhand der Werte von x.
→ f(x) durch eine Funktion $\hat{f}(x)$ schätzen (approximieren) ($y \approx \hat{f}(x)$)

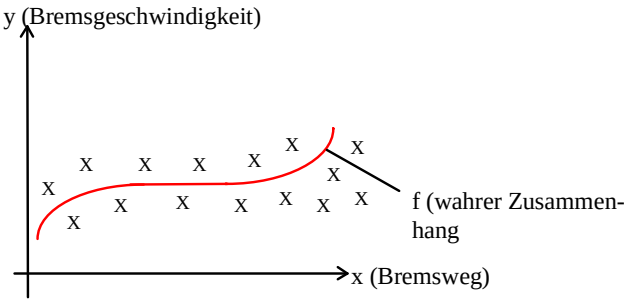
Beispiele:

X	Y	ε
Bremsweg	Bremsgeschwindigkeit	Straßenverhältnisse, Zustand der Bremsen, Reaktionsverhalten
Geschwindigkeit	Benzinverbrauch	.....
Kohlenstoffgehalt einer Stahlsorte	Reißfestigkeit	.....
Durchmesser eines Flaschenhalses	Berstdruck der Flasche	.....
Körpergröße	Intelligenz	.....
Rißthöhe beim Rind	Gewicht	.....

	Geben Sie in der Tabelle Beispiele für die Störgröße ε an!

Verallgemeinerung: Gesucht ist eine Funktion f in mehreren Einflußgrößen mit
$y = f(x_1, x_2, ..., x_k) + \varepsilon$
Wir betrachten hier nur den einfachen Fall einer Einflußgröße, d.h., den Fall k=1.

Methode zur Bestimmung der Schätzung $\hat{f}$ für f:



1. Bestimme Beobachtungspaare (x_i, y_i), $i = 1, \ldots, n$ und trage sie in ein Koordinatensystem ein!
2. Lege anhand der 'Punktwolke' einen Ansatz für den Funktionstyp von f fest!

Hier z.B.: Polynom 2. oder 3. Grades:

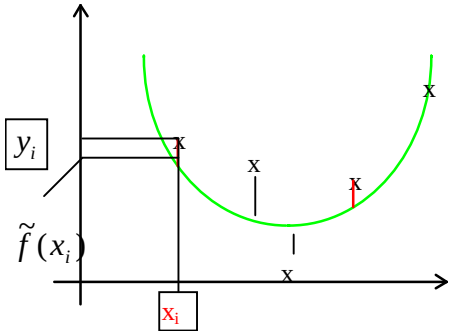
$\tilde{f}(x) = a_0 + a_1 x + a_2 x^2$ Polynom 2. Grades oder

$\tilde{f}(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ Polynom 3. Grades

Allgemeine Form für solche Ansätze in der ***linearen Regressionsanalyse***:

$\tilde{f}(x) = \sum_{\nu=0}^{m} a_\nu g_\nu(x)$, wobei $g_\nu(x)$ bekannte Funktionen sind; $a_\nu \in \mathbb{R}$ unbekannte, noch zu bestimmende Parameter; $m \in N$ fest.

3. Sei $\Im_m = \left\{ \tilde{f}_m \mid \tilde{f}_m(x) = \sum_{\nu=0}^{m} a_\nu g_\nu(x),\ a_\nu \in \Re \right\}$ die Menge aller Funktionen, die dem Ansatz genügen.
Suche nun in der Menge $\Im_m$ aller Funktionen, die diesem Ansatz genügen, die beste!



Polynom 2. Grades

Gütekriterium: Wir suchen die Funktion $\hat{f}_m(x)$ unseres Ansatzes, die zu allen Beobachtungen y_i den kleinsten Abstand hat, d.h. für die gilt:

$$\sum_{i=1}^{n} \left(y_i - \hat{f}_n(x_i) \right)^2 = \min_{\tilde{f}_m \in \Im_m} \sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i) \right)^2$$

Def.: $u_i = y_i - \tilde{f}_m(x_i)$ heißt *Residuum* an der Stelle x_i.

Die Summe $\sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i) \right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\sum_{\nu=1}^{m} a_\nu g_\nu(x_i) \right) \right)^2$ heißt *Residual Sum of Squares (RSS)* von $\tilde{f}_m(x_i)$.

