DIH SÜD - DIGITAL INNOVATION HUB SÜD

Deskriptive Datenanalyse

Wie man Unternehmensdaten deskriptiv und explorativ analysiert

07.11.2022

Der DIH SÜD wird unterstützt von:

FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Deskriptive Datenanalyse

Wie man Unternehmensdaten deskriptiv und explorativ analysiert

DIPL.-ING. HERMANN KATZ

Institut für Wirtschafts- und Innovationsforschung POLICIES

Forschungsgruppe „Datenanalyse und statistische Modellierung“

JOANNEUM RESEARCH Forschungsgesellschaft mbH

hermann.katz@joanneum.at

Graz, 07. November 2022

Programm

- Motivation und Vorbemerkungen
- Einführung in die Darstellung von Datenmaterial
- Vorgangsweise / Strategie bei datengestützten Fragestellungen
- Grundlagen der deskriptiven Statistik
- Graphische Darstellungsvarianten
- Kennzahlen und zugehörige Interpretationen
- Vergleich zwischen deskriptiver Statistik und Inferenzstatistik
- Zusammenfassung

Zeitplan

09:00 – 09:30: Vorstellung und Erwartungshaltungen

09:30 – 10:00: Einführung in die Darstellung von Datenmaterial

10:00 – 11:00: Vorgangsweise bei datengestützten Fragestellungen

11:00 – 11:30: Kaffeepause

11:30 – 13:00: Grundlagen der deskriptiven Statistik

13:00 – 14:00: Mittagspause

14:00 – 15:00: Graphische Darstellungsvarianten

15:00 – 15:30: Kaffeepause

15:30 – 16:15: Kennzahlen und zugehörige Interpretationen

16:15 – 16:30: Vergleich zwischen deskriptiver Statistik und Inferenzstatistik

16:30 – 17:00: Anwendungsbeispiel, Zusammenfassung und Resümee

Vorstellung

- Hermann Katz – Studium der Technischen Mathematik / Statistik
- Seit über 25 Jahren im Bereich Datenanalyse tätig
- Direktorstellvertreter des Instituts für Wirtschafts- und Innovationsforschung - POLICIES
- Leitung der Forschungsgruppe Datenanalyse bei JOANNEUM RESEARCH
- Leitung von vielen anwendungsorientierten Projekten
- Autor bzw. Mitautor von zahlreichen Publikationen
- Langjähriger Vortragender an Universitäten und Fachhochschulen

Einführung in die Darstellung von Datenmaterial

Der DIH SÜD wird unterstützt von:

Motivation

- Digitalisierung führt zu Datenüberflutung
- Analyse von Daten wird zum Erfolgsfaktor für Unternehmen
- Sind aber tatsächliche große Datenbestände der Schlüssel zum Erfolg?
- Informationen werden aus Daten extrahiert
- Zur Verbesserung von Prozessen und Produkten sind meist nur wesentliche Informationen nötig!
- Einsatz von statistischen Methoden ist bei der möglichen Reduktion der Datenmenge ein wichtiges Instrument
- Big Data – viele Daten beinhalten oft große Redundanzen!
- Smart Data – Reduktion auf das Wesentliche ist Erfolgsfaktor!

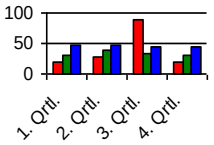
Vorbemerkungen

- Daten sind Mittel zum Zweck
- Datengewinnung ist aber nur ein Teil bei der Analyse von Daten
- Fragestellung steht im Vordergrund
- Die Reduktion von großen Datenmengen auf aussagekräftige kleiner Submengen keine triviale Angelegenheit

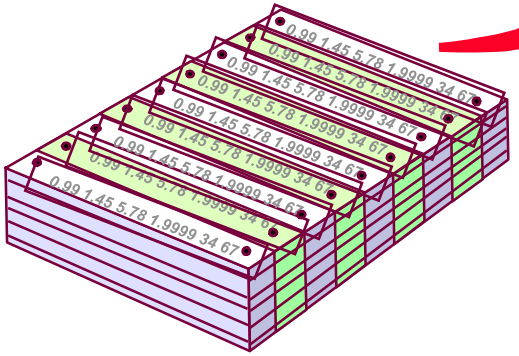
  → Ohne Erfahrung sollte man immer Fachleute einbeziehen

Daten ≠ Information

Information



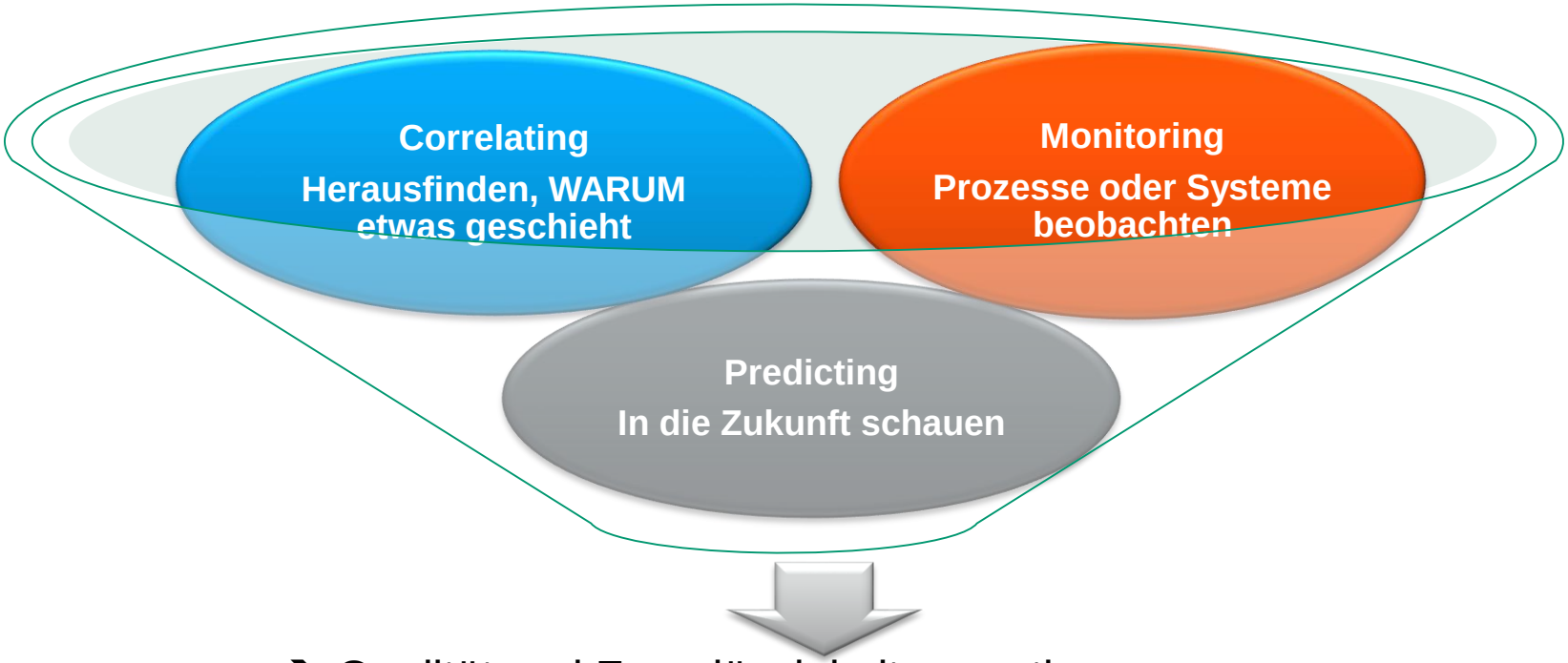
Statistische Werkzeuge



Datenanalyse als Teil der Unternehmenskultur

- Datenanalyse sollte integraler Bestandteil in Unternehmen werden
  - Analyse von Daten generiert Wissen
  - Wissen verbessert Marktposition
  - Alleinstellungsmerkmale können identifiziert werden
  - Vorteile gegenüber Mitbewerbern
- Detaillierte Analyse der Daten auf Basis von konkreten Fragestellungen
  - Deskriptive und explorative Analyse der Daten
  - Einsatz von statistischen Methoden
  - Entwicklung von zielführenden Softwaretools
- Erworbenes Wissen führt zu Verbesserungen
  - Verbesserung von Produktqualität
  - Einleitung von konkreten Präventivmaßnahmen
  - Steigerung der Ressourceneffizienz

Die Möglichkeiten statistischer Methoden …



- ➔ Qualität und Zuverlässigkeit garantieren
- ➔ Durchsatz erhöhen
- ➔ Kosten reduzieren
- ➔ Verkaufszahlen und Gewinne steigern
- ➔ Bessere Entscheidungen treffen

Deskriptive Statistik

Deskriptive (beschreibende) Statistik:

- Instrumentarium zur Beschreibung von Daten
- Vorstufe zur schließenden Statistik

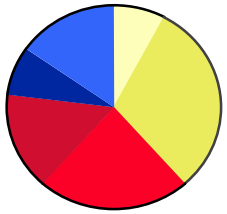
Ziel:

Beschreibung, Strukturierung, Verdeutlichung, Darstellung umfangreichen, unübersichtlichen Datenmaterials

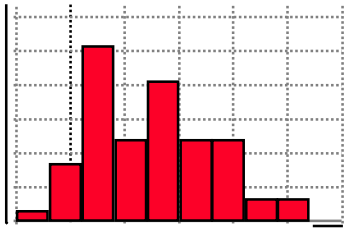
Methoden:

- Graphische Darstellungen
- Kennzahlen (Maßzahlen)

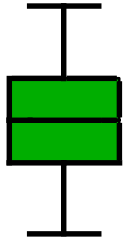
Deskriptive Statistik



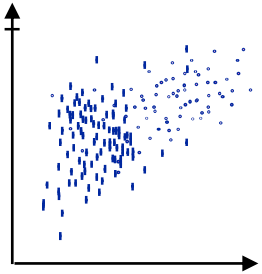
Piechart



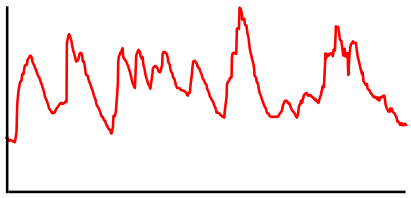
Histogramm



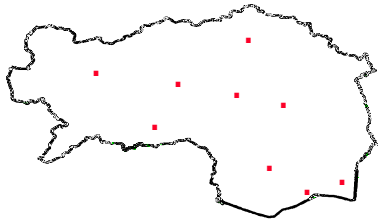
Boxplot



Scatterplot



Zeitreihe

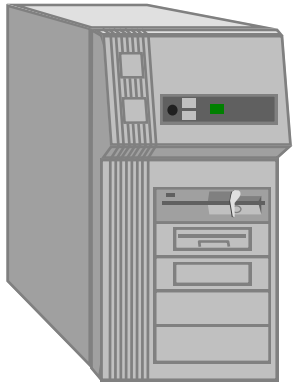


Map

Definitionen

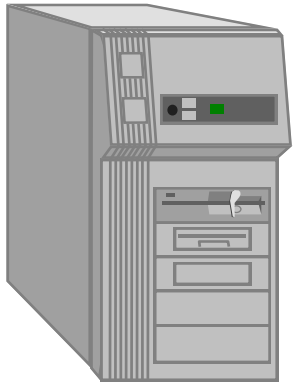
- Datengewinnung
  - durch Befragung von Personen, Messungen
- Untersuchungseinheiten
  - befragte Personen oder Objekte, an denen Messungen durchgeführt werden
- Merkmal
  - Größen, auf die sich die Fragen oder Messungen beziehen
- Merkmalswert
  - Festgestellter Wert des Merkmals an der Untersuchungseinheit

Merkmalswerte insgesamt = Daten



Beispiel

- Datengewinnung
  - **Fragebogen** auf Webpage
- Untersuchungseinheiten
  - **Kunden**, die eine Bestellung vornehmen
- Merkmal
  - **Kundenzufriedenheit**
- Merkmalswert
  - **Zustimmung zur elektronischen Zusendung von Infos (ja/nein)**
  - **Beurteilung anhand einer Skala von 1 - 6**
  - **Zeit, die Kunden auf der Webpage verbringen (sek)**



Art der Merkmale ist entscheidend!

VORGANGSWEISE BEI DATENGESTÜTZTEN FRAGESTELLUNGEN

DIHSÜD

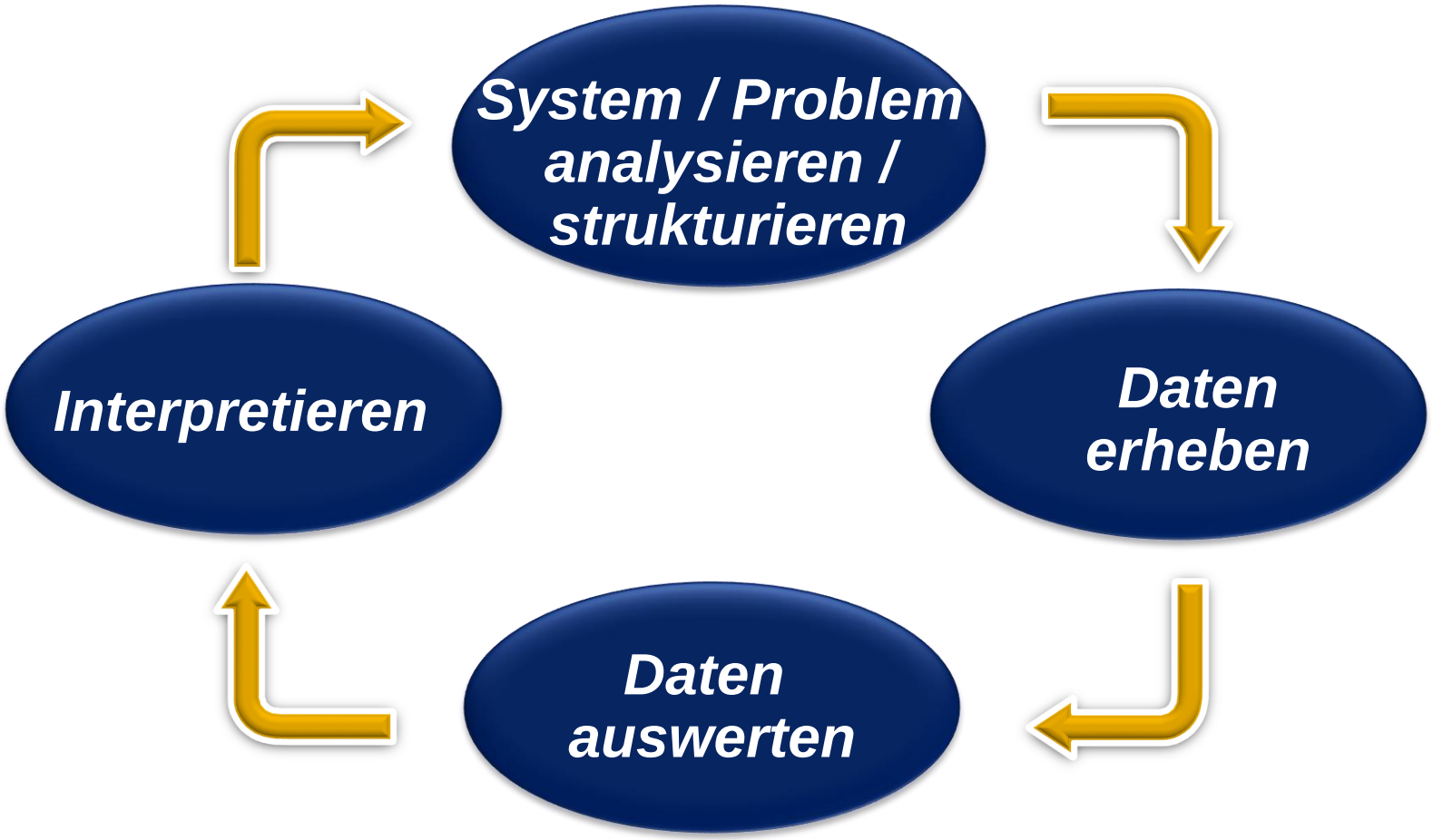
Der DIH SÜD wird unterstützt von:

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Wissenschaftliche Vorgangsweise in der Angewandten Statistik



Unser Vorgehensmodell …

Ask

- Was ist die spezifische Fragestellung?
- Analysieren der Anforderungen
- Analysieren des Prozesses bzw. des Systems

Get

- Daten aus bestehenden Quellen gewinnen
- Neue Datenquellen erschließen
- Daten fusionieren und prozessieren

