II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

Aufgabe der beschreibenden Statistik:

Große und unübersichtliche Datenmengen so aufbereiten, dass wenige aussagekräftige Kenngrößen und/oder Graphiken entstehen, in denen die gesamte Datenmenge „fokussiert“ ist.

Beispiele:

Gesamtnote eines Zeugnisses, in die die Einzelnoten u.U. mit unterschiedlicher Gewichtung eingehen.

Graphiken zur Darstellung der Verteilung der Daten

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

Beispiel: Befragung von 60 Hörern einer Statistik-Vorlesung nach:

1. Familienstand
2. Studienrichtung
3. Interesse am Vorlesungsgegenstand (außerordentlich interessiert, sehr interessiert, interessiert, kaum interessiert, gar nicht interessiert)
4. Anzahl der Geschwister
5. Anzahl der bereits studierten Hochschulsemester
6. Körpergröße
7. Körpergewicht
8. Weglänge von der Wohnung zur Hochschule

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

Begriffe:
Beobachtungsmenge (auch „statistische Masse“)

Gesamtheit der befragten Hörer (60 Personen) der Statistik-Vorlesung
Die Beobachtungsmenge muss räumlich, zeitlich und sachlich präzise definiert werden,
z.B.:

- Räumlich: Ausbildung in Aachen
- Zeitlich: WS 19/20
- Sachlich: MATSEs in Ausbildung im zweiten Ausbildungsjahr

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

- Mögliche statistische Massen:
  Natürliche Personen, Sachen (Maschinen, Produkte,..), Institutionen (Betriebe, Städte, Länder,...), Ereignisse (Maschinenausfälle, Geburten, Todesfälle,...)

- Beobachtungseinheit:
  Ein einzelner Hörer der Statistik-Vorlesung

- Beobachtungsmerkmal:
  Erfragte Eigenschaft

- Merkmalsausprägung: (auch „Merkmalswert")
  Mögliches Ergebnis bei der Beobachtung eines Merkmals

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

Offensichtlich müssen verschiedene „Typen“ von Merkmalen unterschieden werden:

Der Familienstand wird anders charakterisiert als die Körpergröße, und das Interesse am Vorlesungsgegenstand hat eine andere „Skala“ als die Studienrichtung!

Merkmalstypen:

Qualitative Merkmale

Die Werte brauchen keine physikalische Einheit, nochmal unterschieden:

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

- Qualitativ-nominale Merkmale:
Merkmalsausprägungen sind nur dem Namen nach unterscheidbar, drücken aber keinerlei Wertung oder Intensität aus.
In unserem Beispiel: Familienstand, Studienrichtung

- Qualitativ-ordinale Merkmale
(auch „Rang-Merkmale“):
Merkmalsausprägungen können zusätzlich noch in eine inhaltlich sinnvolle Rangordnung gebracht werden, aber keine definierte Skala.
In unserem Beispiel: Interesse am Vorlesungsgegenstand

Quantitative Merkmale

(auch „metrische" oder „kardinale" Merkmale)

- Quantitativ – diskrete Merkmale:

Merkmale, die nur bestimmte, auf der Zahlengeraden getrennt liegende Werte annehmen können.
I.d.R. die natürlichen Zahlen 0,1,2,3,.. die durch einen Zählprozess entstehen; dazwischen können keine Werte angenommen werden.

In unserem Beispiel: Anzahl der Geschwister, Zahl der bereits studierten Semester

II. Beschreibende Statistik
II.1 Merkmale und wichtige Begriffe

- Quantitativ – stetige Merkmale:

Werden durch Messung gewonnen und können jeden Wert innerhalb eines sinnvollen Intervalles annehmen.

In unserem Beispiel: Körpergröße, Körpergewicht, Weglänge von der Wohnung zur Hochschule

Ein diskretes Merkmal X

- Urliste: Liste, die direkt bei der Datenerhebung entsteht. Unübersichtlich!
- Darstellung der Häufigkeitsverteilung des Merkmals X in Form einer Häufigkeitstabelle

Bezeichnungen:

Absolute Häufigkeit des Merkmalswertes a_i:
n_i = Anzahl des Vorkommens des Merkmalswertes a_i bei den n beobachteten Merkmalswerten

$$0 \leq n_i \leq n; \quad \sum_i n_i = n$$

II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse

Relative Häufigkeit des Merkmalswertes a_i:

$$h_i := \frac{n_i}{n} = \frac{\text{Absolute Häufigkeit}}{\text{Anzahl der Beobachtungen}}; \quad 0 \leq h_i \leq 1; \quad \sum_i h_i = 1$$

Beispiel:

Befragung von 60 erfolgreichen Studienabsolventen zum Merkmal X: „Anzahl Fachsemester bis zum Diplom"

Urliste:

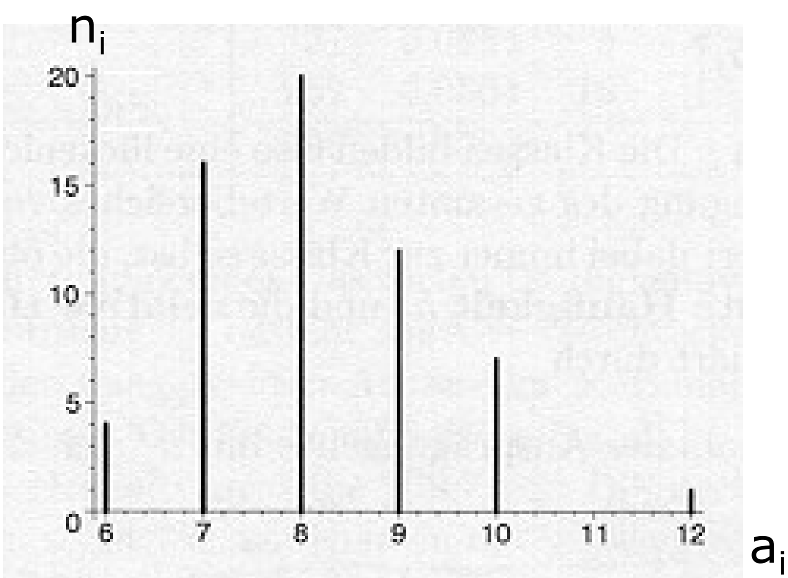
9	8	7	7	8	10	6	8	8	7	9	7
10	8	8	9	7	8	9	10	6	10	8	9
9	7	7	8	8	7	8	7	7	8	8	8
10	7	10	9	8	6	9	7	8	7	9	12
9	8	9	6	12	8	7	8	9	7	8	7

Häufigkeitstabelle:

Semester-zahl	Strichliste	Häufigkeit	
a_i		absolut n_i	relativ h_i
6	||||	4	0.0667
7	||||| ||||| ||||| |	16	0.2667
8	||||| ||||| ||||| |||||	20	0.3333
9	||||| ||||| ||	12	0.2000
10	||||| |	6	0.1000
12	||	2	0.0333
$\sum$		60	1.0000

Stabdiagramm:

Grafische Darstellung der unklassierten Häufigkeits-verteilung, absolute oder relative Häufigkeit der Merkmalsausprägung wird aufgetragen



<u>Häufige Fragestellung</u>: Welcher Anteil der Beobachtungsmenge liegt unterhalb oder oberhalb einer bestimmten Grenze, bzw. zwischen zwei Grenzen?