Die Lösung $\hat{f}_m(x)$ des Extremwertproblems $\min_{\tilde{f}_m \in \Im_m} \sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i) \right)^2$ heißt *lineare Regressionsfunktion*.

Die Bestimmung der Regressionsfunktion $\hat{f}(x)$ erfordert die Lösung eines Extremwertproblems einer Funktion in mehreren Variablen:

$$\min_{\tilde{f}_m \in \Im_m} \sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i)\right)^2 = \min_{(a_0,\ldots,a_m)} \underbrace{\sum_{i=1}^{n} \left(y_i - \left(\sum_{\nu=0}^{m} a_\nu g_\nu(x_i) \right) \right)^2}_{\chi(a_0,a_1,\ldots,a_m)}$$

$$= \min_{(a_0,a_1,\ldots,a_m)} F(a_0, a_1, \ldots, a_m)$$

$\Rightarrow$ Bei unserem Extremwertproblem handelt es sich also um die Extremwertbestimmung einer Funktion (F) in mehreren Veränderlichen $(a_0, \ldots, a_m)$.

3.4.2 Extremwertbestimmung einer Funktion in mehreren Veränderlichen - Exkurs

3.4.2.1 Partielle Ableitungen

Sei $f: \Re^n \to \Re$ eine Funktion in n Veränderlichen, $f(x_1, \ldots, x_n) = y$.

Def.: Die Ableitung von f nach einer Komponente x_i (die anderen Variablen x_j, $j \neq i$ werden als Konstanten verarbeitet) heißt (*erste) partielle Ableitung* von f nach x_i.

Bezeichnung: $\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)$ oder $f_{xi}((x_1, \ldots, x_n)$

Beispiel 18: $f(x,y) = x^2 + y \cdot e^{2x}$

$$\frac{\partial f}{\partial x}(x,y) = 2x + 2y \cdot e^{2x}$$

$$\frac{\partial f}{\partial y}(x,y) = e^{2x}$$

Die partiellen Ableitungen können wiederum abgeleitet werden:

$\frac{\partial^2 f}{\partial x_i \partial x_j}(x_1, \ldots, x_n)$ heißt *zweite partielle Ableitung* von f nach x_j.

Bezeichnung: $\frac{\partial^2 f(x_1,\ldots,x_n)}{\partial^2 x_i} := \frac{\partial^2 f}{\partial x_i \partial x_i}(x_1, \ldots, x_n)$

Beispiel 19: $\frac{\partial^2 f_{(x,y)}}{\partial x \partial y} = 2 \cdot e^{2x}$, $\frac{\partial^2 f_{(x,y)}}{\partial^2 x} = 4 \cdot e^{2x}$

$\frac{\partial^2 f_{(x,y)}}{\partial y \partial x} = 2 \cdot e^{2x}$, $\frac{\partial^2 f_{(x,y)}}{\partial y^2} = 0$

	Was fällt Ihnen bei den beiden gemischten Ableitungen auf ?

Bemerkung:

Es gilt folgender Satz (von Schwarz): Die Reihenfolge der Ableitung ist bei partiellen Aleitungen der Ordnung >1 egal.

(Das bedeutet zum Beispiel, daß die Matrix aller zweiten partiellen Ableitungen einer Funktion symmetrisch ist !)

3.4.2.2 Extremwertberechnung einer Funktion in n Veränderlichen

Sei $f:\Re^n \to \Re$ eine Funktion in n Veränderlichen, y= $f(x_1,...,x_n)$ und seien

$\frac{\partial f}{\partial x_i}(x_1,...,x_n)$, i = 1,...,n, die erste Ableitung von f nach x_i und

$H = \left(\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x_1,...,x_n) \right) \right)_{i=1,...,n}^{j=1,...,n}$ die Matrix der zweiten partiellen Ableitung von f.

Def.: Eine Matrix H vom Typ (n, n) (quadratisch) heißt positiv (bzw. negativ) definit, falls $\forall \vec{h} \in \Re^n$, mit $\vec{h} \neq \vec{0}$ gilt:

$$\vec{h}^T H \vec{h} > 0 \quad (bzw. \vec{h}^T H \vec{h} < 0)$$

Bezeichnung: H positiv definit: $H \succ 0$
H negativ definit: $H \prec 0$

In allen anderen Fällen heißt H indefinit.