Explore

- Plausibilität prüfen
- Analysieren
- Visualisieren

Model

- Detektion
- Klassifikation
- Prognose
- Optimierung
- Validierung

Communicate

- Neue Erkenntnisse vermitteln
- Präsentation
- Report
- Software-Tool

Systemanalyse - Planung

- Formulierung der sachspezifischen Fragestellung
  - Ziele festlegen
  - Fachwissen mit statistischem Know-how verbinden
- Merkmale festlegen
  - Skalierung, Eigenschaften
- Datenquellen konkretisieren
  - Fragebogendesign
  - Pretest
  - Projektdurchführung
- Grundgesamtheit – Stichprobe
  - Auswahl an einer repräsentativen Stichprobe

Datengewinnung

- Erhebungs- oder Versuchsplanung
  - Analyse der Auswahlgrundlage
  - Fragebogendesign
  - Pretest
  - Versuchsplanung
  - Festlegung des Stichprobenumfanges
  - Organisatorischer Ablauf der Datenerhebung – ev. Pilotstudie
- Datensammlung
- Dateneingabe – Online-Befragungen
- Überprüfung der Korrektheit der Daten

Statistische Auswertung

- Kritische Analyse der Urdaten
- Deskriptive und exploratorische Datenanalyse
  - Tabellen
  - Grafiken
  - Kennzahlen
- Inferenzstatistische Aussagen
  - Aussagen bezüglich der Grundgesamtheit
  - Überprüfung von Vermutungen (Hypothesen)
  - Modellierung z. B. Regression

Sachspezifische Entscheidungsfindung

- Aufbereitung der statistischen Ergebnisse für Entscheidungsfindung
- Sachspezifische Interpretation der Ergebnisse
- Ableitung von Maßnahmen
- Ev. Detailstudien

KFV Standard Reporting

Ask

Auswertung und Darstellung bemerkenswerter Aspekte der österr. Unfallstatistiken

Get the Data

KFV, Statistik Austria, internationale Datenbanken

Explore the Data

Tabellen und Grafiken

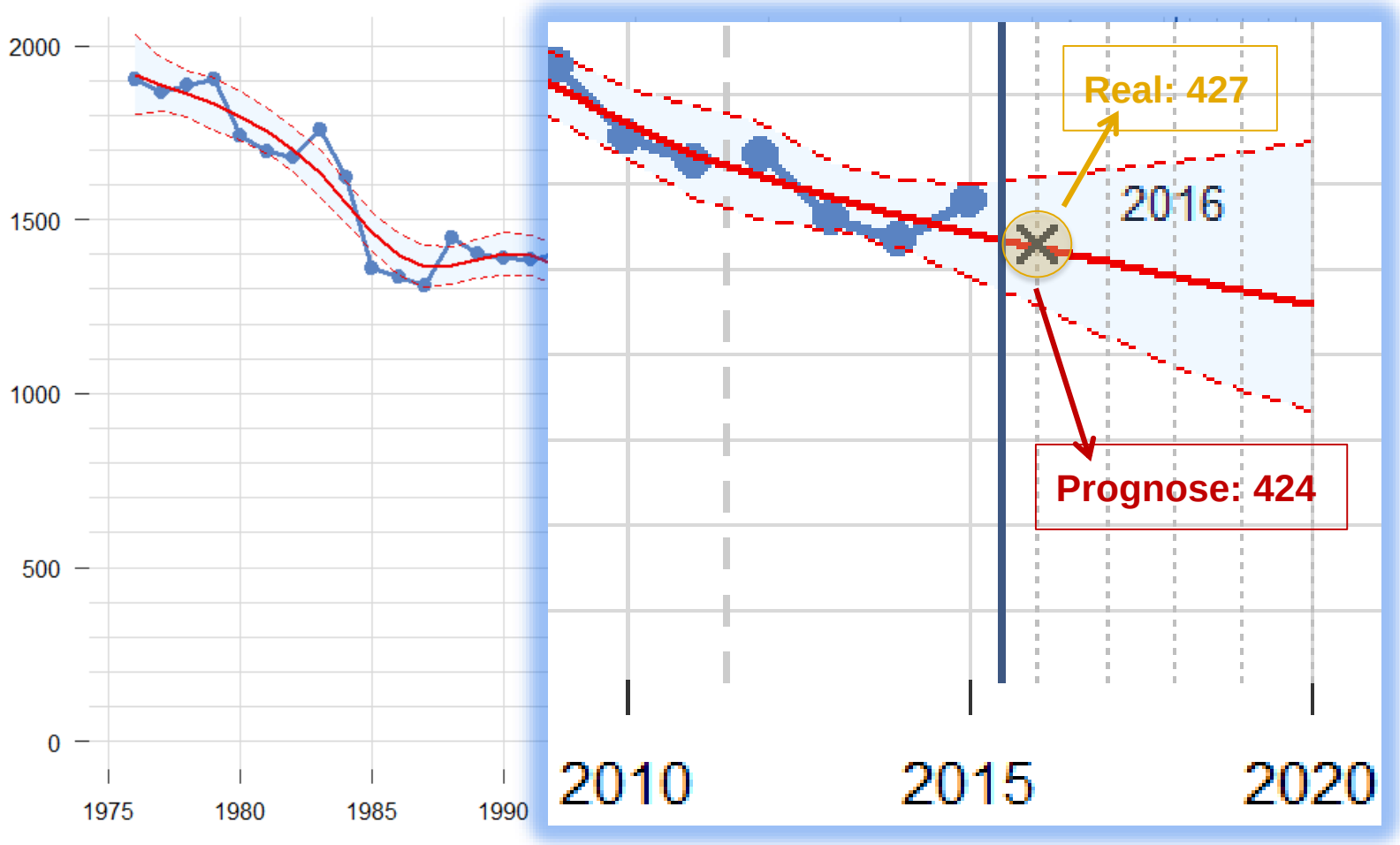
Model the Data

Vorhersagen in Zukunft

Communicate and visualize the result

Bericht

KFV Standard Reporting - Prognose der Todesfälle im Straßenverkehr



Grundlagen der deskriptiven Statistik

Der DIH SÜD wird unterstützt von:

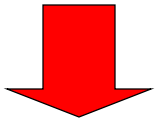
FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

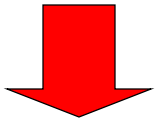
Gliederung der Statistik

Beschreiben



Deskriptive Statistik

Schlüsse ziehen



Inferenzstatistik

Deskriptive Statistik

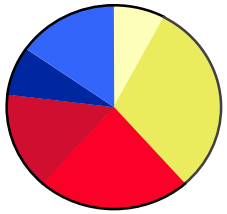
- Deskriptive (beschreibende) Statistik
  - Instrumentarium zur Beschreibung von Daten
  - Vorstufe zur schließenden Statistik

Ziel:

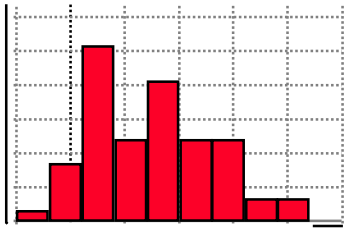
Beschreibung, Strukturierung, Verdeutlichung, Darstellung umfangreichen, unübersichtlichen Datenmaterials

- Methoden:
  - Grafische Darstellungen
  - Kennzahlen (Maßzahlen)

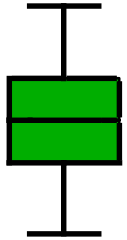
Unterschiedliche Grafiken



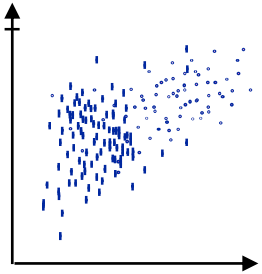
Piechart



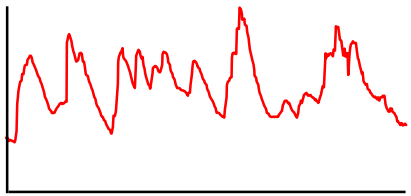
Histogramm



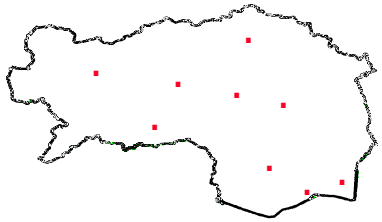
Boxplot



Scatterplot



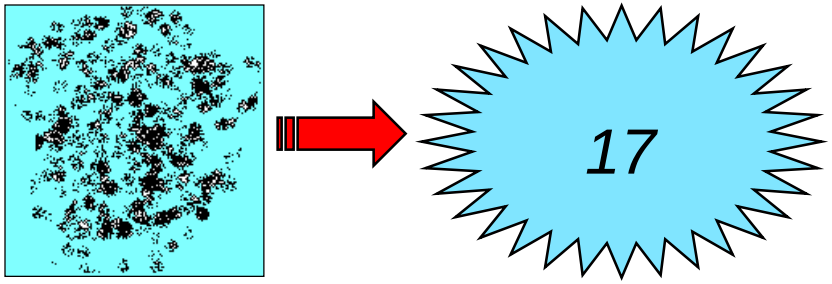
Zeitreihe



Map

Allgemeine Kennzahlen

Sinn von Kennzahlen ...



- *Reduzierung von Komplexität*
- *Verdichtung von Information*

... die Sache auf den Punkt bringen!

Skalierung

Skalentyp	Nominalskala
Definierte Relationen und Operationen	$\neq =$
Zulässige Transformationen	bijektive Transformation wie Umbenennen, Permutation
Beispiele für Merkmale	Familienstand, Geschlecht, Postleitzahl, Artikelbezeichnung, Religionszugehörigkeit
Merkmalsausprägung	Namen, Symbole, Codes

rot blau grün

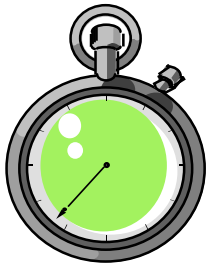
Skalierung

Skalentyp	Ordinalskala
Definierte Relationen und Operationen	$\neq = < >$
Zulässige Transformationen	isotone oder rangerhaltende Transformationen
Beispiele für Merkmale	Zeugnisnoten, Sozialstatus, Produktgüteklassen, Mercalli-Erdbebenskala
Merkmalsausprägung	Ordinalzahlen (i.d.R. ganze Zahlen)

gut
besser
G . . . . .

Skalierung

Skalentyp	Kardinalskala (Metrische Skala)
Definierte Relationen und Operationen	$\neq = < > + - * /$
Zulässige Transformationen	Ähnlichkeitstransformationen $y=ax$ mit $a>0$
Beispiele für Merkmale	Temperatur, geographische Höhe, Messungen im cm-g-sec-System, Anzahlen
Merkmalsausprägung	reelle Zahlen (stetig oder diskret)



Skalierungsart

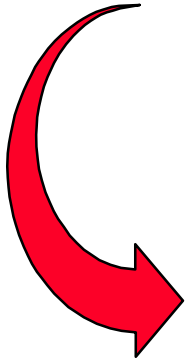
- Militärdienstgrad
- Alter
- Kinderzahl
- erlernter Beruf
- Zensur
- Religionszugehörigkeit
- Gewicht
- Güteklasse von Obst
- Einkommen
- Vereinszugehörigkeit

Klassifizierung von Merkmalen

- Stetige Merkmale
  - Ausprägungen können jeden beliebigen Wert in einem Bereich annehmen
  - metrische Merkmale
- Diskrete Merkmale
  - Ausprägungen können nur endlich viele Werte in einem Bereich annehmen
  - nominale und ordinale Merkmale

Grundgesamtheit und Stichprobe

- Wie komme ich zu meinen Untersuchungseinheiten?
- Wie wähle ich sie aus?
- Was ist eine Stichprobe?
- Was ist eine repräsentative Stichprobe?



Abgrenzung der Gesamtheit
aller Untersuchungseinheiten

Grundgesamtheit / Population

Population: Gesamtheit aller gleichartigen statistischen Objekte, die hinsichtlich eines Merkmales untersucht werden.

Prozentueller Anteil an Arbeitslosen in Österreich

- Grundgesamtheit?
  - Arbeitsfähige Bevölkerung zwischen 17 und 65, die in Österreich leben
  - Arbeitsfähige Bevölkerung zwischen 17 und 65, die Österreicher sind
  - Gesamtbevölkerung

Trügerische Beispiele

„Dann hätte ich gerne noch ein Bier“.

Reaktion eines Alkoholikers auf die Information, daß 70 % aller Verkehrsunfälle im nüchternen Zustand verursacht werden.

„Der weitaus größte Teil von Gewaltverbrechen in den USA geschieht in Küche, Wohn- und Schlafzimmern .... in New York schläft man daher nachts sicherer im Central Park als zu Hause im Bett“.

Stichprobe

- Stichprobe

  Auswahl von Untersuchungseinheiten aus der Grundgesamtheit.

Repräsentativität: In der Stichprobe muß die gesamte Inhomogenität der Grundgesamtheit enthalten sein.

Graphische Darstellungsvarianten

Der DIH SÜD wird unterstützt von:

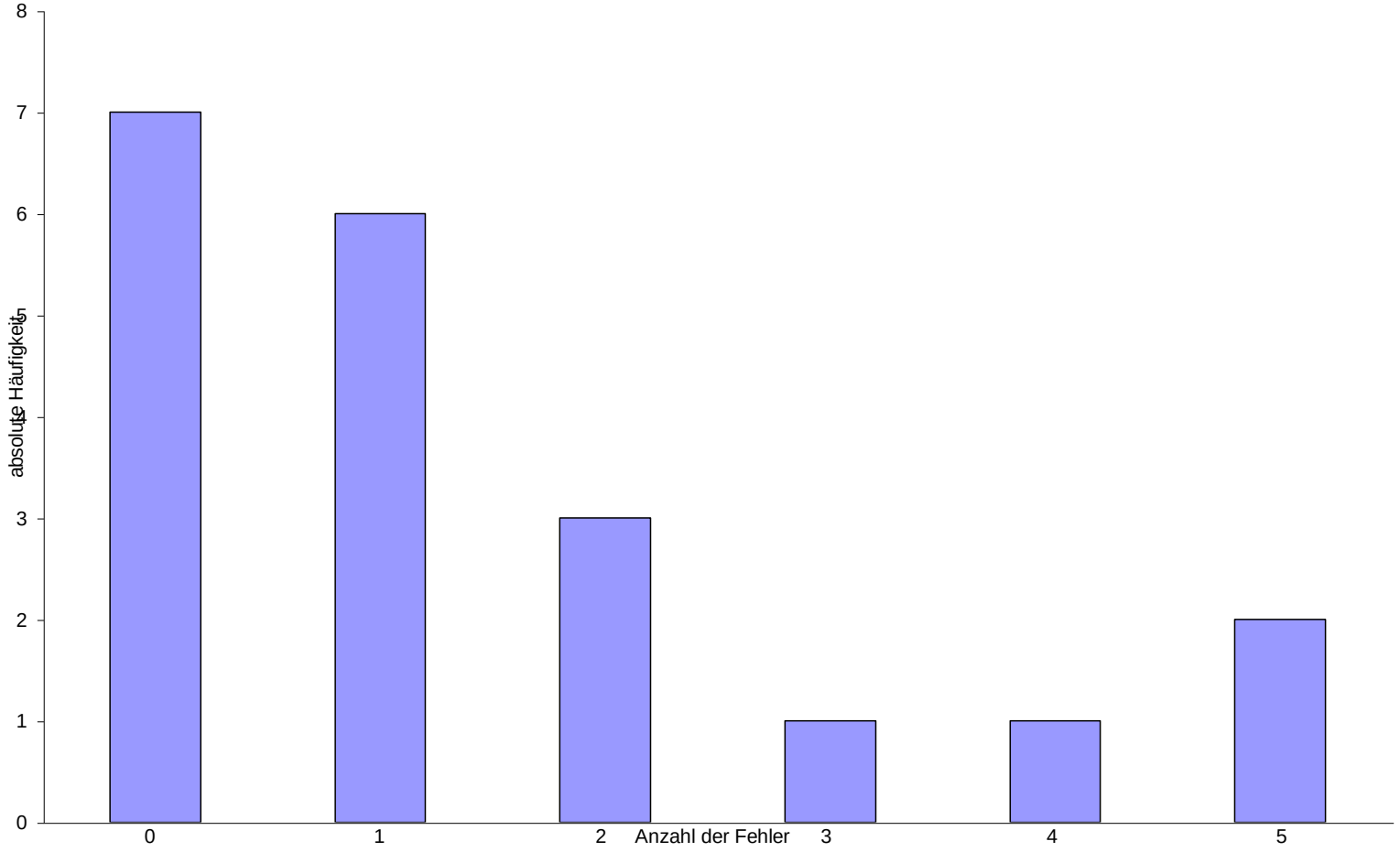
Beispiel

n = 20 Computer sollen auf Fehler untersucht werden.
Bei jedem Computer wird die Anzahl der Fehler registriert.