Bezeichnungen:

Absolute Summenhäufigkeit:

$$G(x) := (\text{Anzahl der Beobachtungen} \leq x) := \sum_{i;a_i \leq x} n_i \qquad x \in R$$

Relative Summenhäufigkeit

(empirische Verteilungsfunktion):

i	Semester-zahl a_i	rel. Häufigkeit einfach h_i	rel. Häufigkeit kumuliert H_i
1	6	0.0667	0.0667
2	7	0.2667	0.3333
3	8	0.3333	0.6667
4	9	0.2000	0.8667
5	10	0.1000	0.9667
6	12	0.0333	1.0000

$$H(x) := \frac{G(x)}{n} = \sum_{i;a_i \leq x} \frac{n_i}{n} = \sum_{i;a_i \leq x} h_i$$

<u>Beispiel zu den Semesterzahlen:</u>

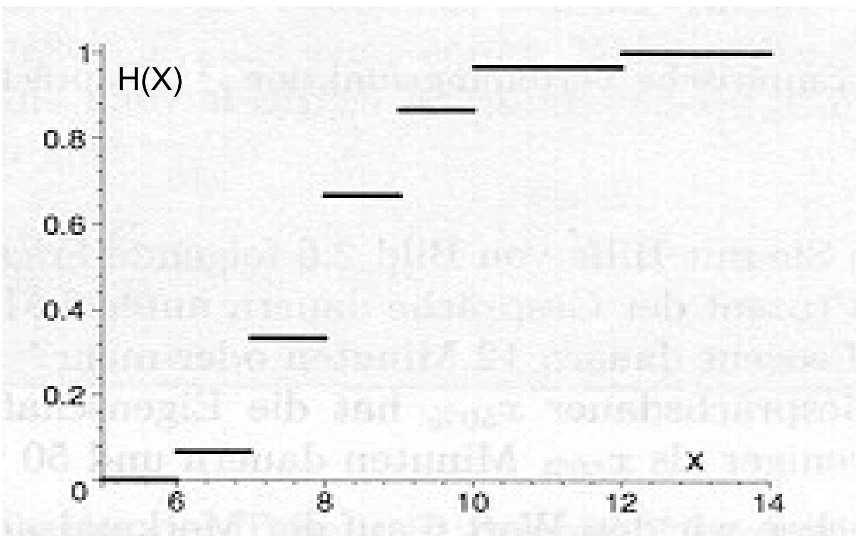
- Anteil mit höchstens 9 Semestern: $H_4=0,8667$
- Anteil mit 8 oder mehr Semestern:
  $1-H_2=1-0,3333=0,6667$
- Anteil mit 7 bis 9 Semestern:
  $H_4 - H_1 = 0,8667-0,0667=0,8000$

i	Semesterzahl a_i	rel. Häufigkeit einfach h_i	rel. Häufigkeit kumuliert H_i
1	6	0.0667	0.0667
2	7	0.2667	0.3333
3	8	0.3333	0.6667
4	9	0.2000	0.8667
5	10	0.1000	0.9667
6	12	0.0333	1.0000

II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse

Empirische Verteilungsfunktion:

Rechtsseitig stetige Verteilungsfunktion mit den Merkmalswerten als Sprungstellen und ihren relativen Häufigkeiten als Sprunghöhen



Bei stetigen Merkmalen andere Darstellung üblich!

Ein stetiges Merkmal X

- Sehr viele verschiedene Merkmalsausprägungen, u.U. sogar bei allen n beobachteten Einheiten verschiedene Werte:
„Klassierte" Häufigkeitsverteilung ist sinnvoll.
- Für den Gewinn an Übersichtlichkeit zahlt man mit einem Informationsverlust, denn über die Verteilung der Werte innerhalb einer Klasse ist dann nichts mehr bekannt.

Alle n Werte in einem Intervall $[a, b]$; Einteilung des Intervalls in disjunkte Klassen $A_1, \ldots, A_k$; $A_i = (a_{i-1}, a_i]$; $a = a_0 < a_1 < \ldots < a_k = b$;

i.a. äquidistant; $\alpha_i = \dfrac{a_i + a_{i-1}}{2}$ als Klassenmitten

II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse

Absolute Klassenhäufigkeit der Klasse A_i:
n_i = Anzahl des Vorkommens der Klasse A_i bei den n beobachteten Merkmalswerten

$$0 \le n_i \le n; \quad \sum_i n_i = n$$

Relative Klassenhäufigkeit der Klasse A_i:

$$h_i := \frac{n_i}{n} = \frac{\text{Absolute Klassenhäufigkeit}}{\text{Anzahl der Beobachtungen}}; \quad 0 \le h_i \le 1; \quad \sum_i h_i = 1$$

Relative Häufigkeitsdichte
der Stichprobe bei Klasseneinteilung:

$$h(x) := \frac{h_i}{a_i - a_{i-1}} = \frac{\text{Relative Klassenhäufigkeit}}{\text{Klassenbreite}}; \quad x \in (a_{i-1}; a_i]$$

II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse

Der Graph von h(x) ist ein „Histogramm“:
Darstellung einer Häufigkeitsverteilung durch die Errichtung von Rechtecken über die Klassen einer Zerlegung der Merkmalswerte, deren **Flächen** proportional zu den (relativen) Klassenhäufigkeiten sind.

Faustregel für die Klassenzahl k: $\boxed{5 \leq k \leq 20 \quad und \quad k \approx \sqrt{n}}$

John/Q-DAS:
Die Anzahl der Klassen liegt zwischen der Quadrat- und Kubikwurzel von n. Es werden möglichst glatte Klassengrenzen gebildet.