Satz: Sei $f:\Re^n \to \Re$ eine Funktion in n Veränderlichen für die die 1. und 2. partiellen Ableitungen

$$\frac{\partial f}{\partial x_i}(x_1,...,x_n) \quad , \quad i = 1,...,n \quad \text{und} \quad \frac{\partial^2 f}{\partial x_i \partial x_j}(x_1,...,x_n) \quad , \quad i,j = 1,...,n \,,$$

existieren.

Sei $H = \left(\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x_1,...,x_n) \right) \right)_{i=1,...,n}^{j=1,...,n}$ die Matrix der 2. partiellen Ableitungen.

Dann gilt:

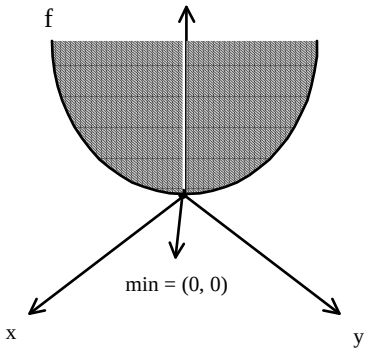
$x^* = \left(x_1^*, \ldots, x_n^*\right)$ ist Extremwert von f, falls gilt:

1.) $\frac{\partial f}{\partial x_i}\left(x^*\right) = 0$ für alle i = 1,...,n.

2.) H ist positiv oder negativ definit.

Gilt $H \succ 0$,so ist x^* Minimum von f und gilt $H \prec 0$,so ist x^* Maximum von f.

Beispiel 20: $f_{(x,y)} = x^2 + y^2$



1. Ableitung:

$$\left.\begin{aligned} \frac{\partial f_{(x,y)}}{\partial x} &= 2x \overset{!}{=} 0 \\ \frac{\partial f_{(x,y)}}{\partial y} &= 2y \overset{!}{=} 0 \end{aligned}\right\} x = y = 0$$

$(x^*, y^*) = (0, 0)$ ist extremwertverdächtig!

2. Matrix der zweiten Partiellen Ableitung bilden:

$$H := \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Sei $\vec{h} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \in \Re^2$

$\Rightarrow$

$$\vec{h}^T H \vec{h} = (h_1, h_2) \cdot \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

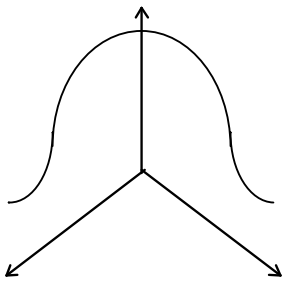
$$= (2h_1, 2h_2) \cdot \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$= 2h_1^2 + 2h_2^2$$

$$= 2 \cdot \left(h_1^2 + h_2^2\right) > 0 \qquad \forall \vec{h}, \vec{h} \neq 0!$$

$\Rightarrow H \succ 0$ $\Rightarrow$ (0, 0) <u>Minimum</u>

<u>Beispiel 21:</u> $f_{(x,y)} = e^{-\left(x^2+y^2\right)}$



1.

$$\frac{\partial f}{\partial x}(x,y) = -e^{-\left(x^2+y^2\right)} \cdot 2x \overset{!}{=} 0$$

$$\frac{\partial f}{\partial y}(x,y) = -e^{-\left(x^2+y^2\right)} \cdot 2y \overset{!}{=} 0$$

$\Rightarrow x = y = 0 \quad \left(\text{da } e^{-\left(x^2+y^2\right)} \neq 0\right) \quad \forall \; x, y$

$\Rightarrow x = y = 0$ extremwertverdächtig

2.

$$\frac{\partial^2 f}{\partial x^2}(0,0) = -2 \cdot e^{-\left(x^2+y^2\right)} + (-2x) \cdot \left(-e^{-\left(x^2+y^2\right)} \cdot 2x\right)\Big|_{0,0}$$

$$= -2$$

$$\frac{\partial^2 f}{\partial y^2}(0,0) = -2$$

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = -2x \cdot \left(-2y \cdot e^{-\left(x^2+y^2\right)} \right) \Big|_{(0,0)}$$

$$= 0$$

$$= \frac{\partial^2 f}{\partial x \partial y}$$

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}_{/(0,0)} = \underline{\underline{\begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}}}$$

$$h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$h^T Hh = (h_1, h_2) \cdot \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$= (-2h_1, -2h_2) \cdot \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = -2h_1^2 - 2h_2^2$$

$$= -2 \cdot \left(h_1^2 + h_2^2\right)$$

d.h. $H \prec 0$ (negativ definit) $\Rightarrow \underline{\underline{(0,0)\text{MAX}}}$