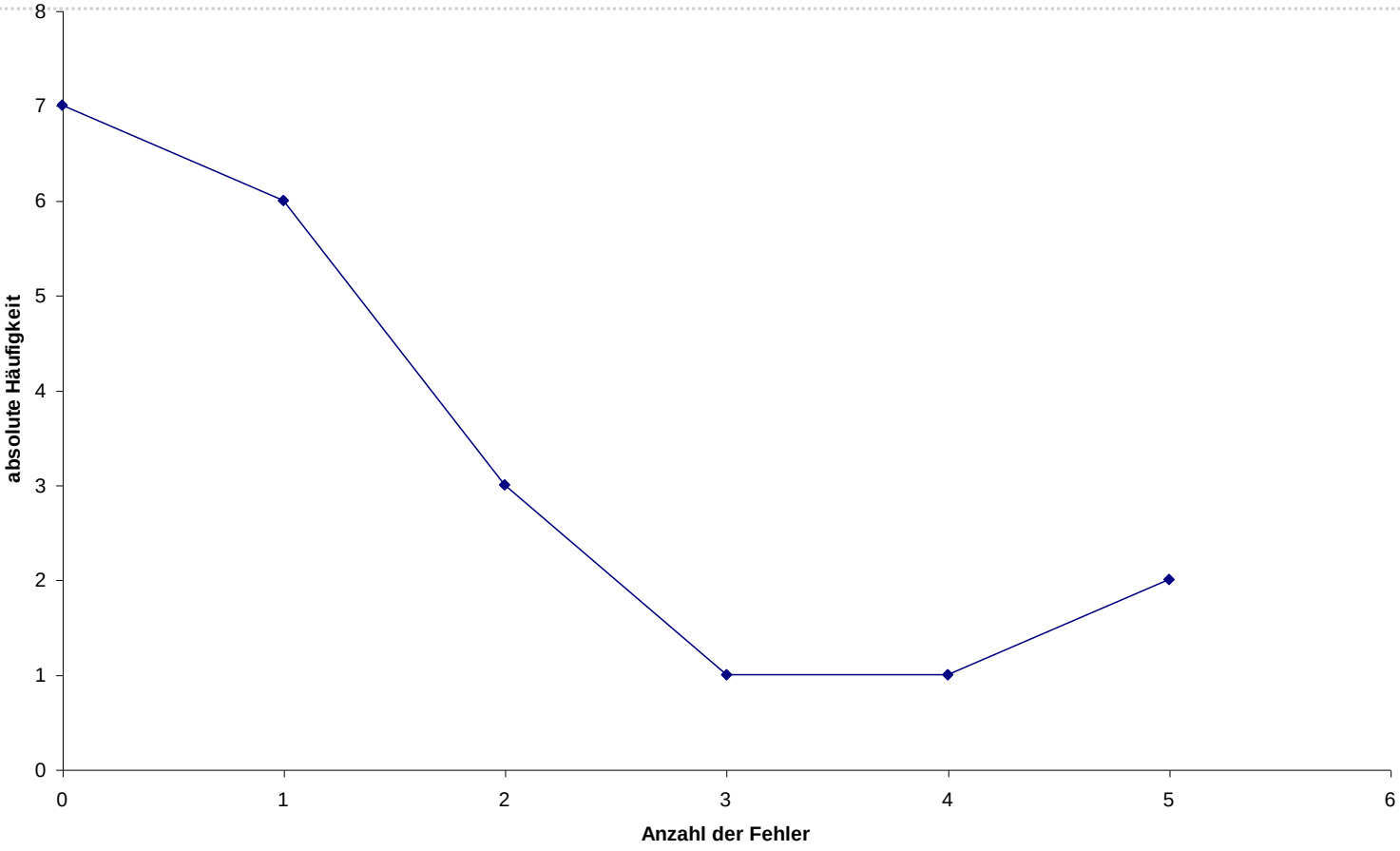
```
1,0,0,3,1,5,1,
2,2,0,1,0,5,2,
1,0,0,4,0,1
```

Anzahl der Fehler	H_j	h_j
0	7	0,35
1	6	0,30
2	3	0,15
3	1	0,05
4	1	0,05
5	2	0,10
Gesamt	20	1,00

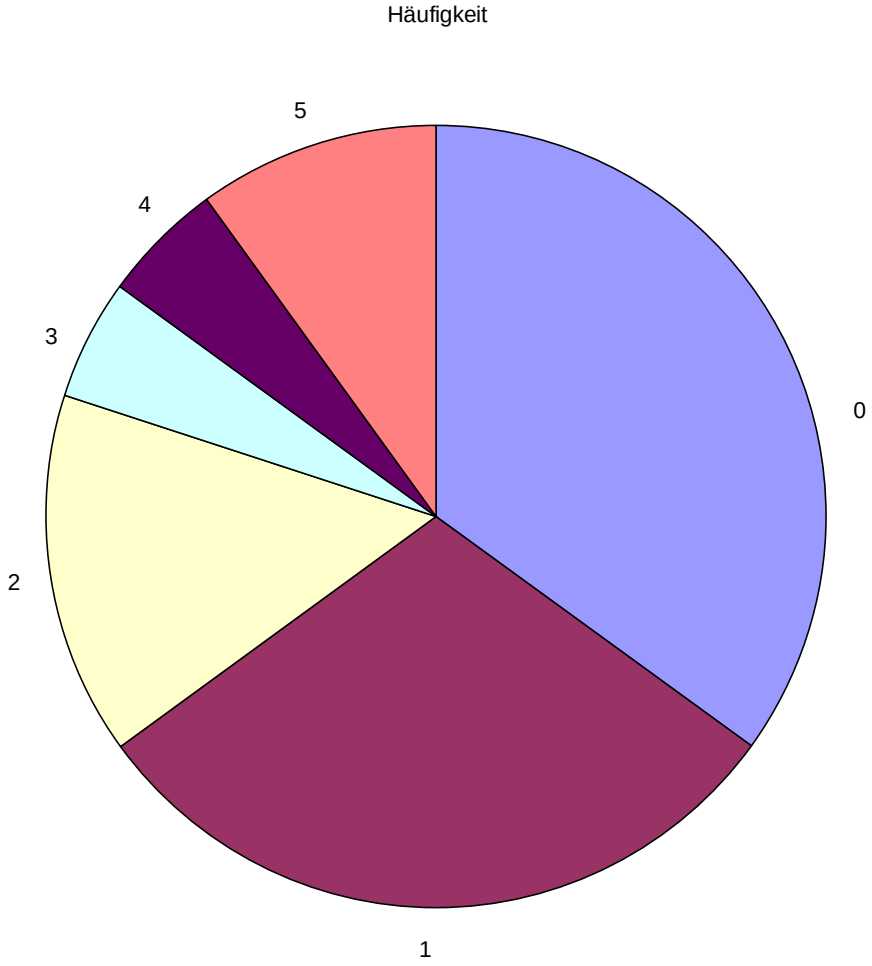
Stabdiagramm



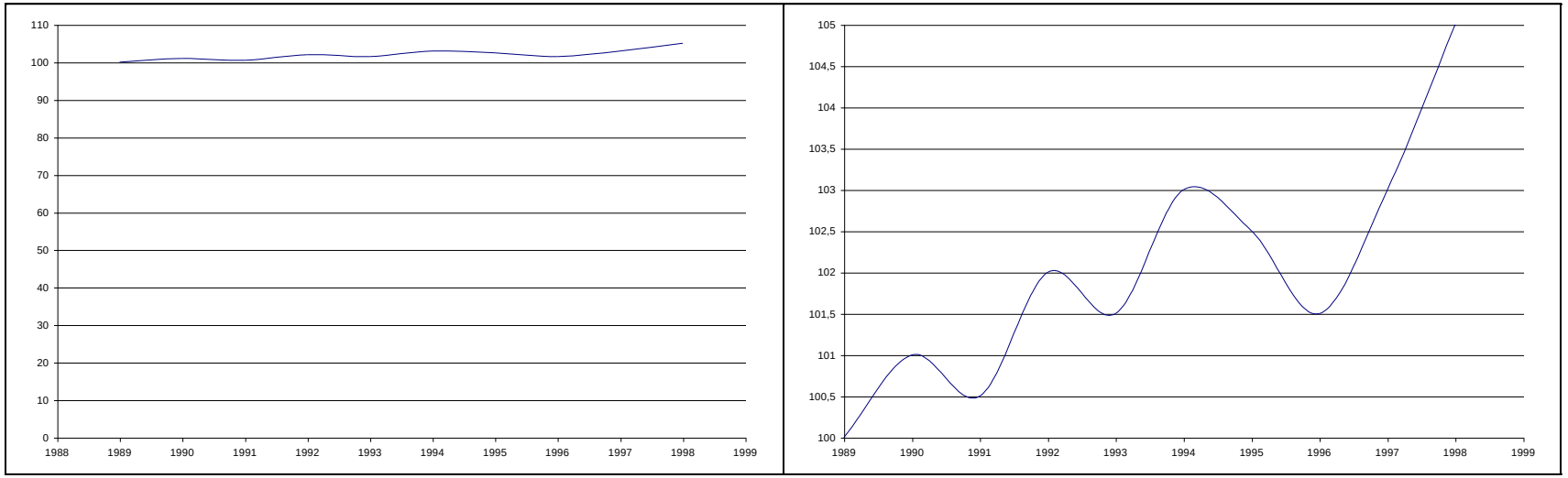
Häufigkeitspolygon



Kreisdiagramm

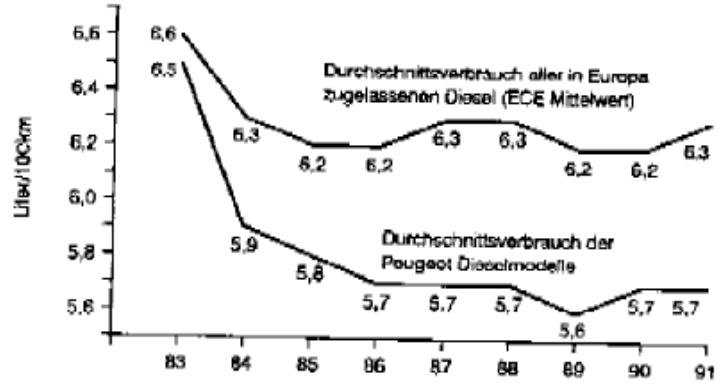


Manipulierte Grafik

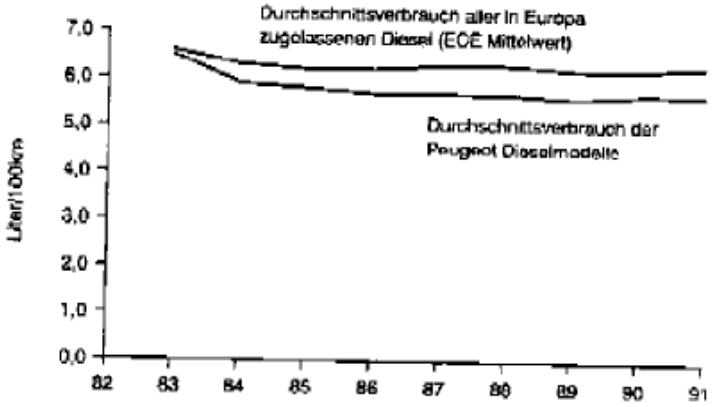


Achsenschnitt

Manipulierte Grafik



Die phänomenale Sparsamkeit der Peugeot-Motoren: Schein ...



... und Sein

Empirische Verteilungsfunktion

Absolute Summenhäufigkeit

$$\sum_{i=1}^{j} H(a_i)$$

Relative Summenhäufigkeit

$$\frac{1}{n}\sum_{i=1}^{j} H(a_i)$$

empirische Verteilungsfunktion =
Summenhäufigkeitsfunktion

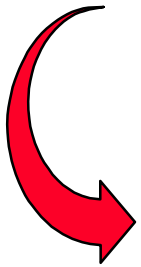
$$S_n(x) = \frac{1}{n}\sum_{i=1}^{j} H(a_i) = \sum_{i=1}^{j} h(a_i)$$

Beispiel - Summenhäufigkeitsfunktion

Anzahl der Fehler	$h(a_j)$	$S_n(x)$
0	0,35	0,35
1	0,30	0,65
2	0,15	0,80
3	0,05	0,85
4	0,05	0,90
5	0,10	1,00

Klasseneinteilung

42,8	38,5	41,6	38,9	39,1	37,1	39,4	39,6
36,9	40,2	36,8	39,1	42,4	47,5	36,3	44,9
41,2	44,5	38,5	39,3	41,9	40,6	40,1	37,4
39,3	35,6	38,3	40,6	40,1	38,6	41,9	41,6
40,4	38,4	37,0	38,5	41,8	38,9	39,7	41,9
35,7	39,0	37,6	37,0	38,2	37,3	38,6	38,5
37,6	38,5	32,1	40,8	35,8	43,9	43,5	36,3
43,5	34,5	40,8	40,6	36,2	35,9	39,1	44,9
35,6	35,8	44,5	34,8	45,9	42,4	45,9	35,3
36,8	36,1	42,3	37,2	38,4	38,8	43,8	46,8



Für stetige Merkmale (n > 50)
KLASSENEINTEILUNG

Klasseneinteilung

- Die Klasseneinteilung muß alle Beobachtungswerte umfassen (also in der ersten Version auch die ungewöhnlichen Werte bzw. Ausreißer).
- Die Klassengrenzen sind so zu wählen, daß die Beobachtungswerte eindeutig den Klassen zugeordnet werden können
- Die Klassenmitte repräsentiert die übrigen Meßwerte der Klasse.
- Je kleiner die Klassenanzahl um so größer die Klassenbreite und um so größer ist der Informationsverlust.
- Je größer die Klassenanzahl, um so mehr kommt die nicht interessierende Wirkung von Zufallseinflüssen zur Geltung.

Klasseneinteilung

- Die Erfahrung führt zu folgenden Faustregeln:

  Bezeichnet R die *Spannweite* oder *Range* (Differenz zwischen größtem und kleinstem Wert) und k die aus der Ungleichung $k \leq 5 \log 10\ n$ (n: Anzahl der Beobachtungswerte) gewonnene Klassenzahl, so schätzt man die Klassenbreite d aus $d = \frac{R}{k}$

  Faustregel: $k = \sqrt{n}$ $\quad 5 < k < 20$

  Trotzdem bleibt die Klassenanzahl subjektiv beeinflusst!

- Man vermeide (wenn möglich) offene Verteilungsenden.
- Man wähle die Klassenbreiten möglichst gleich lang.
- Null-Linie beachten.