DIN 55302-1:
Klassierungsmodell, bei dem die Forderung für die Mindestanzahl der Klassen nach DIN 55302-T2 erst ab n = 100 erfüllt wird. Bei kleinerem Stichprobenumfang ergibt sich die Anzahl der Klassen aus der Quadratwurzel von n.

DIN 55302-1/Q-DAS:
Die Mindestanzahl der Klassen ist auch bei einem Stichprobenumfang von n < 100 auf 10 festgelegt.

Sturges/CNOMO:
Modell nach der französischen CNOMO-Norm.

Quelle:Q-DAS GmbH

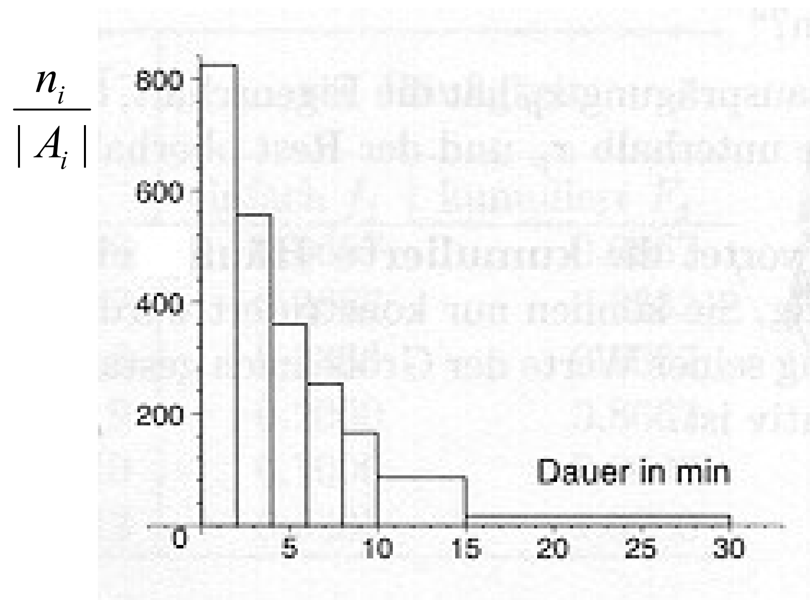
II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse

Beispiel: Von 5000 Telefonaten wurde in einer Telefonzentrale die Dauer in Minuten gemessen

Klassierte Häufigkeitstabelle:

i	Klasse A_i	n_i	h_i	$\lvert A_i \rvert$
1	0–2	1650	0.3300	2
2	2–4	1111	0.2222	2
3	4–6	720	0.1440	2
4	6–8	508	0.1016	2
5	8–10	332	0.0664	2
6	10–15	427	0.0854	5
7	15–30	252	0.0504	15
Σ		5000	1.0000	

Histogramm:



Wäre die Höhe der Rechtecke nicht proportional zur (relativen) Häufigkeitsdichte, sondern zur (relativen) Häufigkeit, würden Klassen mit großer Breite überproportional erscheinen und es entstünde ein falscher optischer Eindruck!

Welcher Anteil der Beobachtungsmenge liegt unterhalb oder oberhalb einer bestimmten Grenze, bzw. zwischen zwei Grenzen?

Bezeichnungen:

Absolute Summenhäufigkeit:

$$G(x) := \text{Anzahl der Beobachtungen} \leq x; \quad x \in R$$

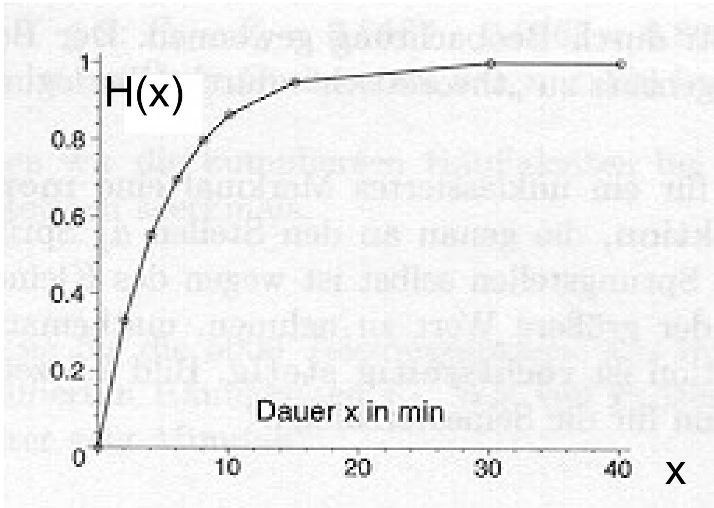
Relative Summenhäufigkeit (empirische Verteilungsfunktion):

$$H(x) := \frac{G(x)}{n}$$

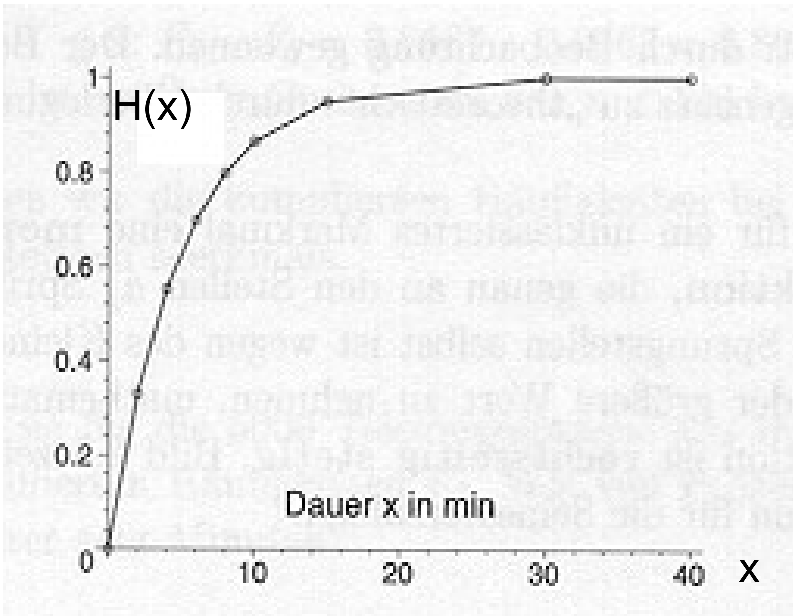
Klassierte Häufigkeitstabelle mit rel. Summenhäufigkeit H:

i	Klasse A_i	relativ, h_i	relativ, kumuliert H_i
1	0–2	0.3300	0.3300
2	2–4	0.2222	0.5522
3	4–6	0.1440	0.6962
4	6–8	0.1016	0.7978
5	8–10	0.0664	0.8642
6	10–15	0.0854	0.9496
7	15–30	0.0504	1.0000

Empirische Verteilungsfunktion:



II. Beschreibende Statistik
II.2 Darstellung der Beobachtungsergebnisse



Bei der empirischen Verteilungsfunktion bilden die Klassenobergrenzen mit ihren zugeordneten relativen Summenhäufigkeiten die Stützpunkte, die durch Strecken verbunden werden. Dabei wird $H(a_0)=0$ gesetzt.