3.4.3 Bestimmung der Regressionsfunktion, die kleinste Quadrat-Schätzung

Sei $\Im_m = \left\{ \tilde{f}_m \mid \tilde{f}_{m(x)} = \sum_{\nu=0}^{m} a_\nu g_\nu(x),\ a_\nu \in \Re \right\}$, $g_\nu(x)$ – bekannt, $\nu = 0,...,m$, die Menge aller Regressionsfunktionen zu einem bestimmten Typ (Ansatz).

$\Rightarrow$ Ansatz:

$$y = a_0 g_0(x) + a_1 g_1(x) + \cdots + a_m g_m(x) + \varepsilon$$

für die Beobachtungen gilt also:

$$\begin{aligned} y_1 &= a_0 g_0(x_1) + a_1 g_1(x_1) + \cdots + a_m g_m(x_1) + \varepsilon_1 \\ &\vdots \\ y_n &= a_0 g_0(x_n) + a_1 g_1(x_n) + \cdots + a_m g_m(x_n) + \varepsilon_n \end{aligned} \qquad \textbf{(1)}$$

In Matrizenschreibweise läßt sich dieser Ansatz (1) wie folgt schreiben:

$$\vec{y} = G\vec{a} + \vec{\varepsilon}$$

wobei:

$$G = \begin{pmatrix} g_0(x_1) & \cdots & g_m(x_1) \\ \vdots & & \vdots \\ g_0(x_n) & \cdots & g_m(x_n) \end{pmatrix} = \left(\left(g_\nu(x_i)\right)\right)_{i=1,\ldots,n}^{\nu=0,\ldots,m}$$

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \vec{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \quad \text{und} \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Satz: Sei G wie oben definiert und rg(G) = m+1.

Die Parameter $\hat{a}_\nu$ der Lösung $\hat{f}_m(x) = \sum_{\nu=0}^{m} \hat{a}_\nu g_\nu(x)$ (Regressionsfunktion) des Extremwertproblems $\min_{\tilde{f}_m(x)\in\Im} \sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i)\right)^2$ ergeben sich als Lösung des

Normalengleichungssystems : $\left(G^T G\right)\vec{a} = \left(G^T \vec{y}\right)$.

Wir erhalten also die Parameter $\hat{a}_\nu$ der Regressionsfunktion wie folgt:

$$\vec{\hat{a}} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_m \end{pmatrix} = \left(G^T G\right)^{-1} G^T \vec{y}.$$

Def.: Die Lösung $\hat{a} = \begin{pmatrix} \hat{a}_0 \\ \vdots \\ \hat{a}_m \end{pmatrix}$ des Normalengleichungssystems heißt *gewöhnliche kleinste Quadrat-Schätzung* von $\vec{a}$ KQS (**O**(rdinary)**L**(east)**S**(quare)**E**(stimater)).

Beweis des Satzes: Wir müssen das Extremwertproblem

$$\min_{\tilde{f}_m\in\Im} \sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i)\right)^2 = \min_{(a_0,a_1,\ldots,a_m)\in\Re^{m+1}} \underbrace{\sum_{i=1}^{n} \left(y_i - \left(a_0 g_0(x_i) + \cdots + a_m g_m(x_i)\right)\right)^2}_{F(a_1,a_1,\ldots,a_m)}$$

$$= \min_{(a_0,a_1,\ldots,a_m)\in\Re^{m+1}} F(a_0,a_1,\ldots,a_m)$$

lösen.