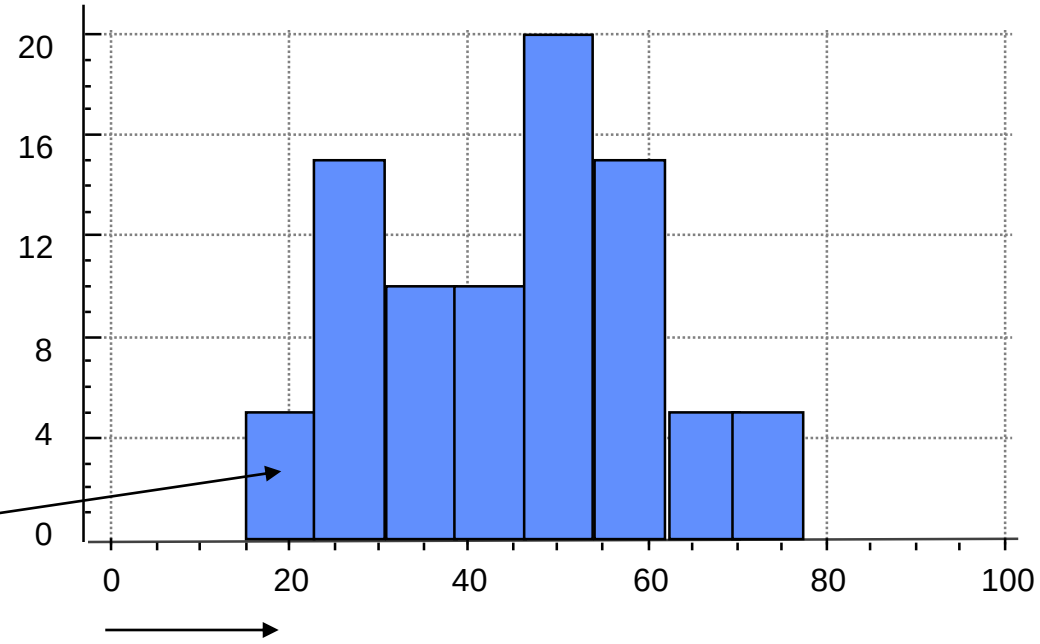
Klasseneinteilung - Beispiel

Klasse k_i		Klassenmitte	Rel. Häufigkeit	Anzahl
von [	bis)	x_i	$h(x_i)$	n_i
31.1 -	33.1	32.1	0.012	1
33.1 -	35.0	34.0	0.025	2
35.0 -	36.9	35.9	0.175	14
36.9 -	38.8	37.9	0.250	20
38.8 -	40.8	39.8	0.225	18
40.8 -	42.7	41.7	0.150	12
42.7 -	44.6	43.6	0.087	7
44.6 -	46.5	45.6	0.050	4
46.5 -	48.5	47.5	0.025	2
			1.000	80

Histogramm

Zweck: Graphische Darstellung einer Häufigkeitsverteilung

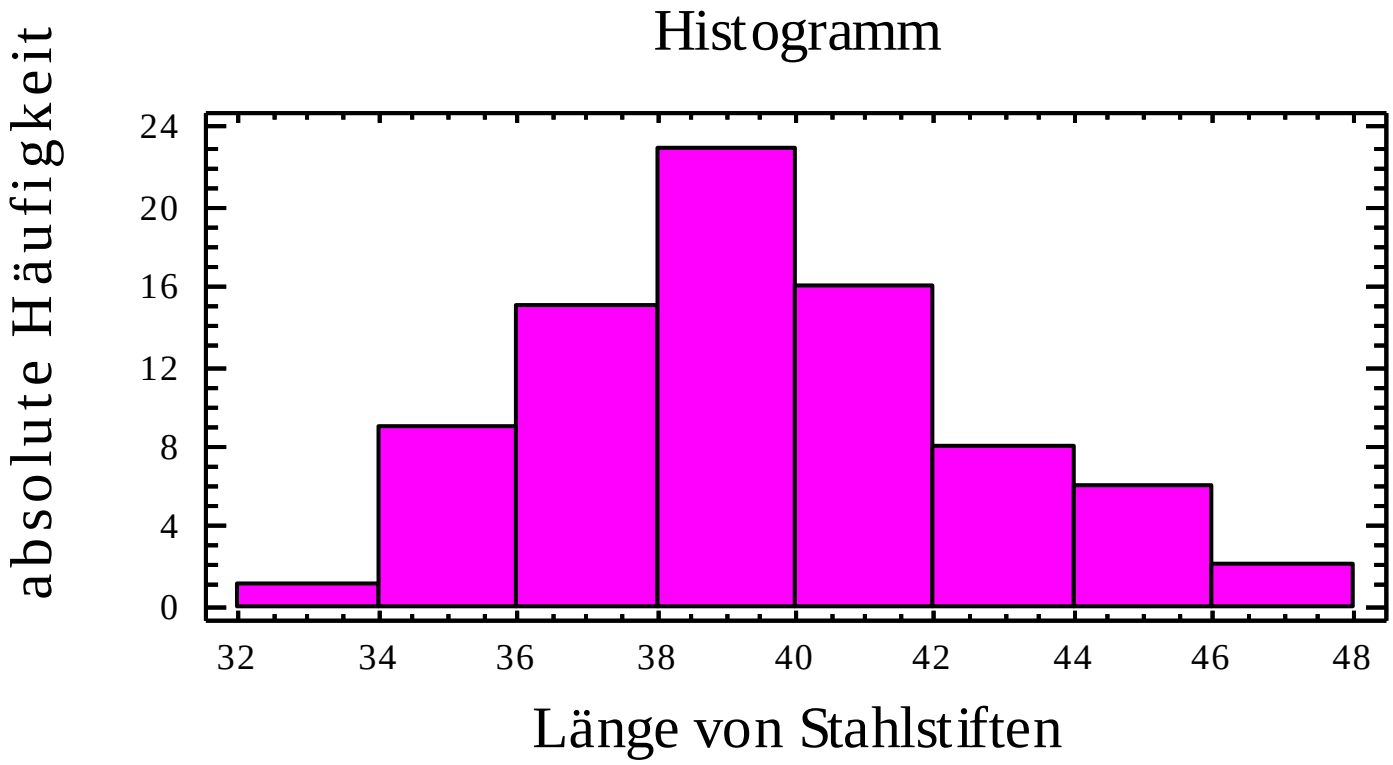
Vertikale Achse:
Absolute oder relative (%) Häufigkeit



Fläche der Rechtecke:
Proportional zur Häufigkeit in der entsprechenden Klasse

Horizontale Achse:
Wertebereich des Merkmals in Klassen eingeteilt

Histogramm - Beispiel



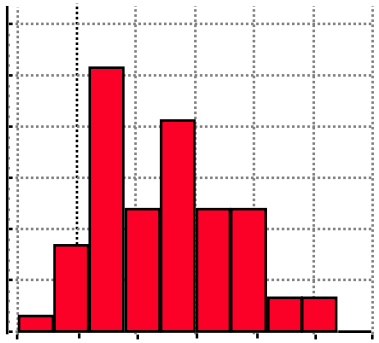
Histogramm

Beobachtungsanzahl (n) > 50: Klasseneinteilung

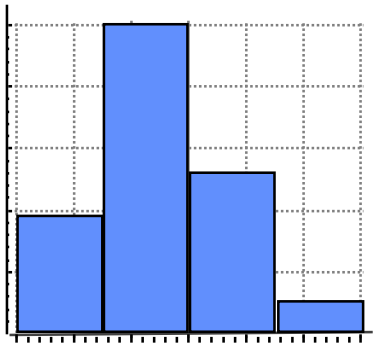
Klassenanzahl (k) und Klassenbreite (d) geeignet wählen!

Faustregeln: $k = \sqrt{n}, \quad 5 < k < 20$

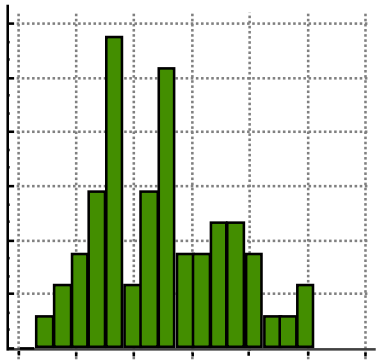
$$d = \frac{\text{größter - kleinster Wert}}{k}$$



Richtig



Falsch
zu wenige
Klassen



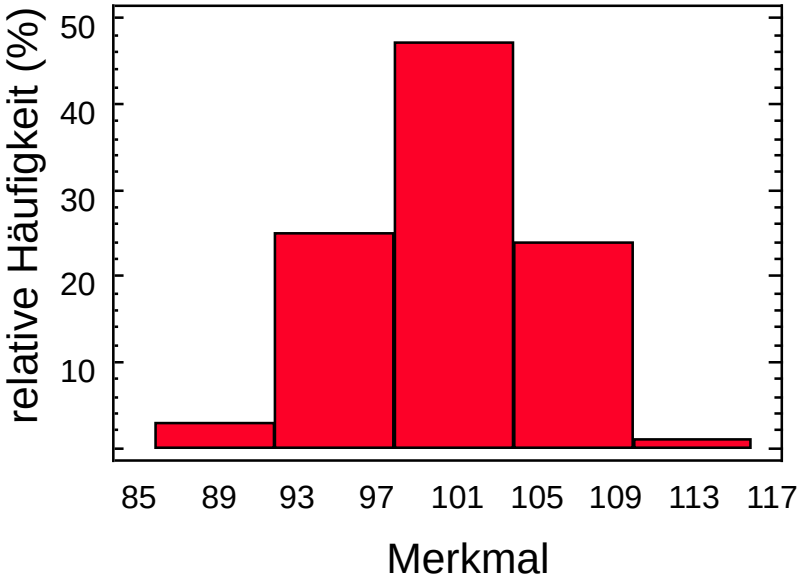
Falsch
zu viele
Klassen

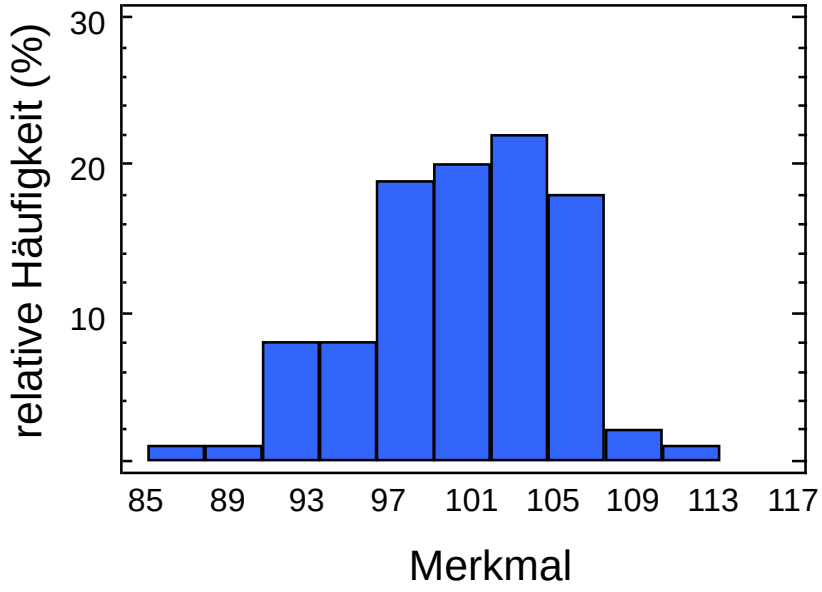
Histogramm

Achtung:

Die Klassenanzahl beeinflußt das Bild der Häufigkeitsverteilung

Selbe Daten, anderes Verteilungsbild





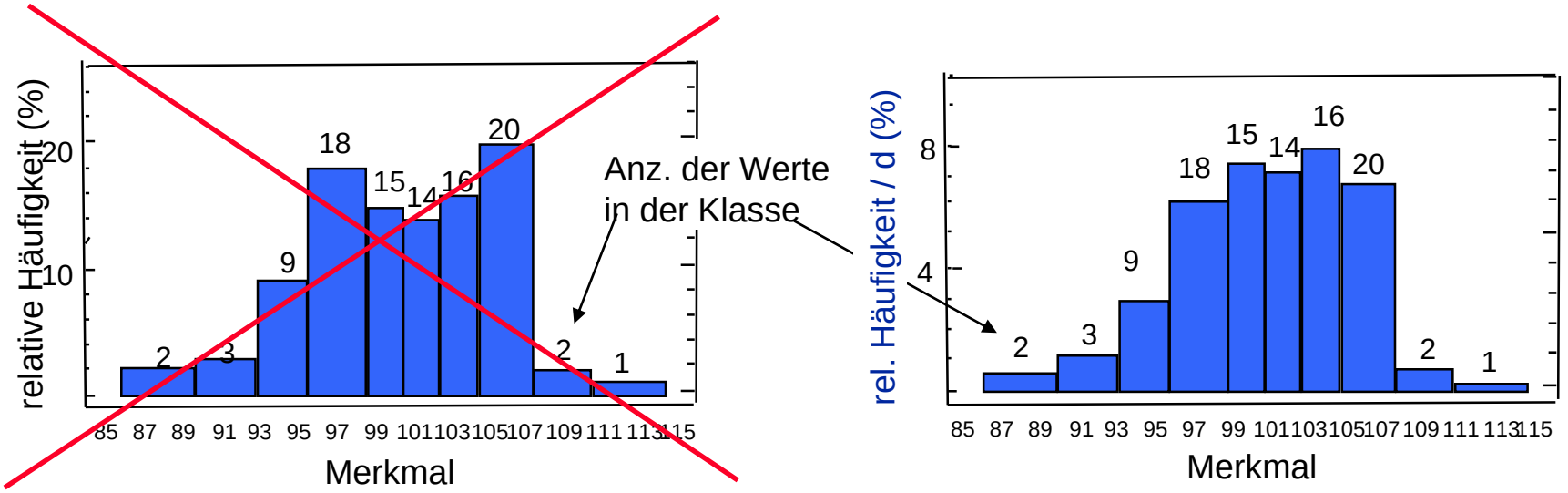
Histogramm

Flächentreue:

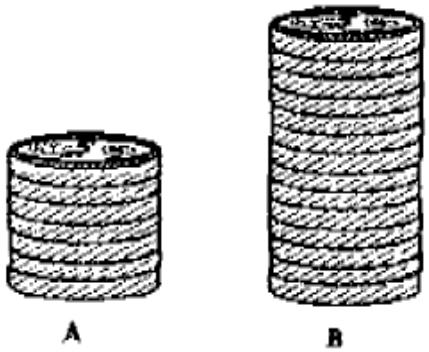
Häufigkeit muß proportional zur
Fläche des Rechteckes sein, und nicht zur Höhe!

Bei ungleicher Klassenbreite $f = \frac{\text{relative Häufigkeit}}{\text{Klassenbreite}}$

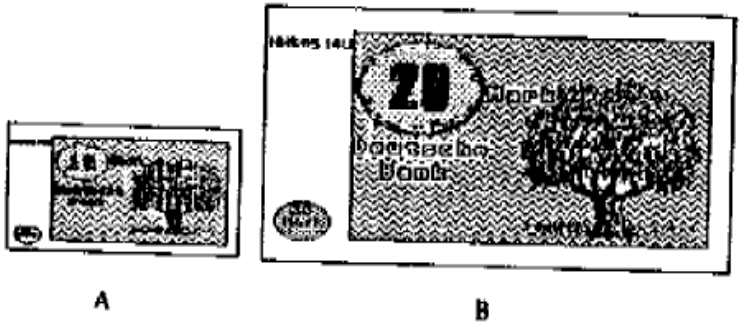
auf der vertikalen Achse auftragen!



Manipulierte Grafiken



Der rechte Stapel symbolisiert ein doppelt so hohes Einkommen



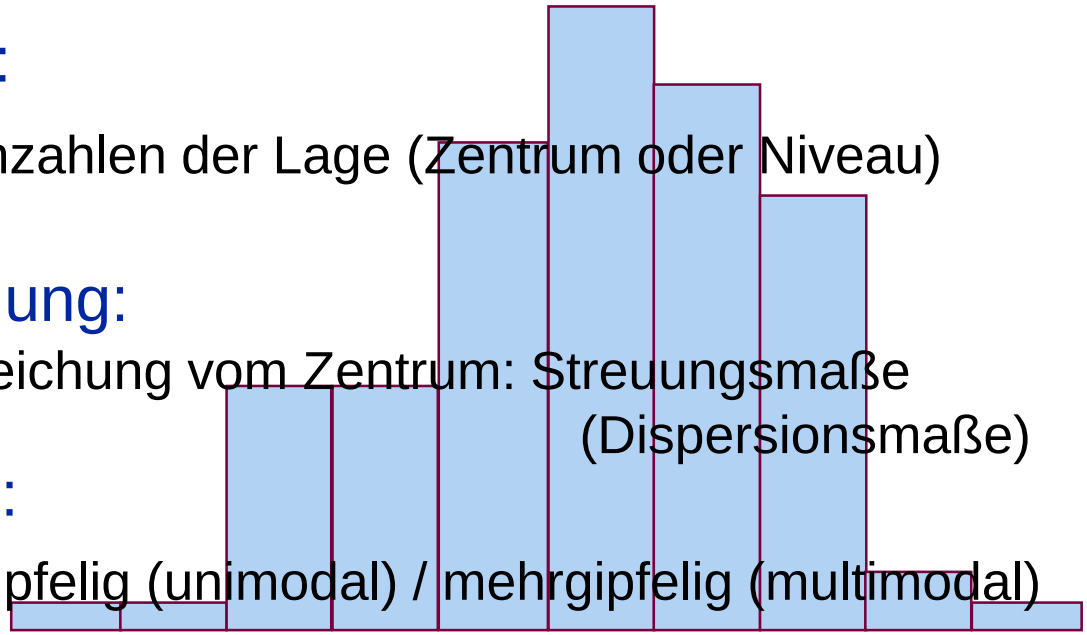
Irreführend: Die Fläche des rechten Scheins ist viermal größer als die des linken

Kennzahlen und zugehörige Interpretationen

Der DIH SÜD wird unterstützt von:

Häufigkeitsverteilungen

Charakterisierung einer Häufigkeitsverteilung



Lage:

- Kennzahlen der Lage (Zentrum oder Niveau)

Streuung:

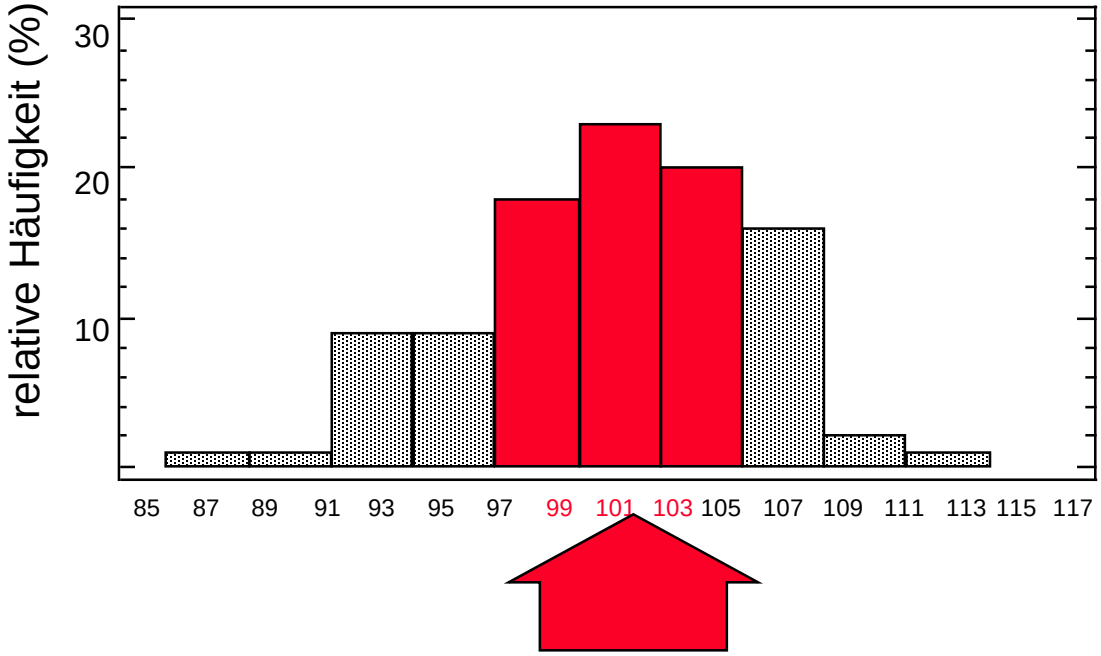
- Abweichung vom Zentrum: Streuungsmaße (Dispersionsmaße)

Form:

- Eingipfelig (unimodal) / mehrgipfelig (multimodal)
- Schiefe (Unsymmetrie): linksschief, rechtsschief, symmetrisch
- Wölbung (Steilheit, Kurtosis): steilgipfelig, flachgipfelig, normal

Lagekennzahlen

Kennzahlen (Maßzahlen) der Lage charakterisieren das Zentrum und das Niveau einer Häufigkeitsverteilung



Gegeben: Stichprobe vom Umfang n mit Einzelwerten

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

Lagekennzahlen

Arithmetisches Mittel:

Mittelwert eines metrisch skalierten Merkmales

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = (x_1 + x_2 + ... + x_n)/n$$

Eigenschaften:

- Empfindlich gegenüber außergewöhnlichen Werten (Ausreißern)
- Für mehrgipfelige und sehr schiefe Häufigkeitsverteilungen nicht geeignet
- Bei diskreten Merkmalen muß der berechnete Mittelwert in der Realität nicht auftreten
- Arithmetisches Mittel ist nicht immer sinnvoll

Lagekennzahlen

α-Quantile (Ordnungsstatistiken):

Lagemaße einer Häufigkeitsverteilung

$$Q_\alpha = \begin{cases} x_{[j]}, & \text{falls } n\cdot\alpha \text{ keine ganze Zahl} \\ & (j: \text{auf } n\cdot\alpha \text{ folgende ganze Zahl}) \\ \frac{1}{2}\left(x_{[j]} + x_{[j+1]}\right), & \text{falls } n\cdot\alpha \text{ ganze Zahl } (j=n\cdot\alpha) \end{cases}$$

$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$: der Größe nach geordnete Beobachtungen

Eigenschaften:

- $n\alpha$ Werte der geordneten Datenmenge kleiner oder gleich Q_α, $n(1-\alpha)$ Werte größer oder gleich Q_α
- Für metrisch und ordinal skalierte Merkmale geeignet

Spezielle Quantile: Median ($Q_{0.5}$), Minimum (Q_0), Maximum (Q_1), Unteres Quartil ($Q_{0.25}$), Oberes Quartil ($Q_{0.75}$)

Spezielle Ordnungsstatistiken

- N=4, Z=1,3 :Quartile
  - Unterhalb des ersten (unteren) Quartils $Q_{1/4}$ liegen also gerade 1/4 oder 25% aller Beobachtungen der geordneten Reihe, oberhalb 3/4 oder 75%. Analog liegen unterhalb des oberen Quartils $Q_{3/4}$ 75% und oberhalb 25% aller Werte.
- N=2, Z=1: Median
  - siehe nächste Folie
- N=100, Z=1,2, ... , 99: Perzentile
  - Das Z-te Perzentil $P_{Z/100}$ ist der Wert der geordneten Reihe, unterhalb dessen gerade Z% und oberhalb (100-Z)% aller Beobachtungen liegen. Die Quartile sind damit 25% - bzw. 75%-Perzentile, der Median das 50%-Perzentil.

Lagekennzahlen

Median:

Maß für die Lage des Zentrums einer Häufigkeitsverteilung

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

$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$: der Größe nach geordnete Beobachtungen

Eigenschaften:

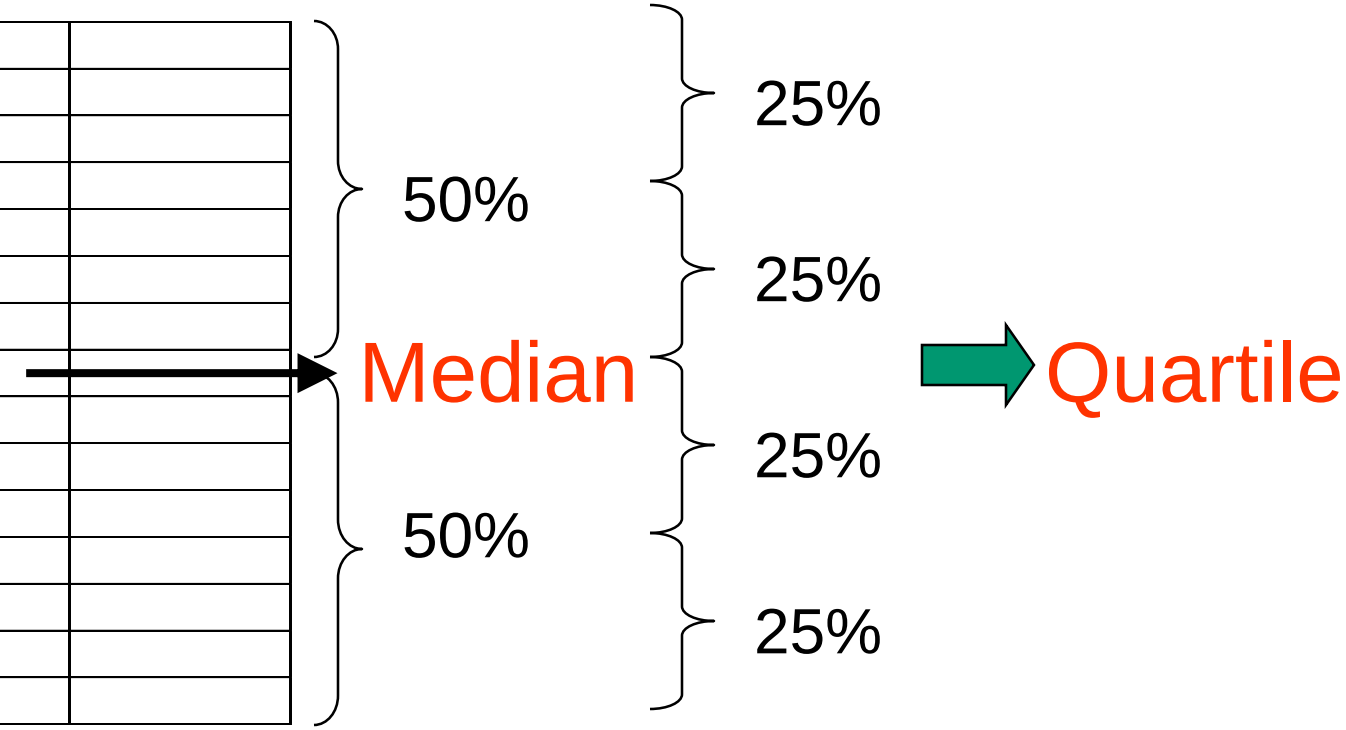
- Teilt die geordnete Datenreihe in zwei gleich große Teile
- Für metrisch und ordinal skalierte Merkmale geeignet
- Unempfindlich gegenüber außergewöhnlichen Werten (Ausreißern)
- Für mehrgipfelige und sehr schiefe Häufigkeitsverteilungen nicht unbedingt geeignet

Vom Wert zum Rang....

Vgl. Ski-Rennen: Zeiten werden (1) **geordnet** und (2) in **Rangliste** gebracht.

Bsp.: Originaldaten: 10 14 9 13 8 12 13 12 7 14 13 14 12 11 11

7	1.	
8	2.	
9	3.	
10	4.	
11	5.	
11	6.	
12	7.	
12	8.	
12	9.	
13	10.	
13	11.	
13	12.	
14	13.	
14	14.	
14	15.	



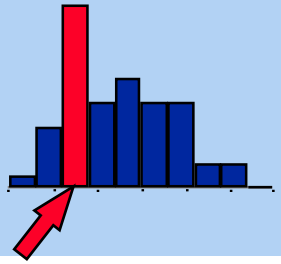
Lagekennzahlen

Modalwert (Mode):

Maß für die Lage des Zentrums einer Häufigkeitsverteilung

x_{mod}: Häufigster Wert der Beobachtungsreihe mit

$$h(x_{mod}) \geq h(a_j)$$



a_1,....., a_k: die verschiedenen Merkmalsausprägungen

$h(a_j)$: Häufigkeit der Merkmalsausprägung a_j

Eigenschaften:

- Für schiefe und mehrgipfelige (mehrere Modalwerte!) Häufigkeitsverteilungen geeignet
- Für nominal skalierte Merkmale geeignet
- Bei Klasseneinteilung: Klassenmitte der Klasse mit der größten Häufigkeit

Lagekennzahlen

Durchschnittsgehalt
40 000 S pro Monat!!
8 000 S in der Probezeit, danach mehr.

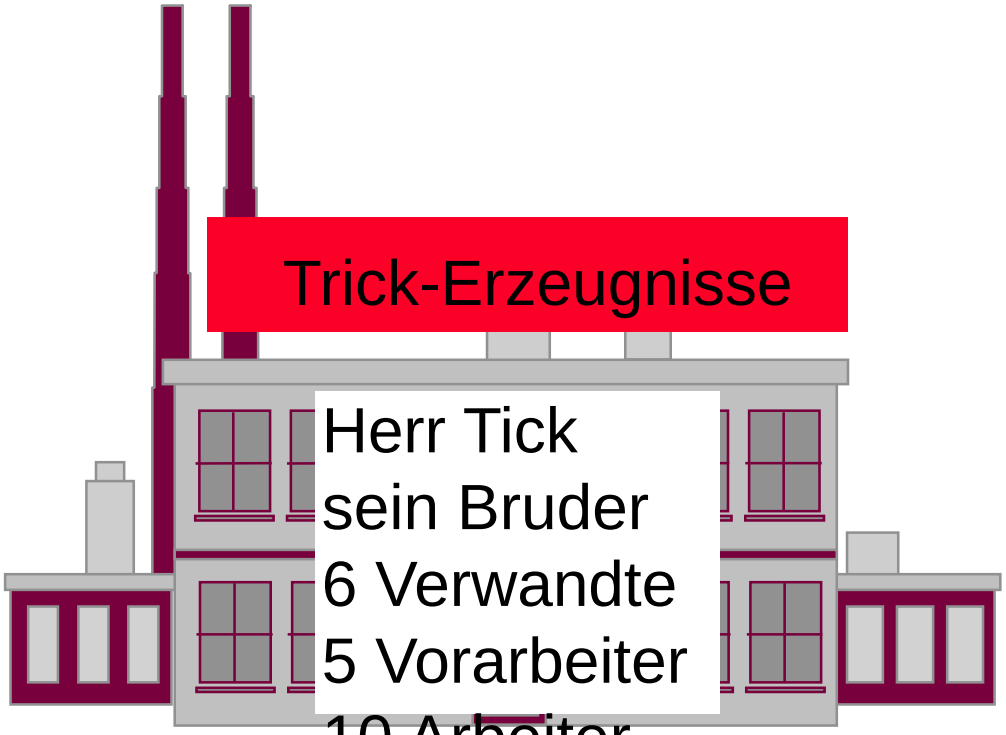
Lagekennzahlen

?!?!?

Jeder Arbeiter verdient

10 000 S

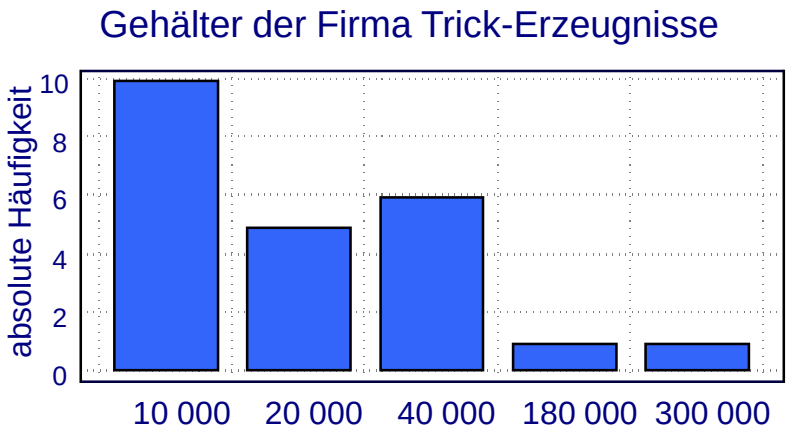
pro Monat!!



Lagekennzahlen

Lösung "Trick-Erzeugnisse"

Herr Trick:	300 000 S
Bruder:	180 000 S
6 Verwandte:	40 000 S
5 Vorarbeiter:	20 000 S
10 Arbeiter:	10 000 S