II. Beschreibende Statistik
II.3 Statistische Maßzahlen

Mit statistischen Maßzahlen sollen die gewonnenen Daten komprimiert werden, d.h. die Charakterisierung der Daten erfolgt durch einige typische Kennwerte. Dafür benötigt man Lageparameter (Lagemaßzahlen) und Streuungsparameter (Streuungsmaßzahlen), sowie bei mehrdimensionalen Merkmalen auch Abhängigkeitsmaße.

II.3.1 Lageparameter

<u>Arithmetisches Mittel</u>
(Stichprobenmittel, empirischer Erwartungswert)
Nur bei quantitativen Merkmalen!
Aus der Urliste mit x_i als Ausprägung des i-ten Elements:

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

Aus der unklassierten Häufigkeitstabelle:

$$\overline{x} = \frac{a_1 \cdot n_1 + a_2 \cdot n_2 + \ldots a_m \cdot n_m}{n} = \sum_{j=1}^{m} a_j \cdot \frac{n_j}{n} = \sum_{j=1}^{m} a_j \cdot h_j$$

Aus der klassierten Häufigkeitstabelle:
Näherungsweise möglich, indem man die Merkmalsausprägung durch die Klassenmitte ersetzt!

$$\overline{x} \approx \frac{1}{n} \sum_{i=1}^{k} n_i \cdot \alpha_i = \sum_{i=1}^{k} h_i \cdot \alpha_i$$

Median (Zentralwert)

Aus der Urliste

(x_i Ausprägung des i-ten Elements)
Bildung der geordneten Stichprobe:

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

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n \text{ gerade} \end{cases}$$

„Vorläufige Definition", später allgemeiner!

Aus der unklassierten Häufigkeitstabelle

Merkmalsausprägungen a_i der Größe nach sortieren, betrachte die relative Summenhäufigkeit H_i:

- Wird bei der Merkmalsausprägung a_i der Wert H=0,5 für die relative Summenhäufigkeit zum erstenmal überschritten, so ist dies der Median.
- Selten: Wird H_i=0,5 bei a_i genau erreicht, so ist der Median das arithmetische Mittel aus a_i und a_{i+1}.

Beispiel: Studiendauer in Fachsemestern

i	Semester-zahl a_i	rel. Häufigkeit einfach h_i	rel. Häufigkeit kumuliert H_i
1	6	0.0667	0.0667
2	7	0.2667	0.3333
3	8	0.3333	0.6667
4	9	0.2000	0.8667
5	10	0.1000	0.9667
6	12	0.0333	1.0000

$$\tilde{x} = 8$$

Aus der klassierten Häufigkeitstabelle:

- Man sucht in der Häufigkeitstabelle die „Einfallsklasse", in der zum erstenmal der Wert 0,5 für die relative Häufigkeitssumme erreicht oder überschritten wird.
- Innerhalb dieser Klasse wird der Median mit linearer Interpolation ermittelt.

Einfallsklasse $A_j = (a_j; b_j]$

$$\tilde{x} = a_j + \frac{0{,}5 - H_{j-1}}{H_j - H_{j-1}} \cdot (b_j - a_j)$$

Beispiel: Dauer von Telefonaten

	Klasse A_i	relativ, h_i	relativ, kumuliert H_i
1	0–2	0.3300	0.3300
2	2–4	0.2222	0.5522
3	4–6	0.1440	0.6962
4	6–8	0.1016	0.7978
5	8–10	0.0664	0.8642
6	10–15	0.0854	0.9496
7	15–30	0.0504	1.0000

Einfallsklasse: $A_2 = (2;4]$

$$\tilde{x} = 2 + \frac{0{,}5 - 0{,}3300}{0{,}5522 - 0{,}3300} \cdot (4-2) \approx 3{,}53$$

Modalwert (Häufigster Wert)

Großer Vorteil: Im Gegensatz zum arithmetischen Mittel auch bei nominalen Merkmalen:

- Der Modalwert ist diejenige Merkmalsausprägung mit der größten (absoluten oder relativen) Häufigkeit.

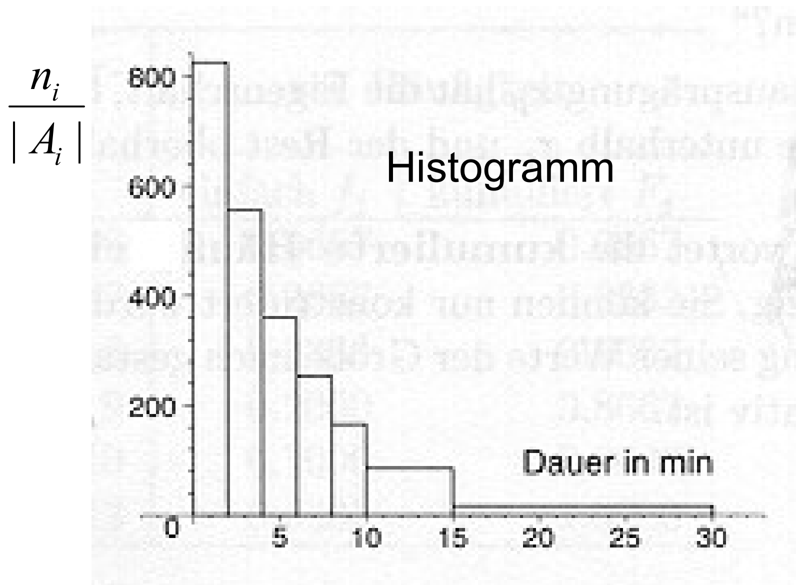
Beispiel: Studiendauer in Fachsemestern

i	Semester-zahl a_i	rel. Häufigkeit einfach h_i	rel. Häufigkeit kumuliert H_i
1	6	0.0667	0.0667
2	7	0.2667	0.3333
3	8	0.3333	0.6667
4	9	0.2000	0.8667
5	10	0.1000	0.9667
6	12	0.0333	1.0000

Modalwert $a_3=8$

• Bei klassierten Daten ist die Modalklasse diejenige Klasse mit der größten Besetzungsdichte

Beispiel: Dauer von Telefonaten



Modalklasse = [0;2]

II.3.2 Quantile

Ein p-Quantil x_p soll die Beobachtungen in einen Anteil p kleiner als x_p und einen Anteil 1-p größer als x_p aufteilen, d.h. die empirische Verteilungsfunktion H sollte an der Stelle x_p den Wert p annehmen: $H(x_p)=p$.