$\Rightarrow$ Hier handelt es sich um ein Extremwertproblem einer Funktion F in m+1

Veränderlichen:

1. Berechnung der ersten partiellen Ableitungen und „Null“ setzen:

$$\frac{\partial F}{\partial a_0}(a_0,...,a_m) = \sum_{i=1}^{n} 2\left(y_i - \left(a_0 g_o(x_i)\right) + \cdots + a_m g_m(x_i)\right) \cdot \left(-g_0(x_i)\right) = 0$$

$$\frac{\partial F}{\partial a_1}(a_0,...,a_m) = \sum_{i=1}^{n} 2\left(y_i - \left(a_0 g_0(x_i)\right) + \cdots + a_m g_m(x_i)\right) \cdot \left(-g_1(x_i)\right) = 0$$

$$\vdots$$

$$\underbrace{\frac{\partial F}{\partial a_m}(a_0,...,a_m) = \sum_{i=1}^{n} 2\left(y_i - \left(a_0 g_0(x_i)\right) + \cdots + a_m g_m(x_i)\right) \cdot \left(-g_m(x_i)\right) = 0}_{\text{Gleichungssystem } (m+1)\times(m+1)}$$

Dieses Gleichungssytem in Matrix-Schreibweise ergibt:

$$-2 \cdot G^T \vec{y} + 2 \cdot G^T G \vec{a} = \vec{0}$$

und wir erhalten nach Umstellung das Normalengleichungssystem:

$$\underline{\underline{G^T G \vec{a} = G^T \vec{y}}}$$

Damit ist die Lösung $\hat{\vec{a}} = \left(\mathrm{G}^{\mathrm{T}}\mathrm{G}\right)^{-1}\mathrm{G}^{\mathrm{T}}\vec{\mathrm{y}}$ extremwertverdächtig.

2. Untersuchung der Matrix H der zweiten partiellen Ableitungen:

$$\frac{\partial^2 F}{\partial a_j \partial a_l} = \sum_{i=1}^{n} -2 g_l(x_i)\left(-g_j(x_i)\right) \qquad \forall\ j,l \in \{0,...,m\}$$

$$= 2 \cdot \sum_{i=1}^{n} g_l\left(x_i g_j(x_i)\right)$$

$$\Rightarrow \mathrm{H} = \left(\left(\frac{\partial^2 \chi}{\partial \mathrm{a}_j \partial \mathrm{a}_l}\right)\right) = 2 \cdot \mathrm{G}^{\mathrm{T}}\mathrm{G} \qquad \text{– Matrix der zweiten partiellen Ableitung}$$

Sei $\vec{h} \in \Re^{m+1}$. Es gilt:

$$\vec{\mathrm{h}}^{\mathrm{T}}\mathrm{H}\vec{\mathrm{h}} = 2\left(\vec{\mathrm{h}}^{\mathrm{T}}\mathrm{G}^{\mathrm{T}}\right)\cdot\left(\mathrm{G}\vec{\mathrm{h}}\right)$$

$$= 2 \cdot \vec{\mathrm{v}}^{\mathrm{T}}\vec{\mathrm{v}} \qquad \text{mit } \vec{\mathrm{v}} := \mathrm{G}\vec{\mathrm{h}}$$

$$= 2 \cdot \left|\vec{\mathrm{v}}\right|^2 \geq 0 \qquad \forall\ \vec{\mathrm{v}}$$

Daraus folgt:

$$\vec{\mathrm{h}}^{\mathrm{T}}\mathrm{H}\vec{\mathrm{h}} = 0 \quad \Leftrightarrow \vec{\mathrm{v}} = \mathrm{G}\vec{\mathrm{h}} = 0 \quad \Leftrightarrow \vec{\mathrm{h}} = \vec{0} \quad (\text{da } \mathrm{rg}(\mathrm{G}) = m+1).$$

Da $\vec{h} \neq \vec{0}$ ist $\vec{h}^T H \vec{h} > 0 \quad \forall \vec{h} \neq \vec{0}$. Also ist H positiv definit und damit ist $\hat{\vec{a}} = \left(G^T G\right)^{-1} G^T \vec{y}$ ein Minimum !

ged.