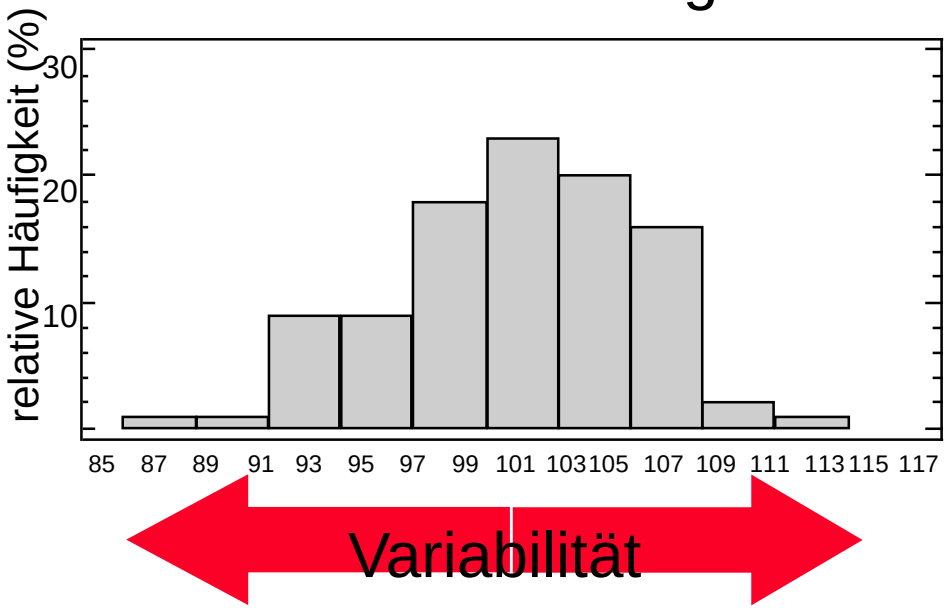


Summe pro Monat: 920 000 S

Arithmetisches Mittel:	40 000 S
Median:	20 000 S
Modalwert:	10 000 S

Streuungskennzahlen

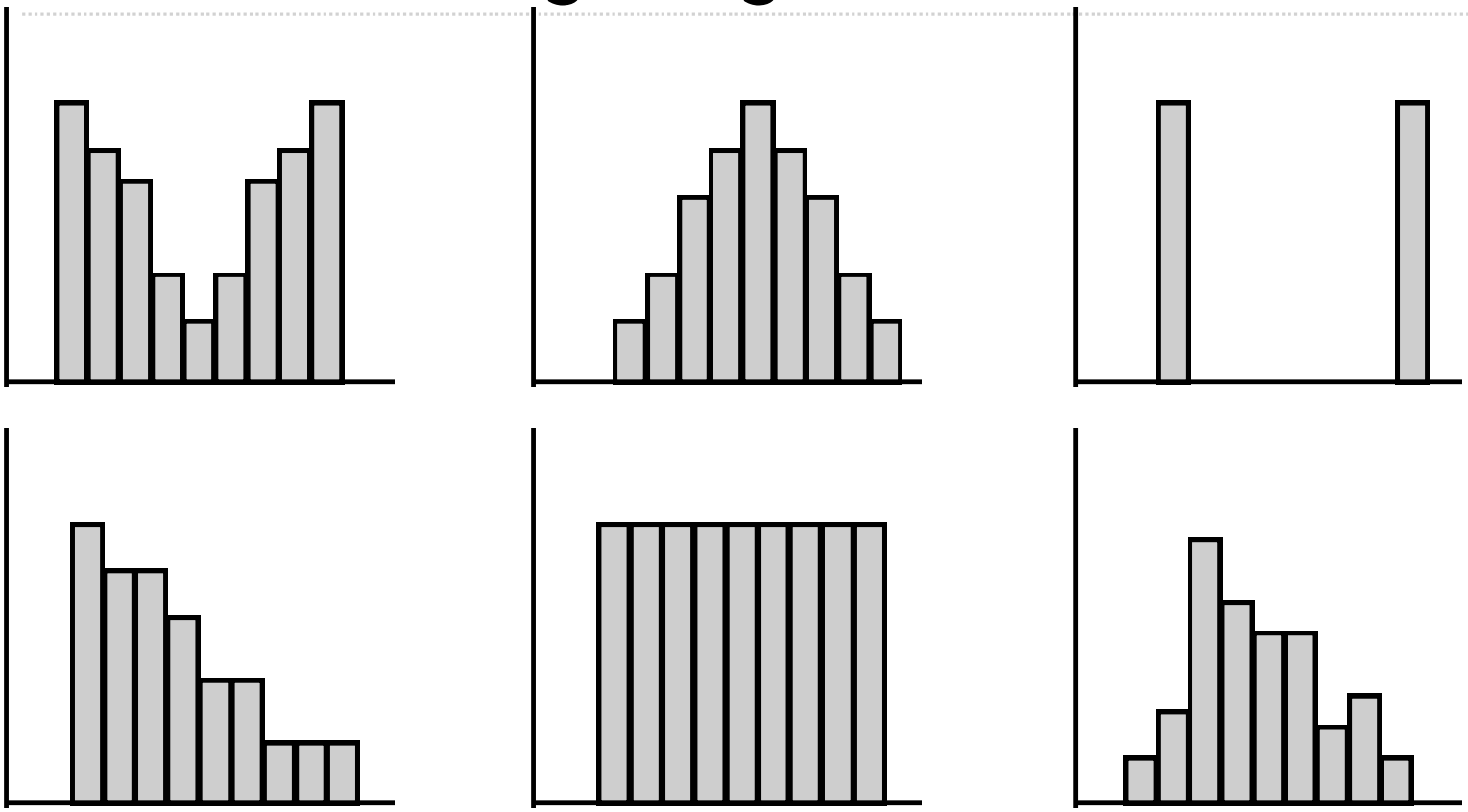
Streuungskennzahlen (-maßzahlen) charakterisieren das Ausmaß der Abweichungen vom Zentrum - die Variabilität - einer Häufigkeitsverteilung



Gegeben: Stichprobe vom Umfang n mit Einzelwerten

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

Verteilung mit gleichem Mittelwert



Streuungskennzahlen

Spannweite (Range):

Differenz zwischen größtem (x_{max}) und kleinstem (x_{min}) Wert einer Beobachtungsreihe

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

Eigenschaften:

- Ist der Streubereich einer Häufigkeitsverteilung (=Wertebereich in dem alle Merkmalswerte einer Beobachtungsreihe liegen)
- Für metrisch und ordinale skalierte Merkmale geeignet
- Wird sehr stark durch Ausreißer beeinflußt
- Spannweiten verschiedener Beobachtungsreihen nur dann vergleichbar, wenn Stichprobenumfang n gleich groß ist!!

Streuungskennzahlen

Stichprobenvarianz:

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

Mittlere quadratische Abweichung vom arithmetischen Mittel

Stichprobenstandardabweichung:

$$s = \sqrt{s^2}$$

Eigenschaften:

- Nur für metrisch skalierte Merkmale geeignet
- Stark durch extreme Werte (Ausreißer) beeinflußt
- Standardabweichung ist zur Interpretation von Ergebnissen besser geeignet als Varianz (s hat gleiche Maßeinheit wie Beobachtungen)
- Interpretation der Standardabweichung nur sinnvoll, wenn x bekannt ist (je größer $\bar{x}$, desto größer ist meist auch s)

Streuungskennzahlen

Interquartilsdistanz (Quartilsabstand):

Differenz zwischen oberem ($Q_{0.75}$) und unterem ($Q_{0.25}$) Quartil einer Beobachtungsreihe

$$IQR = Q_{0.75} - Q_{0.25}$$

Eigenschaften:

- Ist die Größe des Bereichs, in dem ca. 50% aller Werte einer Beobachtungsreihe liegen
- Für metrisch und ordinal skalierte Merkmale geeignet
- Kaum durch Ausreißer beeinflußt
- IQR zum Vergleich der Variabilität verschiedener Beobachtungsreihen besser geeignet als Spannweite

Streuungskennzahlen

Variationskoeffizient:

Verhältnis von Standardabweichung zu arithmetischem Mittel

$$v = \frac{s}{\bar{x}}$$

Eigenschaften:

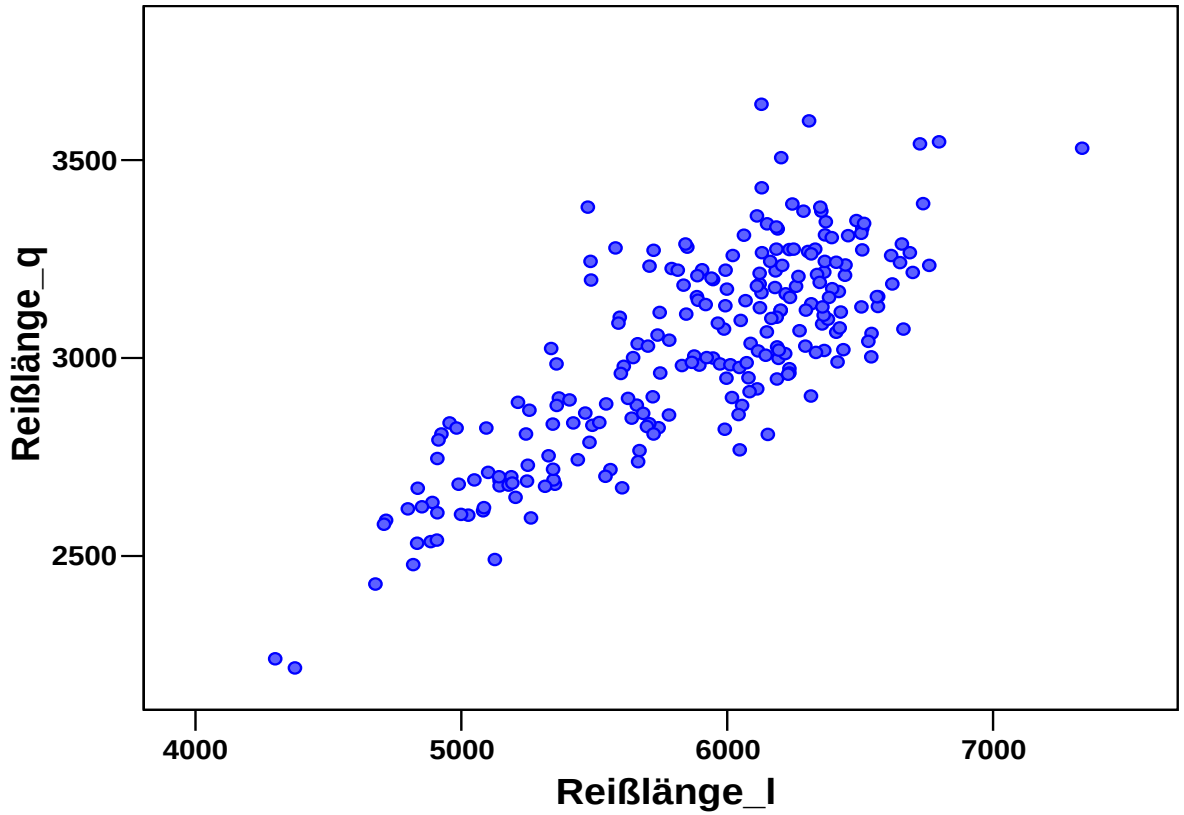
- Nur für metrisch skalierte Merkmale geeignet
- Von $\bar{x}$ bereinigtes Streuungsmaß, kann ohne Nennung von $\bar{x}$ interpretiert werden
- Stark durch Ausreißer beeinflußt
- Eignet sich zum Vergleich der Streuung von Merkmalen mit unterschiedlichen Wertebereichen
- Nur sinnvoll, wenn ausschließlich positive Merkmalswerte auftreten
- v oft in Prozent angegeben (v*100)

Analyse von Zusammenhängen

- univariat ➔ bivariat ➔ multivariat
- grafische Darstellungen
- Kennzahlen für Korrelation
- Kategorielle Daten ➔ Kontingenztafeln
  - Häufigkeiten
  - Zusammenhangsmaße

Scatterplot

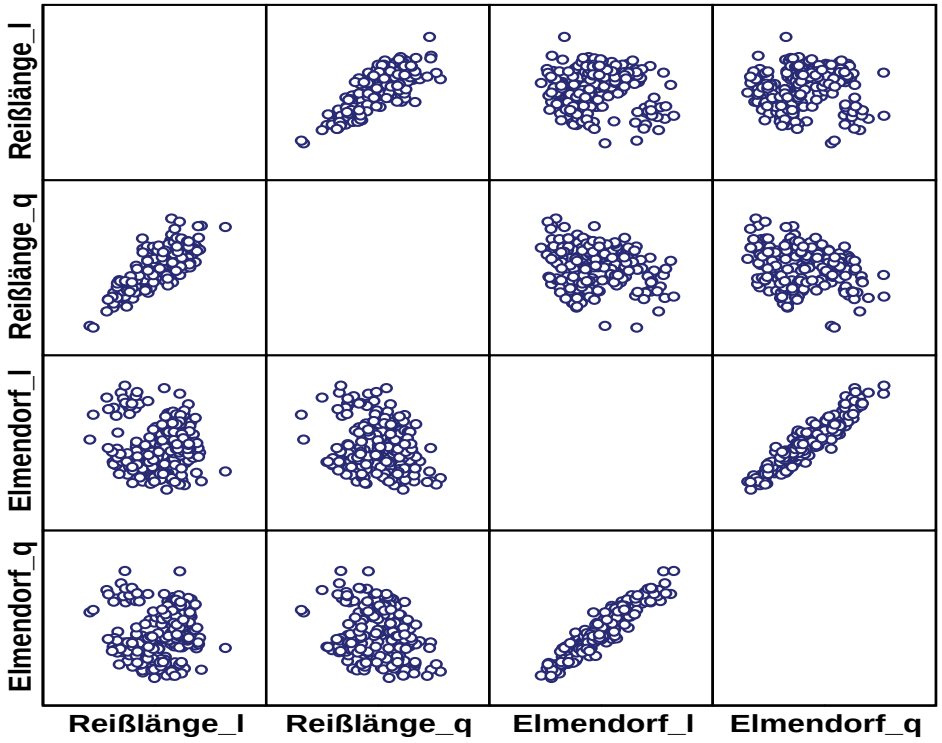
Beispiel: Vergleich der Reißlänge_l mit Reißlänge_q



Scatterplotmatrix (Draftsman-Plot)

Zweck:

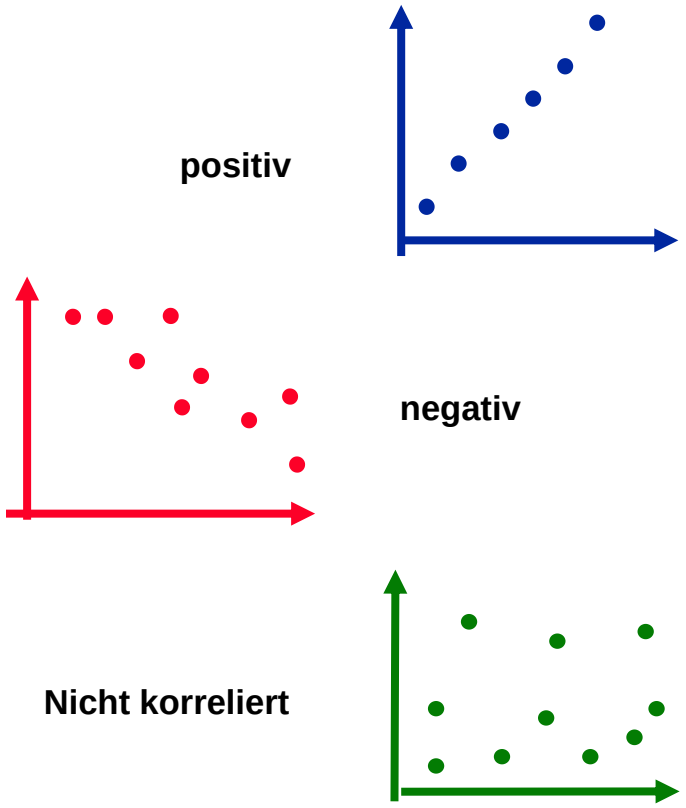
Grafische Darstellung des Zusammenhanges zwischen mehreren Merkmalen



Scatterplot für jede Zweierkombination von Merkmalen

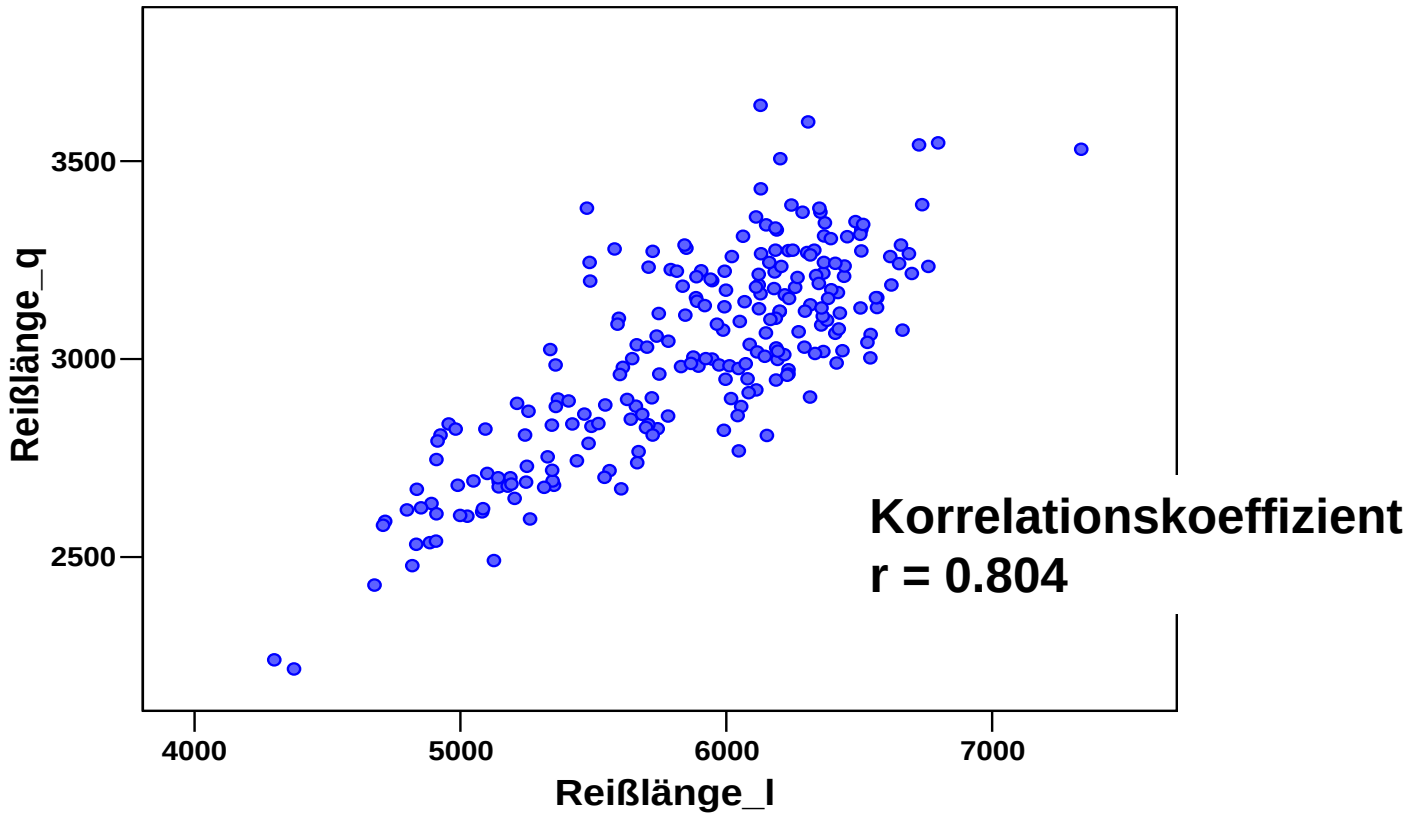
Korrelationskoeffizienten

- **Grad des (linearen) Zusammenhangs zwischen X und Y**
- **Metrisch skalierte Merkmale X und Y**
  - Pearson - Korrelationskoeffizient
- **Ordinal skalierte Merkmale X und Y**
  - Spearman - Rangkorrelationskoeffizient
- **Wertebereich: $-1 \leq r \leq 1$**
  - **Positiv korreliert**
    - $r > 0$
    - Direkt proportional
  - **Negativ korreliert**
    - $r < 0$
    - Indirekt proportional
  - **unkorreliert**
    - $r \approx 0$
    - kein linearer Zusammenhang