Da die empirische Verteilungsfunktion H nur endlich viele verschiedene Werte annimmt, fordert man stattdessen, dass H an der Stelle x_p den Wert p „überspringt“:

Definition:

Sei $0 < p < 1$; x_p heißt (empirisches) p-Quantil

$\Leftrightarrow H(x_p^-) \leq p \wedge H(x_p^+) = H(x_p) \geq p$

Wird p von H nicht angenommen, existiert genau

ein p-Quantil x_p mit $H(x_p^-) < p \wedge H(x_p) > p$

Wird p von H angenommen, existiert ein p-Quantil-Intervall.

Aus der geordneten Stichprobe:

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

$$x_p = \begin{cases} x_{([np]+1)} & \text{falls np nicht ganzzahlig} \\ \left[x_{(np)}; x_{(np+1)}\right] & \text{falls np ganzzahlig; p-Quantil-Intervall} \end{cases}$$

Bemerkung: $x_{(np)} = x_{(np+1)}$ ist möglich!

Beispiel: Studiendauer in Fachsemestern

Aus der unklassierten Häufigkeitstabelle

i	Semesterzahl a_i	rel. Häufigkeit einfach h_i	rel. Häufigkeit kumuliert H_i
1	6	0.0667	0.0667
2	7	0.2667	0.3333
3	8	0.3333	0.6667
4	9	0.2000	0.8667
5	10	0.1000	0.9667
6	12	0.0333	1.0000

$$x_{0,1} = a_2 = 7$$

Oberes Quartil: $x_{0,75} = 9$

Unteres Quartil: $x_{0,25} = 7$

Aus der klassierten Häufigkeitstabelle:
(Analog zur Bestimmung des Medians!)

- Man sucht in der Häufigkeitstabelle die „Einfalls-klasse", in der zum erstenmal der Wert p für die relative Häufigkeitssumme erreicht oder über-schritten wird.
- Innerhalb dieser Klasse wird das Quantil mit linearer Interpolation ermittelt.

Einfallsklasse $A_j = (a_j;b_j]$

$$x_p = a_j + \frac{p - H_{j-1}}{H_j - H_{j-1}} \cdot (b_j - a_j)$$

<u>Beispiel</u>:
Dauer von Telefonaten

i	Klasse A_i	relativ, h_i	relativ, kumuliert H_i
1	0–2	0.3300	0.3300
2	2–4	0.2222	0.5522
3	4–6	0.1440	0.6962
4	6–8	0.1016	0.7978
5	8–10	0.0664	0.8642
6	10–15	0.0854	0.9496
7	15–30	0.0504	1.0000

$$x_{0,8} = ?$$

Einfallsklasse: $A_5 = (8;10]$

$$x_{0,8} = 8 + \frac{0,8 - 0,7978}{0,8642 - 0,7978} \cdot (10 - 8) \approx 8,066$$

II.3.3 Streuungsmaße

Vorsicht mit der Bezeichnung „Streuung": Dieser Begriff kann je nach Lehrbuch/Autor unterschiedliche Bedeutungen haben!

Spannweite (Variationsbreite, Range):

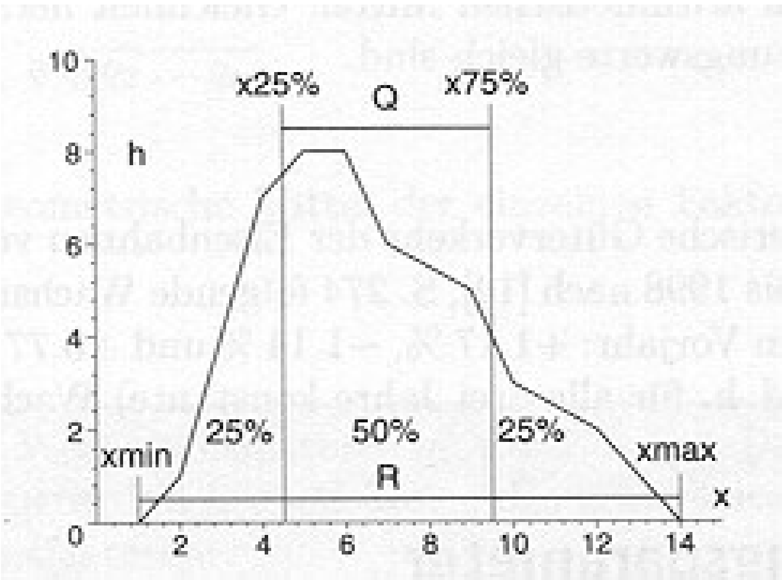
$$\boxed{R = x_{(n)} - x_{(1)} = x_{\max} - x_{\min}}$$

Nachteil:
Ein „Ausreißer" kann den Wert von R stark in die Höhe treiben

Quartilsabstand: $\boxed{Q = x_{0,75} - x_{0,25}}$ „Innere" 50%

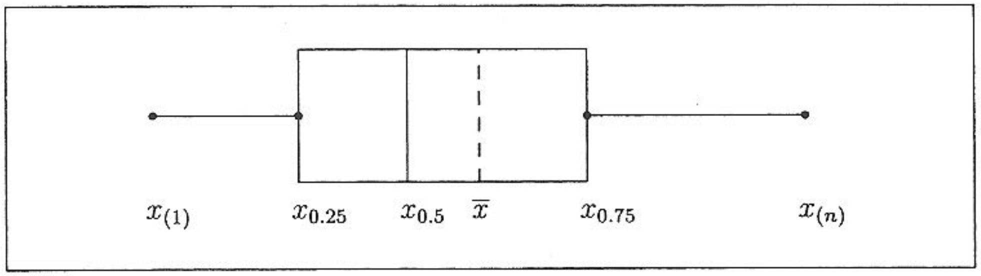
Ausreißer nach oben oder unten werden abgeschnitten, „robusteres" Streuungsmaß als die Spannweite R, reagiert nicht so empfindlich auf Ausreißer.