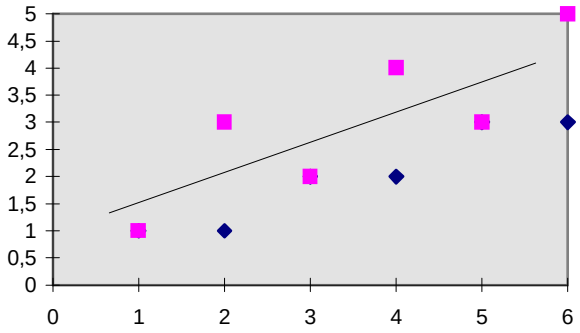
Beispiel 22:

i	1	2	3	4	5	6
x_i	1	1	2	2	3	3
y_i	1	3	2	4	3	5

Gesucht: Eine Funktion f mit $f(x) \approx y$

1.) Anhand der Daten legen wir als Typ von f eine Gerade fest:

$m = 1, \quad n = 6; \quad y = a_0 + a_1 x$



$$\vec{y} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 & , & x_1 \\ 1 & , & x_2 \\ 1 & , & x_3 \\ 1 & , & x_4 \\ 1 & , & x_5 \\ 1 & , & x_6 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \qquad G = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix}$$

2.) $\hat{a}$ ergibt sich als Lösung von

$$G^T G \vec{a} = G^T \vec{y}$$

$$G^T G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 12 & 28 \end{pmatrix}$$

$$G^T \vec{y} = \begin{pmatrix} 18 \\ 40 \end{pmatrix}$$

$\Rightarrow$ zu lösen ist: $\begin{pmatrix} 6 & 12 \\ 12 & 28 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 18 \\ 40 \end{pmatrix}$

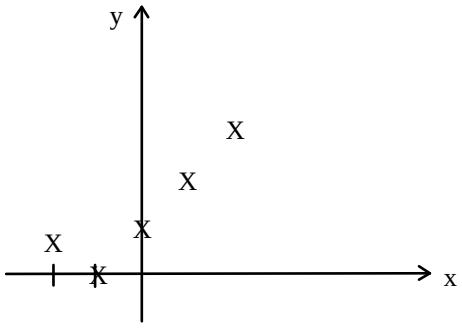
Cramersche Regel: $$a_0 = \frac{\det \begin{vmatrix} 18 & 12 \\ 40 & 28 \end{vmatrix}}{\det \begin{vmatrix} 6 & 12 \\ 12 & 28 \end{vmatrix}} = \frac{18 \cdot 28 - 40 \cdot 12}{6 \cdot 28 - 12 \cdot 12} = \underline{\underline{1}}$$

$$a_1 = \frac{\det \begin{vmatrix} 6 & 18 \\ 12 & 40 \end{vmatrix}}{\det \begin{vmatrix} 6 & 12 \\ 12 & 28 \end{vmatrix}} = \frac{6 \cdot 40 - 12 \cdot 18}{6 \cdot 28 - 12 \cdot 12} = \underline{\underline{1}}$$

3.) Ergebnis: f(x) $\approx \underline{\underline{\hat{f}(x) = 1 + x}}$

Beispiel 23:

x_i	-2	-1	0	1	2
y_i	1	0	1	2	3



1. Ansatz: Wir unterscheiden 2 in Frage kommende Ansätze:
   a) $f_1(x) = a_0 + a_1 x$ (linear)
   b) $f_2(x) = a_0 + a_1 x + a_2 x^2$ (quadratisch)

2. Berechnung der Koeffizienten $a_0, a_1, \ldots$:

Zu a) Für den linearen Ansatz:

$$y \approx a_0 + a_1 x_i \qquad ; i = 1, \ldots, n$$

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\vec{y}} \approx \underbrace{\begin{pmatrix} \underbrace{1}_{g_0(x_i)} & \underbrace{x_1}_{g_1(x_i)} \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{G} \cdot \underbrace{\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}}_{\vec{a}} \qquad \Rightarrow \qquad G = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

$$G^T G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}$$

$$G^T \vec{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

NGS: $\left(G^T G\right)\vec{a} = G^T \vec{y} \Leftrightarrow \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix} \cdot \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

$\Rightarrow$

$$= \begin{pmatrix} \frac{7}{5} \\ \frac{6}{10} \end{pmatrix} = \begin{pmatrix} 1{,}4 \\ 0{,}6 \end{pmatrix}$$