Scatterplot

Beispiel: Vergleich der Reißlänge_l mit Reißlänge_q



Lineare Korrelation

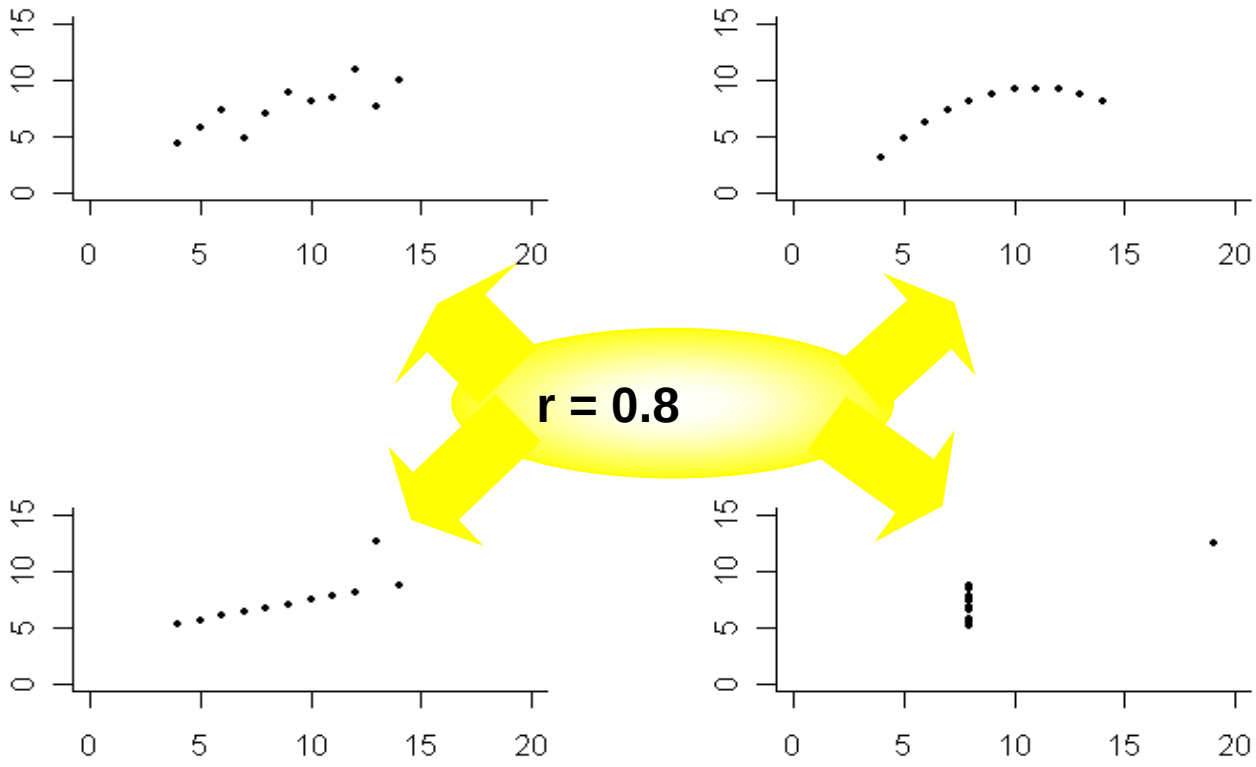
Je stärker der Zusammenhang zwischen *X* und *Y* ist, desto näher liegt r bei ± 1 und desto mehr nähert sich die Punktwolke der Beobachtungswerte (im Scatterplot) einer Geraden mit dem Anstieg ± 1.

! A B E R !

Auf keinen Fall isoliert den Korrelationskoeffizienten berechnen, sondern stets vor dessen Bestimmung den zugehörigen Scatterplot ansehen!!

Korrelationsanalyse - Scatterplot

Immer Scatterplot ansehen!
Korrelationskoeffizient allein kann täuschen



Vergleich zwischen deskriptiver Statistik und Inferenzstatistik

Der DIH SÜD wird unterstützt von:

Gliederung der Statistik

Beschreiben

Deskriptive Statistik

Schlüsse ziehen

Inferenzstatistik

Deskriptive Statistik

Deskriptive (beschreibende) Statistik:

- Instrumentarium zur Beschreibung von Daten
- Vorstufe zur schließenden Statistik

Ziel:

Beschreibung, Strukturierung, Verdeutlichung, Darstellung umfangreichen, unübersichtlichen Datenmaterials

Methoden:

- Graphische Darstellungen
- Kennzahlen (Maßzahlen)

Schließende Statistik

Schließende Datenanalyse:

- "Schätzen und Testen"
- Schließende/beurteilende Statistik

Ziele:

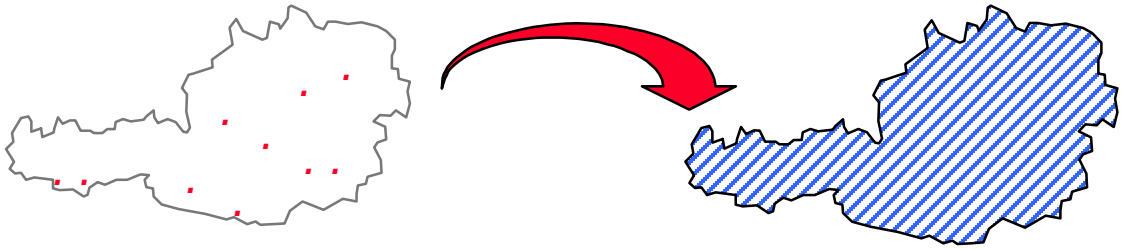
- Überprüfen der Gültigkeit von Hypothesen und Vermutungen
- Statistische Modellbildung

Methoden:

- Konfidenzintervalle
- Statistische Tests
- Korrelations- und Regressionsanalyse
- Varianzanalyse
- Zeitreihenanalyse
- Clusteranalyse u.v.a

Schließende Statistik

Schätzen: *Schluß von einer Stichprobe auf die Gesamtheit*



Testen: *Überprüfen einer Vermutung (Hypothese) über die Population mittels einer Stichprobe*

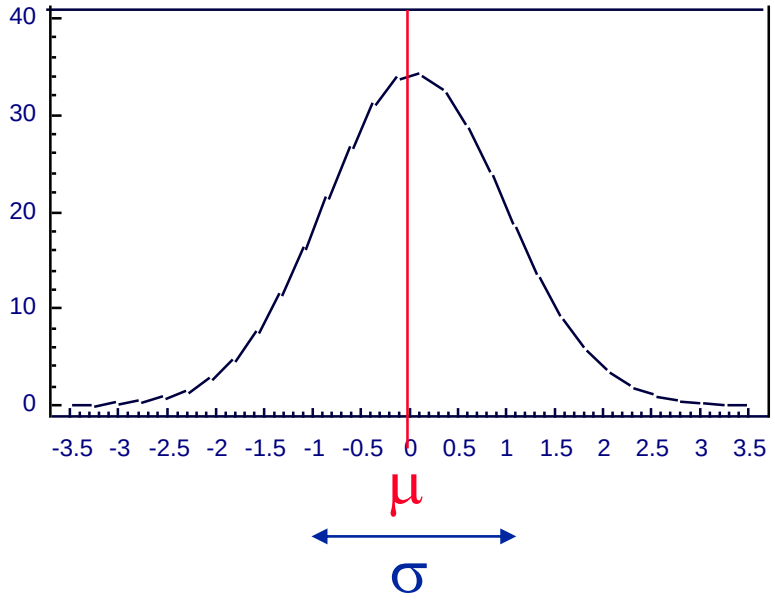
H_0:
Die Österreicher sind im Durchschnitt kleiner als 1,70m

Schätzen von Parametern

- Gegeben
  - Beobachtungen x_1, ..., x_n von Zufallsgrößen X_1, ..., X_n
- Gesucht
  - "möglichst gute" Schätzung eines unbekannten Parameters θ der ***Verteilungen*** der X_i
  - Funktion $\hat{\theta}(x_1,...,x_n)$, die der Zufallsstichprobe x_1, ..., x_n einen Wert zuordnet, der möglichst nahe oder gleich dem wahren Parameter θ ist, für den wir uns interessieren
- Beispiel
  - Anzahl der PCs je Firma
    - Schätzung der mittleren Anzahl je Firma
    - Schätzung der Summe der PCs
  - 10 Zeiten beim 100 m Lauf
    - Schätzung der mittleren Zeit

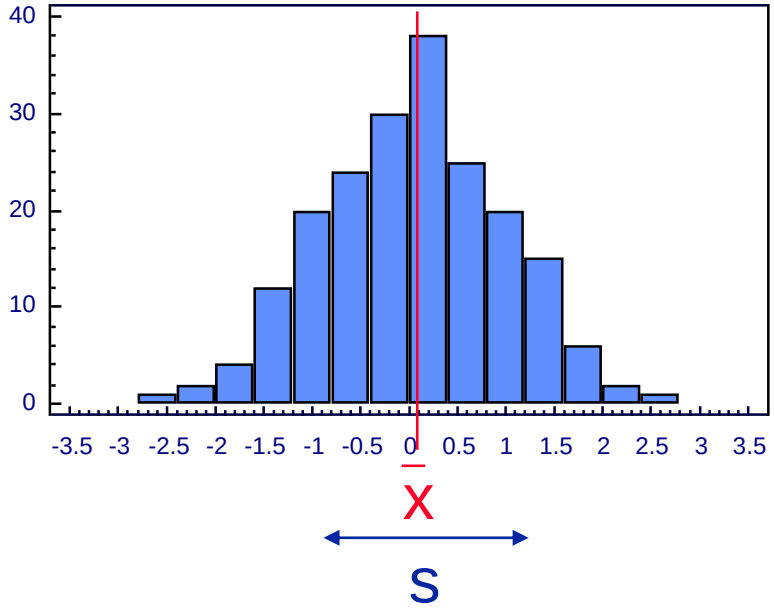
Normalverteilung

Theoretisch:
Population



μ, σ unbekannt

Praktisch:
Stichprobe

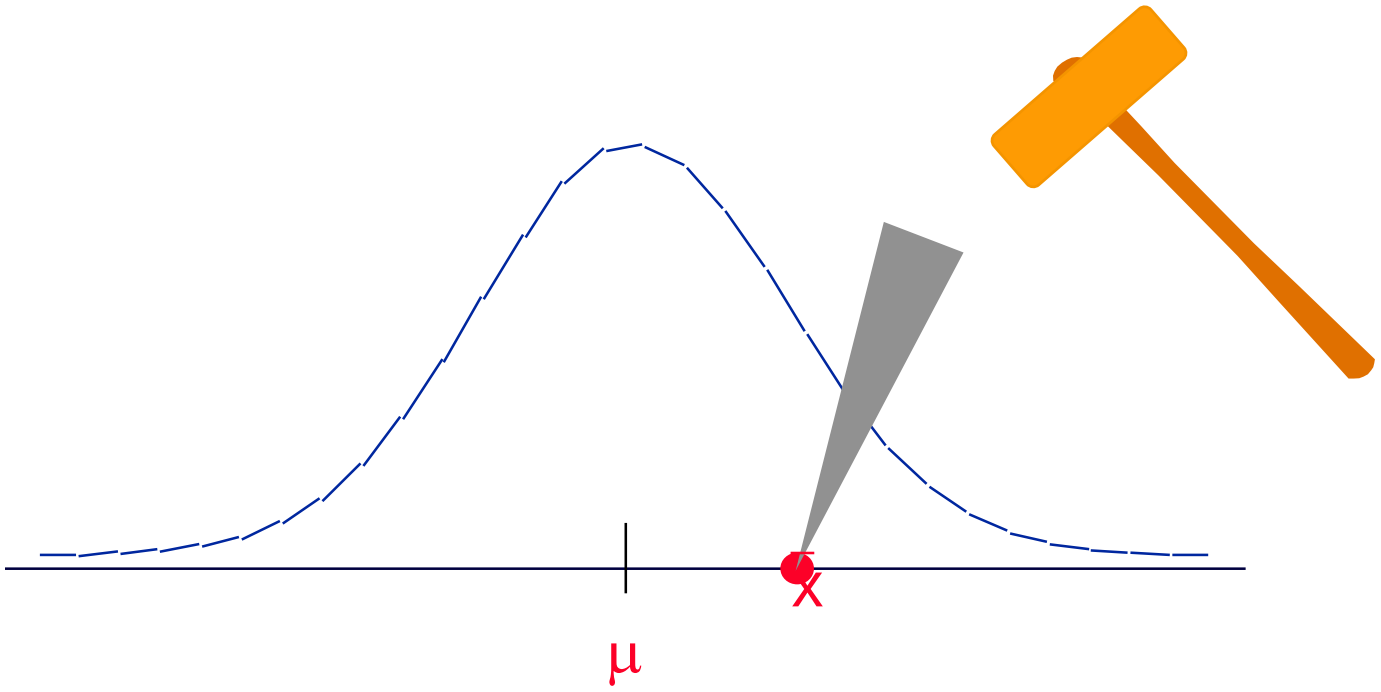


μ, σ durch $\bar{x}$, s schätzen

Schätzen: Punktschätzung

Transsylvanische Methode:

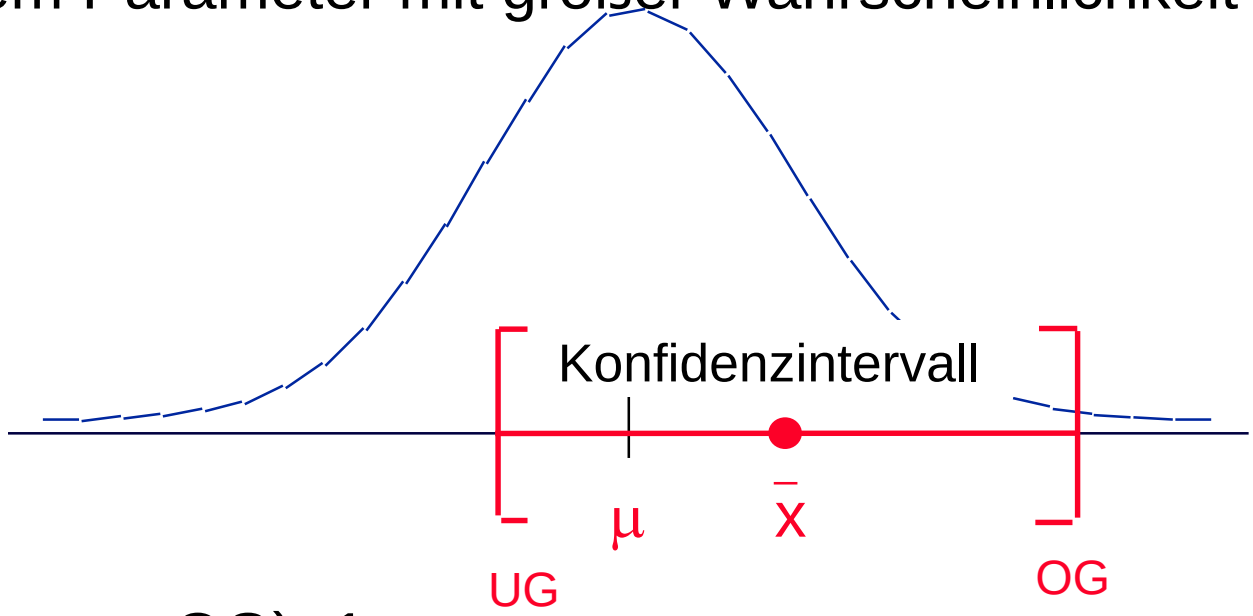
Schätzung von Parametern aus Stichprobe



Schätzen: Intervallschätzung

Konfidenzintervall:

Schätzung eines Bereiches aus Stichprobe,
in dem Parameter mit großer Wahrscheinlichkeit liegen



$P(UG \leq \mu \leq OG)=1-\alpha$

Konfidenzintervall

- [UG, OG] .... Konfidenzintervall
  - je kleiner das Intervall, desto größer ist die Genauigkeit der Schätzung
- Genauigkeit a
  - a=OG-UG
- Sicherheit
  - Wahrscheinlichkeit 1-α, dass der Parameter im Intervall liegt
  - P (UG$\leq \mu \leq$ OG)=1-α
  - α ... Irrtumswahrscheinlichkeit

Beispiel

- Konfidenzintervall (KI) für den Erwartungswert bei normalverteilter Grundgesamtheit

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Wann wird das KI klein?