Beispiel:



„Boxplot"

Quelle: Lehn/Wegmann

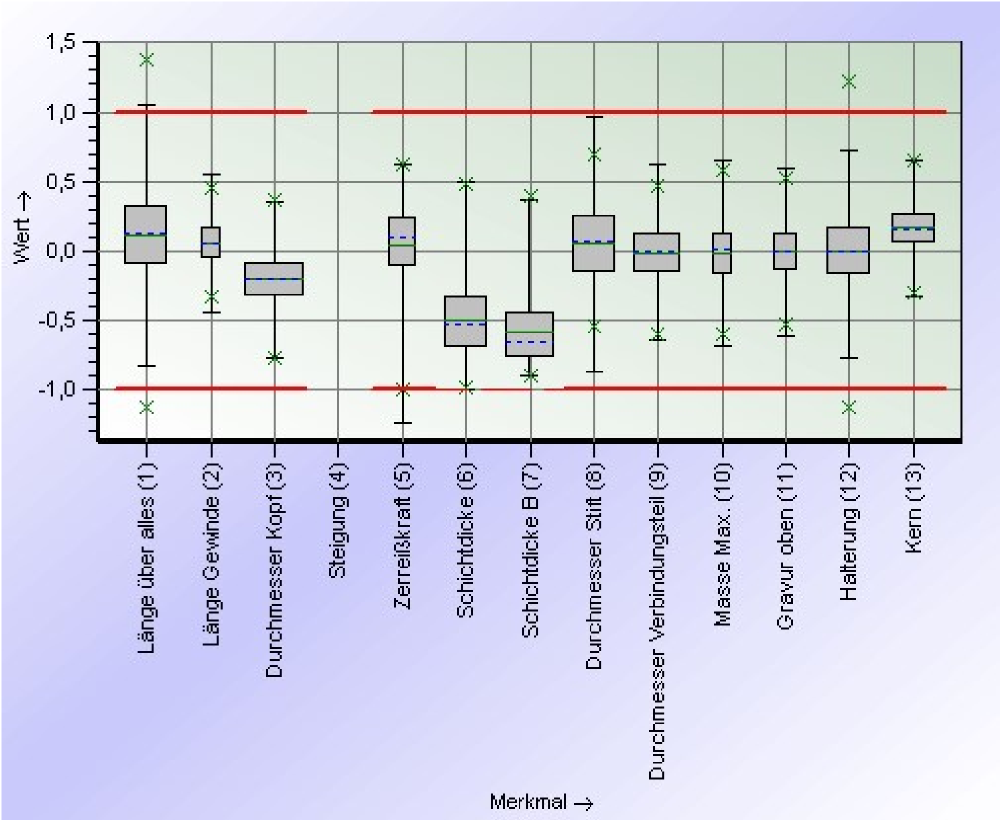


Nicht eindeutig definiert!

II. Beschreibende Statistik
II.3 Statistische Maßzahlen

Boxplots sind besonders geeignet, um mehrere Merkmale schnell in ihrer Lage und Streuung gegeneinander zu vergleichen:

Quelle: Q-DAS GmbH



Nachteil:
Spannweite und Quartilsabstand werden nur von zwei Merkmalsausprägungen bestimmt; was dazwischen passiert, hat auf R und Q keinen Einfluss!

Gesucht:
Streuungsmaß, welches alle x_i berücksichtigt!

Empirische Varianz

Aus der Urliste: $$\boxed{s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right)}$$

Warum „(n-1)“ im Nenner?
S. Eigenschaften von Schätzfunktionen in Kap.III !
Bis dahin: Vorsicht bei der Benutzung eines Taschenrechners!

Aus der unklassierten Häufigkeitstabelle:

$$\boxed{s^2 = \frac{1}{n-1} \cdot \sum_{j=1}^{m} (a_j - \overline{x})^2 \cdot n_j = \frac{1}{n-1} \cdot \left(\sum_{j=1}^{m} a_j^2 \cdot n_j - n \cdot \overline{x}^2 \right)}$$

Aus der klassierten Häufigkeitstabelle: Nur näherungsweise möglich, indem man die Merkmals-ausprägung durch die Klassenmitte ersetzt:

$$\boxed{s^2 \approx \frac{1}{n-1} \cdot \sum_{j=1}^{m} \alpha_j^2 \cdot n_j - \frac{1}{n(n-1)} \left(\sum_{j=1}^{m} \alpha_j \cdot n_j \right)^2}$$

II. Beschreibende Statistik
II.3 Statistische Maßzahlen

Empirische Standardabweichung:

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

(Empirischer) Variationskoeffizient: $\boxed{V := \frac{s}{\overline{x}}}$

Empfindliche Größe zur Beurteilung von Mess-verfahren; insbesondere in der chemischen Analytik zur Akkreditierung von Analysenlaboratorien.

II.3.4 Lineare Regression und Korrelation

Betrachte bei einer statistischen Masse zwei Merkmale: X und Y

Urliste:

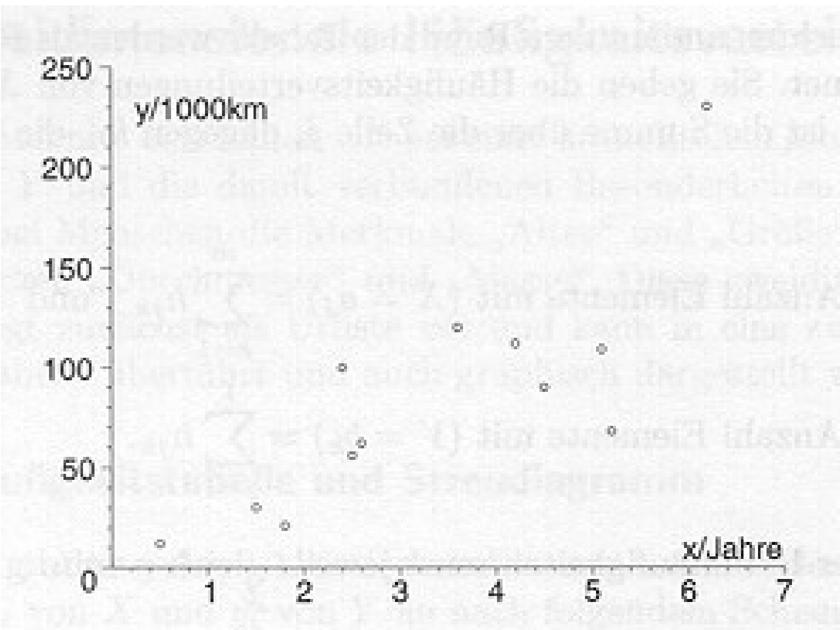
Element Nr.	1	2	...	i	...	n
Ausprägung von X	x_1	x_2	...	x_i	...	x_n
Ausprägung von Y	y_1	y_2	...	y_i	...	y_n

<u>Beispiel</u>:
Alter und Fahrstrecke von Kraftfahrzeugen eines Fuhrparks

Nr.	Alter in Jahren	Strecke in 1000 km	Nr.	Alter in Jahren	Strecke in 1000 km
1	1.5	30	7	1.8	21
2	5.2	68	8	4.2	112
3	4.5	90	9	6.2	230
4	0.5	12	10	3.6	120
5	2.4	100	11	2.5	56
6	2.6	62	12	5.1	109

II.3.4 Lineare Regression und Korrelation

Graphische Darstellung:
Streudiagramm (Punktwolke)



Kein richtungsloser Punkthaufen, sondern wachsende Tendenz mit „Störung“, d.h. es gibt einen Zusammenhang zwischen den beiden Merkmalen!

Quantitative Erfassung des Zusammenhangs?

II.3.4 Lineare Regression und Korrelation

Zur quantitativen Beschreibung des Zusammenhangs werden die folgenden bekannten Größen benötigt:

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2; \quad s_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Empirische Varianzen

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

Arithmetische Mittelwerte

Neue Maßzahl, an der beide Messreihen gleichzeitig beteiligt sind: <u>Empirische Kovarianz</u>

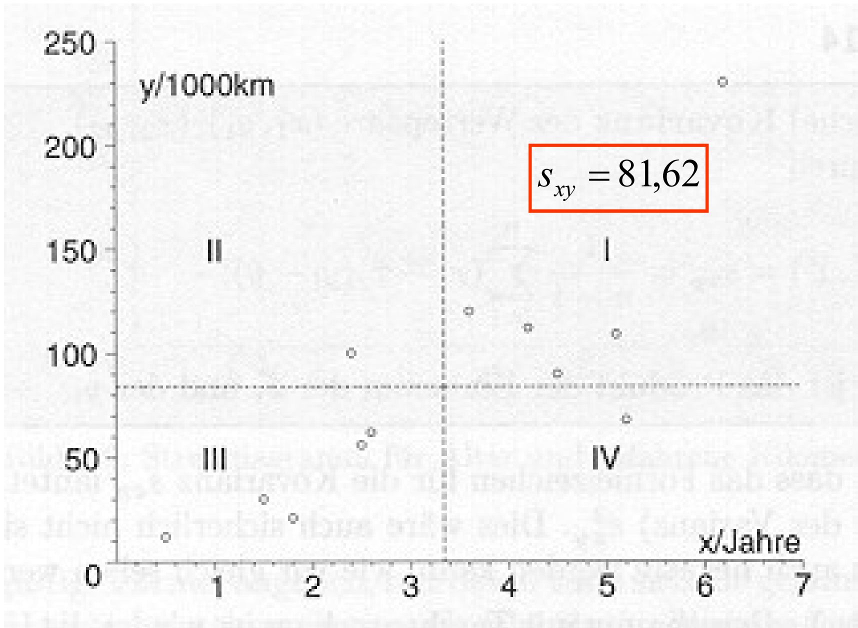
$$\boxed{Cov(X,Y) = s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y} \right)}$$

II.3.4 Lineare Regression und Korrelation

Anschauliche Bedeutung:
Lege Ursprung eines neuen Koordinatensystems nach $(\bar{x}; \bar{y})$

$x_i < \bar{x}; y_i > \bar{y}$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) < 0$

$x_i > \bar{x}; y_i > \bar{y}$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) > 0$



$x_i < \bar{x}; y_i < \bar{y}$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) > 0$

$x_i > \bar{x}; y_i < \bar{y}$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) < 0$

II.3.4 Lineare Regression und Korrelation

- Ist die Kovarianz positiv, so sind große x_i mit großen y_i gekoppelt, positive Korrelation
- Ist die Kovarianz negativ, so sind große x_i mit kleinen y_i gekoppelt, negative Korrelation

⇒ Die Kovarianz ist ein Maß für die „Richtung" des Zusammenhangs

Stärke des Zusammenhangs: Korrelationsrechnung
Art des Zusammenhangs: Regressionsrechnung

II.3.4 Lineare Regression und Korrelation

<u>Korrelationsrechnung</u>

Die Größe der Kovarianz lässt sich nicht sinnvoll interpretieren (z.B. von Einheit abhängig!)

$$\boxed{r_{xy} := \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

Empirischer Korrelationskoeffizient

$$|s_{xy}| \leq s_x \cdot s_y$$

$$|s_{xy}| = s_x \cdot s_y$$

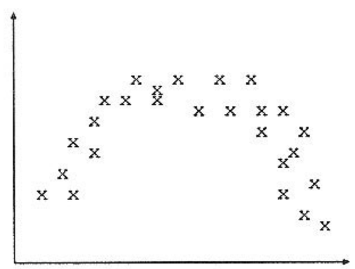
$$\Leftrightarrow y_i = a + bx_i$$

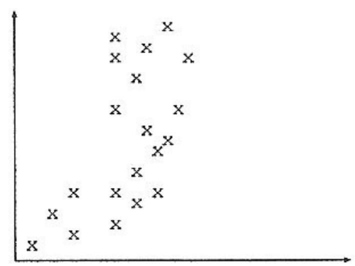
$$r_{x,y} \in [-1;1]$$

$$r_{x,y} \approx \begin{cases} -1 & \text{starker fallender Zusammenhang} \\ 0 & \text{kein Zusammenhang} \\ 1 & \text{starker positiver Zusammenhang} \end{cases}$$

II.3.4 Lineare Regression und Korrelation

r_{xy} macht nur Sinn, falls Zusammenhang linear ist!





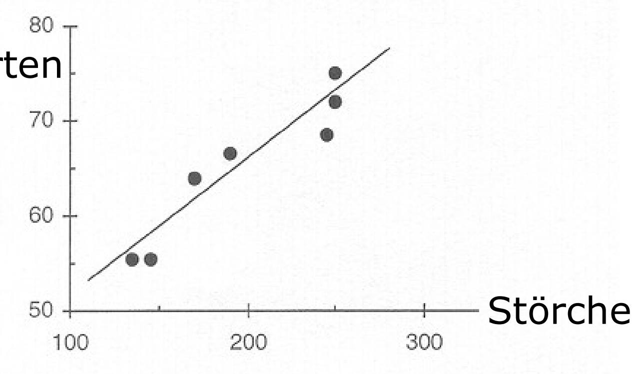
„Rangkorrelation“ verwenden

Falls linearer Zusammenhang ohne Störung, d.h.

$$y = a + b \cdot x \qquad r_{x,y} = \frac{b}{|b|} = \begin{cases} 1 \text{ falls } b > 0 \\ -1 \text{ falls } b < 0 \end{cases}$$

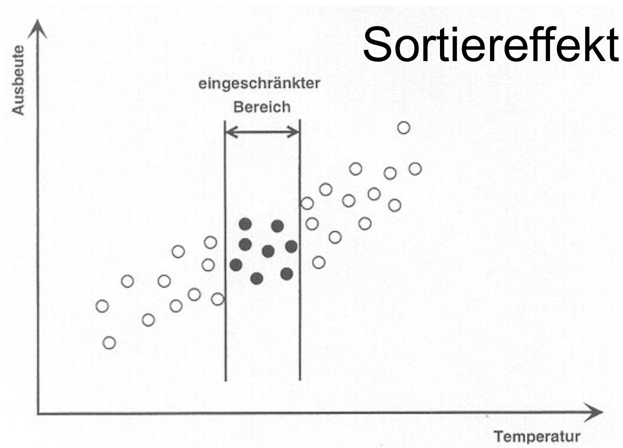
II.3.4 Lineare Regression und Korrelation

„Nonsens-Korrelation“

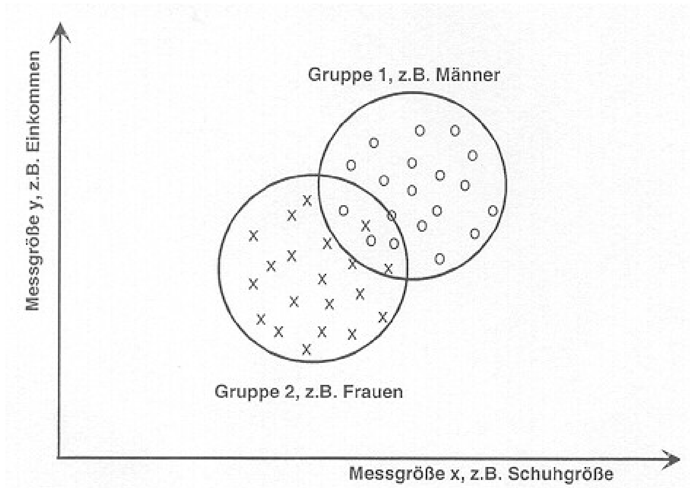


Korrelation zwischen der Bevölkerungszahl in Oldenburg und der Anzahl Störche in den 30er Jahren (nach Box, Hunter, Hunter /1/)

Sortiereffekt



Inhomogenitätskorrelation



Gefahren bei der Interpretation, z.B. „Scheinkorrelationen“

II. Beschreibende Statistik
II.3 Statistische Maßzahlen

II.3.4 Lineare Regression und Korrelation

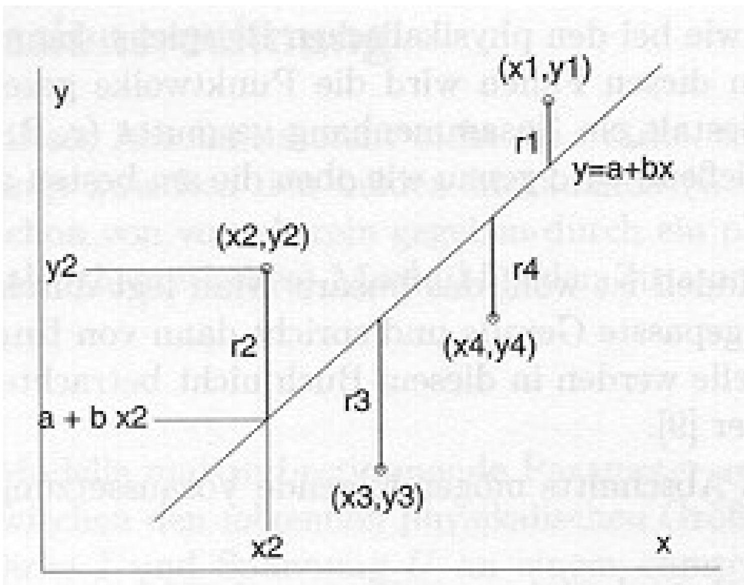
Falls das Streudiagramm einen linearen Zusammenhang rechtfertigt, kann man durch den Punkthaufen eine Gerade legen, die „möglichst gut“ zu den Daten passt.

Regressionsrechnung (nur linear)

- Lineares Modell vorhanden für den Zusammenhang zwischen X und Y
- Messreihe aus n Wertepaaren $(x_i;y_i)$, $y=a+bx$ „optimieren“

II. Beschreibende Statistik
II.3 Statistische Maßzahlen

II.3.4 Lineare Regression und Korrelation



$$S(a,b) := \sum_{i=1}^{n} r_i^2$$

$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \overset{!}{=} \min$$

$$\hat{b} = \frac{s_{xy}}{s_x^2}; \quad \hat{a} = \overline{y} - \hat{b}\overline{x}$$

„Regressionsparameter“

$\boxed{y = \hat{a} + \hat{b}x}$ „Regressionsgerade“

„Residuen“ (Reste) $r_i = y_i - (\hat{a} + \hat{b}x_i); \quad \sum_{i=1}^{n} r_i = 0$

II.3.4 Lineare Regression und Korrelation

„Bestimmtheitsmaß“:

$$B_{xy} = r_{xy}^2$$

Eigenschaften:

$$\boxed{\begin{aligned} & -1 \le r_{xy} \le +1; \quad 0 \le B_{xy} \le 1 \\ & r_{xy} = +1 \Leftrightarrow y_i = a + bx_i; \quad b > 0 \\ & r_{xy} = -1 \Leftrightarrow y_i = a + bx_i; \quad b < 0 \end{aligned}}$$

Beliebte Missverständnisse:

- r_{xy} sagt nichts über die Größe der Geradensteigung aus!
- $r_{xy}=0$ (unkorreliert) bedeutet nur, dass zwischen X und Y kein **linearer** Zusammenhang herrscht, andere Abhängigkeiten sind möglich!
- r_{xy} nahe bei +/-1 bedeutet keinen kausalen Zusammenhang!
- r_{xy} nur für quantitative Merkmale

FH Aachen
Fachbereich Medizintechnik und Technomathematik
Prof. Dr. Horst Schäfer
Heinrich-Mußmann-Str. 1
52428 Jülich
T +49. 241. 6009 53927
horst.schaefer@fh-aachen.de
www.fh-aachen.de