$\Rightarrow$ Beste Gerade: $\underline{\hat{f}_1(x) = 1{,}4 + 0{,}6x}$

Zu b) Für das Polynom 2. Grades:

$$y_i \approx a_0 + a_1 x_i + a_2 x_i^2$$

$$\begin{pmatrix}1\\0\\1\\2\\3\end{pmatrix} \approx \begin{pmatrix}1 & -2 & 4\\1 & -1 & 1\\1 & 0 & 0\\1 & 1 & 1\\1 & 2 & 4\end{pmatrix} \cdot \begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix}$$

$$\Rightarrow G^T G = \begin{pmatrix}1 & 1 & 1 & 1 & 1\\-2 & -1 & 0 & 1 & 2\\4 & 1 & 0 & 1 & 4\end{pmatrix} \cdot \begin{pmatrix}1 & -2 & 4\\1 & -1 & 1\\1 & 0 & 0\\1 & 1 & 1\\1 & 2 & 4\end{pmatrix} = \begin{pmatrix}5 & 0 & 10\\0 & 10 & 0\\10 & 0 & 34\end{pmatrix}$$

$$G^T \vec{y} = \begin{pmatrix}1 & 1 & 1 & 1 & 1\\-2 & -1 & 0 & 1 & 2\\4 & 1 & 0 & 1 & 4\end{pmatrix} \cdot \begin{pmatrix}1\\0\\1\\2\\3\end{pmatrix} = \begin{pmatrix}7\\6\\18\end{pmatrix}$$

$$\text{NGS: } \begin{pmatrix}5 & 0 & 10\\0 & 10 & 0\\10 & 0 & 34\end{pmatrix} \cdot \begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix} = \begin{pmatrix}7\\6\\18\end{pmatrix}$$

$$\Rightarrow a_0 = \frac{\begin{vmatrix}7 & 0 & 0\\6 & 10 & 0\\18 & 0 & 34\end{vmatrix}}{\det\left(G^T G\right)} = \frac{7 \cdot 10 \cdot 34 - 18 \cdot 10 \cdot 10}{5 \cdot 10 \cdot 34 - 10 \cdot 10 \cdot 10} = \frac{58}{70} = \underline{0{,}83}$$

$$a_1 = \frac{\begin{vmatrix}5 & 7 & 10\\0 & 6 & 0\\10 & 18 & 34\end{vmatrix}}{700} = \frac{5 \cdot 6 \cdot 34 - 10 \cdot 6 \cdot 10}{700} = \frac{32}{70} = \underline{0{,}46}$$

$$a_2 = \frac{\begin{vmatrix}5 & 0 & 7\\0 & 10 & 6\\10 & 0 & 18\end{vmatrix}}{700} = \frac{5 \cdot 10 \cdot 18 - 10 \cdot 10 \cdot 7}{700} = \frac{20}{70} = \underline{0{,}29}$$

$\Rightarrow$ Bestes Polynom 2. Grades: $\underline{\hat{f}_2(x) = 0{,}83 + 0{,}48x + 0{,}29x^2}$

Welche Funktion ist die bessere?

Betrachte:

$$\mathrm{RSS}\left(\hat{f}_1\right) = \sum_{i=1}^{5}\left(y_i - \hat{f}_1\left(x_i\right)\right)^2$$

$$\mathrm{RSS}\left(\hat{f}_2\right) = \sum_{i=1}^{5}\left(y_i - \hat{f}_2\left(x_i\right)\right)^2$$

Entscheidung:

$$\mathrm{RSS}\left(\hat{f}_2\right) < \mathrm{RSS}\left(\hat{f}_1\right) \quad \rightarrow \quad \hat{f}_2$$

$$\mathrm{RSS}\left(\hat{f}_2\right) \geq \mathrm{RSS}\left(\hat{f}_1\right) \quad \rightarrow \quad \hat{f}_1$$

	Berechnen Sie $\mathrm{RSS}\left(\hat{f}_1\right)$ und $\mathrm{RSS}\left(\hat{f}_2\right)$ und wählen Sie die beste Regressionsfunktion!

3.5 Übungsaufgaben

1) Geben Sie ein geeignetes Zusammenhangsmaß für die folgenden Merkmalspaare an:

a) Studienfach und Anfangsgehalt in DM bei den Absolventen einer Hochschule.
b) Einstellungsalter und Anfangsgehalt in DM bei Absolventen einer Hochschule.
c) Verdienst in DM und ausgeübter Beruf.
d) Studienfach und Geschlecht.
e) Körpergröße und Plazierung beim 100-Meter-Lauf.
f) % tualer Kohlenstoffgehalt und Güteklasse einer Stahlsorte
g) Leistung (Zensur) in "Statistik" und Schuhgröße
h) Leistung (Zensur) in "Statistik" und Haarfarbe

2) Die 50 Teilnehmer einer Statistikklausur werden nach dem Familienstand geordnet:

	bestanden	nicht bestanden
Ledig	36	4
Verheiratet	4	6

a) Berechnen Sie den Kontingenzkoeffizienten!
b) Berechnen Sie den korrigierten Kontingenzkoeffizienten!

3) In einem psychologischen Experiment wurde die Reaktionszeit von Probanden vom Geben bis zum Erkennen eines optischen Signals gemessen. Die Reaktionszeit wurde dabei in 3 Gruppen eingeteilt:

1 – schnell, 2 - mittel, 3 – langsam.

Eine Untersuchung an 30 Personen ergab folgende Daten : (M-männlich, W-weiblich)

Person	Geschlecht	Reaktionszeit
1	W	1
2	W	2
3	W	2
4	M	3
5	M	1
6	W	3
7	M	2
8	M	1
9	W	3
10	M	1

Person	Geschlecht	Reaktionszeit
11	M	2
12	M	1
13	W	1
14	W	1
15	M	3
16	W	2
17	W	1
18	M	3
19	M	3
20	M	1

Person	Geschlecht	Reaktionszeit
21	W	2
22	W	3
23	M	2
24	M	2
25	W	1
26	W	2
27	W	2
28	W	3
29	M	1
30	M	1

a) Geben Sie an:
- Die relative Häufigkeitsverteilung der Zufallsgröße X=Geschlecht!
- Die relative Häufigkeitsverteilung der Zufallsgröße Y=Reaktionszeitgruppe!

- Die bedingte relative Häufigkeitsverteilung von Y unter der Bedingung: das Geschlecht ist weiblich!

b) Untersuchen Sie mit einer geeigneten Methode der deskriptiven Statistik, ob (zumindest in der Stichprobe) die Reaktionszeit vom Geschlecht abhängt oder nicht!

4.) 10 Studenten veranstalten einen Wettlauf. Die folgende Tabelle enthält die Körpergröße und die Plazierung:

Student	A	B	C	D	E	F	G	H	I	J
Körpergröße	180	170	174	190	165	182	178	169	184	189
Platz	3	7	8	2	10	5	6	9	1	4

Berechnen Sie ein Maß für den Zusammenhang zwischen Körpergröße und Plazierung!

5.) Bei einem Turnwettkampf haben 6 Turner die folgenden Bewertungen an zwei Geräten erzielt:

Turner Nr.	1	2	3	4	5	6
Reck	9,3	8,6	9,1	9,1	9,0	9,5
Stufenbarren	9,1	8,8	9,0	8,9	8,7	9,4

Berechnen Sie den Spearman'schen Rangkorrelationskoeffizienten!
Gibt es einen Zusammenhang? Bewerten Sie den Zusammenhang!

6) Aus 80 Wertepaaren der Merkmale X und Y wurde ein Pearson'scher Korrelationskoeffizient von r = -0,95 berechnet.
Welche der folgenden Aussagen sind richtig?

a) Die Beobachtungswerte streuen eng um eine fallende Gerade.
b) Ein Zusammenhang ist nicht erwiesen, da r < 0 gilt.
c) Die Werte von X und Y sind annähernd umgekehrt proportional zueinander.
d) Berechnet man für die Wertepaare eine Regressionsgerade nach dem Kriterium der kleinsten Quadrate, dann erhält man für b einen negativen Wert.

7.) Gegeben seien die folgenden Beobachtungswerte der gemeinsam auftretenden Merkmale X und Y.

x	0	2	4	6
y	2	0	6	4

Berechnen Sie die Stichprobenkovarianz und den Pearsonschen Korrelationskoeffizienten!

8) Berechnen Sie für die folgenden Beobachtungswerte (x_i , y_i) (i = 1,...,6) eine Regressionsfunktion vom Typ $y = a + bx + cx^2$ nach der Methode der kleinsten Quadrate!

$$(1;6), (2;4), (3;3), (3;5), (4;6), (5;10)$$