ANWENDUNGSBEISPIEL

DIHSÜD

Der DIH SÜD wird unterstützt von:

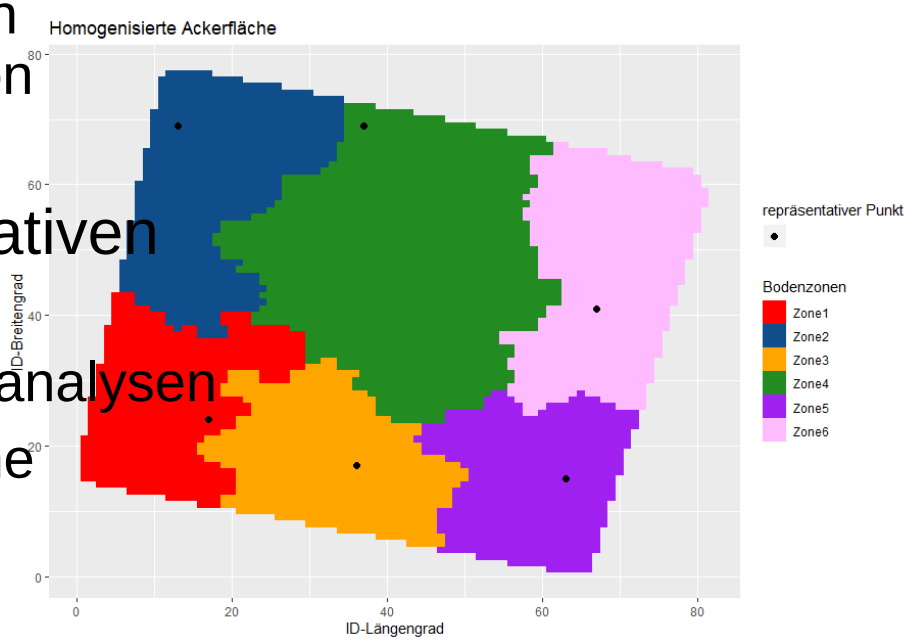
FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

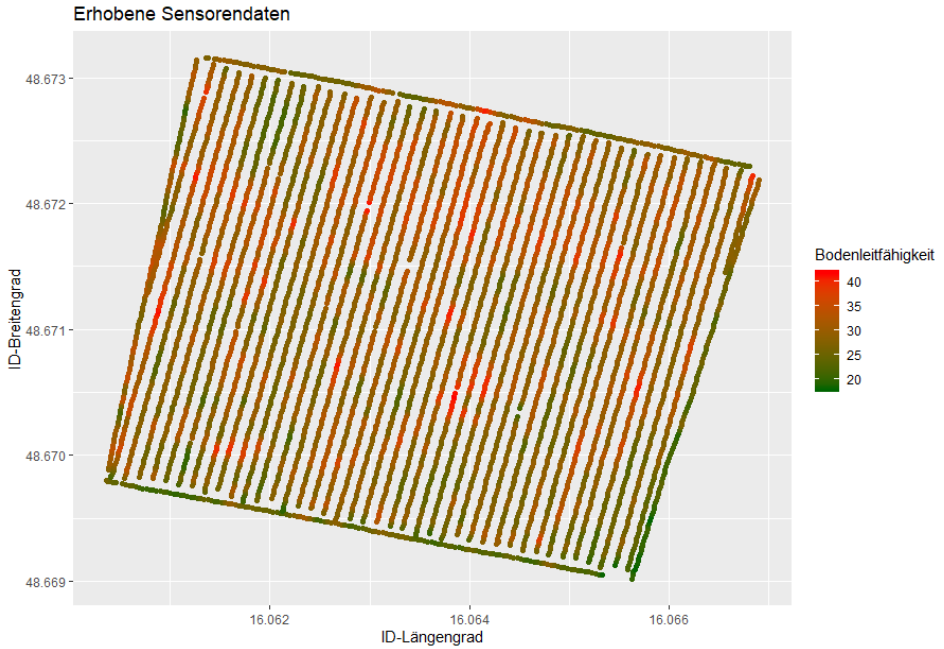
Zielsetzung

- Ressourcenoptimale Bewirtschaftung
  - Identifikation von homogenen Feldabschnitten auf Basis von Bodensensordaten
- Festlegung eines repräsentativen Punktes je Zone
  - Basis für ergänzende Bodenanalysen
  - Einflussgrößen für statistische Modellierung



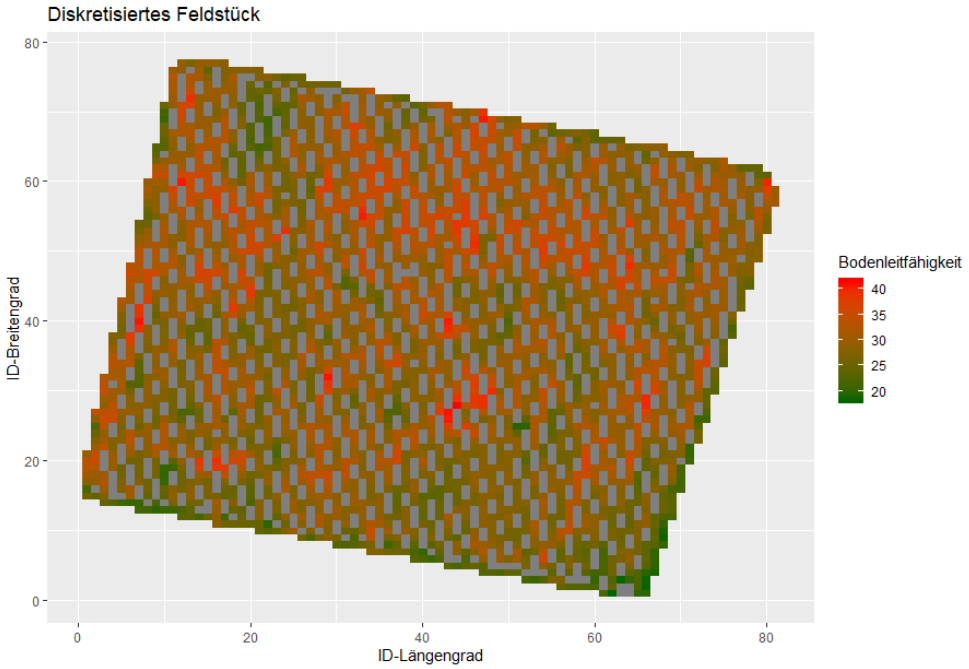
Ausgangssituation

- Erhebung verschiedenster Bodenparameter
  - Bodenleitfähigkeit
  - Rötungswerte des Bodens
  - Säuregehalt des Bodens



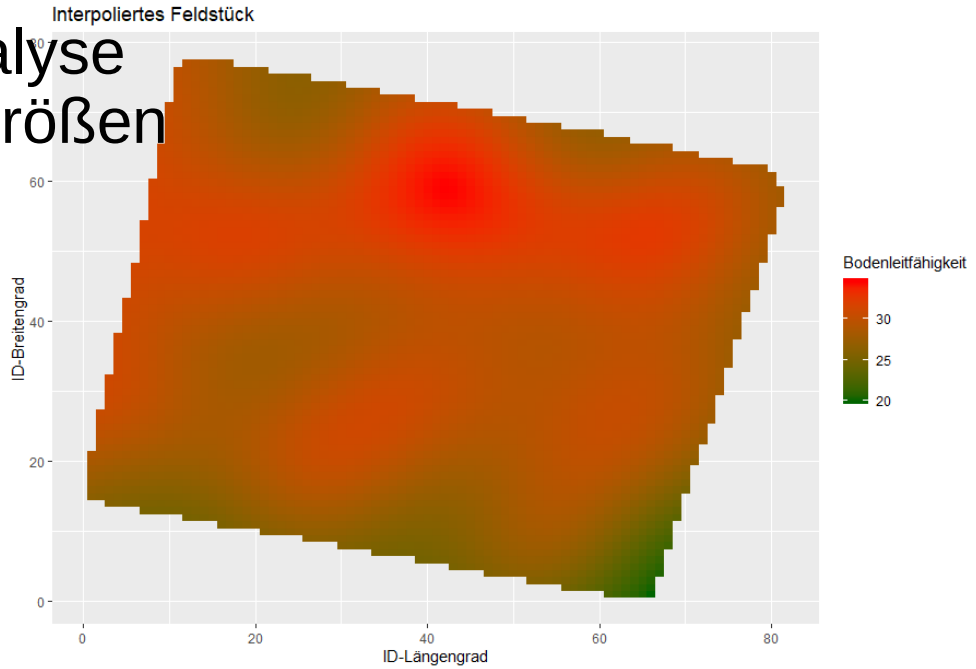
Vorgangsweise I

- Plausibilitätscheck der Sensordaten
- Diskretisierung der Dater in Zellen
  - 6×6 Meter Zellen
- Aggregation der Zellwert



Vorgangsweise II

- Räumliche Interpolation und Glättung der Zellwerte mittels Generalisiertem Additiven Modell
- Durchführung einer hierarchischen Clusteranalyse auf Basis von m Einflussgrößen
  - Anzahl der Cluster n muss angegeben werden



Identifikation des Repräsentativen Punktes

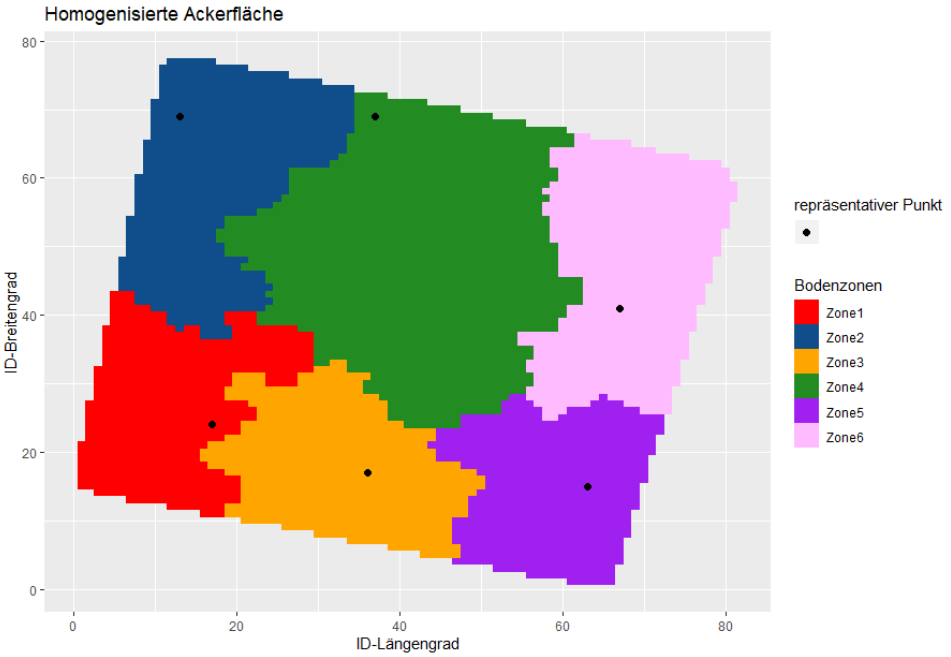
- Bestimmung des Overall-Mean jeder Einflussgröße pro Cluster $(\overline{\overline{x_1}}, \dots, \overline{\overline{x_m}})$
- Berechnung der euklidischen Distanz zwischen Overall-Mean und geglätteten Zellwerten der jeweiligen Zone $(\overline{x_{1_l}}, \dots, \overline{x_{m_l}})$

$$d_l = \sqrt{\sum_{i=1}^{m} (\overline{\overline{x_i}} - \overline{x_{i_l}})^2} \text{ für } l = 1, \dots,$$

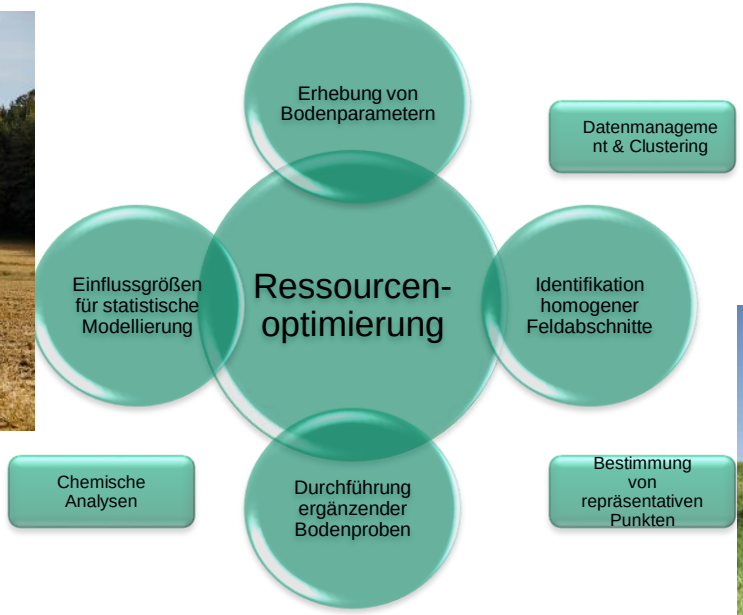
- Wähle jene Zelle mit minimalstem d_l je
  → repräsentative Punkt

Resultate

- Ergänzende Bodenproben an repräsentativen Punkten
  - Daten sind charakteristisc gesamte Zone
- Zoneninformation und Inf der ergänzenden Bodenp sind Basis für Modellierur verschiedenen Zielgröße
  - Düngung
  - Bewässerung
  - Ertrag



Zusammenfassung



Zusammenfassung und Resümee

Der DIH SÜD wird unterstützt von:

FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Zusammenfassung

- Große Datenmengen sind von sich aus kein Erfolgsgarant
- Systematische Vorgangsweise bei datengestützten Fragestellungen ist von enormer Wichtigkeit
- Daten ≠ Informationen
- Datenanalyse führt meist zur Datenreduktion
  - Einsatz von statistischen Methoden zwingend notwendig
  - Führt meist von Big Data zu Smart Data
  - Basis für Verbesserungen
- Datenanalyse kann in folgenden Bereichen eingesetzt werden
  - Identifikation von Zusammenhängen – statistische Modellierung
  - Monitoring
  - Prognose
- „Mache die Dinge so einfach wie möglich – aber nicht einfacher“. (Albert Einstein)

Resümee und Ausblick

- Datenanalyse sollte integraler Bestandteil im unternehmerischen Umfeld sein
- Sammeln und auswerten der „richtigen“ Daten wird zum Erfolgsfaktor
- Wissen aus den Daten wird Differenzierungsmerkmal am Markt

Angewandte Statistik ist eine Zusammenfassung von Methoden, die uns erlauben, vernünftige Entscheidungen im Falle von Ungewissheit zu treffen.

(Abraham Wald)

Danke für Ihre Aufmerksamkeit!

DI Hermann Katz

JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH

POLICIES – Institut für Wirtschafts- und Innovationsforschung
Datenanalyse und statistische Modellierung

Leonhardstraße 59
8010 Graz

Tel.: +43 316 876-1553
Mobil: +43 664 602 876 1553
PC-Fax: +43 316 8769-1553

Email: hermann.katz@joanneum.at

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN