

# Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

*Marcel Dettling*

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

[marcel.dettling@zhaw.ch](mailto:marcel.dettling@zhaw.ch)

<http://stat.ethz.ch/~dettling>

ETH Zürich, 28. April 2010

# Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

## ***Einführung in die Statistik***

### **Was ist Wahrscheinlichkeitsrechnung?**

Man geht von einem Modell aus (welches die “Wahrheit”, d.h. die zugrunde liegenden Phänomene/Mechanismen beschreibt) und macht Aussagen über mögliche Ausgänge von Experimenten, d.h. das Aussehen von Daten.



## Einführung in die Statistik

### Beispiele zur Wahrscheinlichkeitsrechnung?

- 1) W'keit, in 3 Würfelwürfen 2 Sechser zu werfen
- 2) W'keit, dass 2 Personen im Raum gleichentags Geburi haben
- 3) W'keit, dass eine Frau Fusslänge < 25cm hat



## Einführung in die Statistik

### Was ist Statistik?

Man geht von bereits vorliegenden Daten aus und versucht damit, Rückschlüsse auf das erzeugende Modell (die "Wahrheit", d.h. die zugrunde liegenden Phänomene/Mechanismen) zu ziehen.



### ***Statistik: Wozu?***

- Statistische Datenanalyse ist dazu da, um unter der Präsenz von Unsicherheit und Variation dennoch intelligente und informierte Entscheidungen zu treffen.
- Wenn alles exakt vorbestimmt (d.h. deterministisch) ist, so braucht es die Methoden der Statistik nicht.
- Reale Daten weisen aber nahezu immer Unsicherheit und Variation auf.
- die Folgerung daraus können Sie selbst ziehen...

### ***Woher kommen die Daten?***

- Geplante Experimente
  - z.B. Wirksamkeit von Medikamenten
- Beobachtungsstudien
  - z.B. Frischwassermenge
- Umfragen
  - z.B. wieviele % würden einen CVP-Bundesrat wählen?
- Data Mining
  - mit Daten, welche für andere Zwecke erhoben wurden

### ***Aufgaben der Statistik***

- Versuchsplanung
  - welche Daten sollen erhoben werden?
- Deskriptive Statistik
  - wie werden Daten zusammengefasst und visualisiert?
- Schliessende Statistik
  - welche Folgerungen lassen sich aus den Daten ziehen?
- Explorative Statistik / Data Mining
  - lassen sich unerwartete/unbekannte Strukturen finden?

### ***Schliessende Statistik: Beispiel 1***

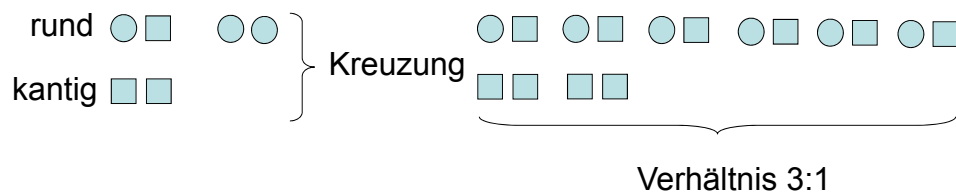
Wirksamkeit von 2 Schlafmitteln A und B. Bei 10 Probanden wurde die durchschnittliche Schlafverlängerung von A vs. B gemessen.

1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4

- andere Probanden, andere Messwerte?!
- Dauer der Schlafverlängerung, plausible Werte?
- glauben wir daran, bzw. ab wann glauben wir daran, dass A besser als B ist?

### Schliessende Statistik: Beispiel 2

Mendels Vererbungsgesetze: 2 Erbsensorten mit runden/kantigen Samen. Dabei werden runde Samen dominant vererbt.



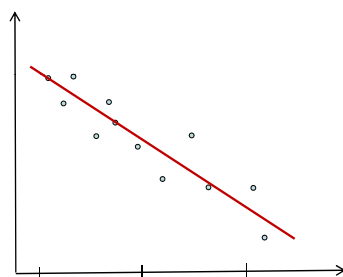
Beobachtung: 5474 runde, 1850 kantige Samen, d.h. 2.96:1

- folgen die Pflanzen dem Vererbungsmodell?
- wieviele Samen müssen gezählt werden, um sicher zu sein?

9

### Schliessende Statistik: Beispiel 3

Korrosion einer Kupfer-Nickel-Legierung in Abhängigkeit vom Eisengehalt. 13 Räder wurden hergestellt und während 60 Tagen gedreht. Gemessen wurde der Gewichtsverlust in mg.



**Beobachtung:** Weniger Korrosion bei zunehmenden Eisengehalt

- ist dieser Effekt gesichert?
- wie gross ist die Abnahme, wenn der Eisengehalt um 1 Einheit erhöht wird?
- wie genau kann man den Gewichtsverlust vorhersagen?

10

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Grundgesamtheit**

Die **Grundgesamtheit** ist die Menge aller für eine Untersuchung relevanten Einheiten/Personen, z.B.

- a) alle Chem/Bio-Studenten des 2. Semesters
- b) alle in der Schweiz wohnhaften Personen
- c) alle Gelatinkapseln eines Produktionsloses
- d) sämtliche Tiere/Pflanzen einer bestimmten Gattung

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Vollerhebung vs. Stichprobe**

Bei einer **Vollerhebung** werden bei jedem Individuum in der *Grundgesamtheit* die Eigenschaften erhoben/abgefragt/gemessen.

Sobald die *Grundgesamtheit* eine gewisse Grösse hat, ist eine *Vollerhebung* meist weder nötig, noch mit vernünftigem Aufwand durchzuführen. Man beschränkt sich deshalb meist auf eine *Stichprobe*.

Eine **Stichprobe** ist eine Teilmenge der *Grundgesamtheit*, die unter bestimmten Gesichtspunkten ausgewählt wurde. Dies kann zufällig, systematisch, willkürlich, etc. sein.

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Repräsentativität**

Wir nennen eine *Stichprobe* **repräsentativ**, wenn sie eine Auswahl aus der *Grundgesamtheit* ist, die alle deren typischen Merkmale getreu deren relativer Häufigkeit aufweist. Das heisst, sie ist ein Abbild der *Grundgesamtheit*.

### **Variable**

Eine **Variable** ist eine Eigenschaft, deren Wert vom einen zum anderen Objekt in der *Grundgesamtheit* bzw. *Stichprobe* wechselt.

- Geschlecht eines/einer Bio/Chem-Studenten/Studentin
- Anzahl defekte Komponenten in einem Auto
- Masse einer Pflanze unter spezifischen Konditionen

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Beispiele zur Repräsentativität**

- Fragestellung:** Wie lange sind sie täglich im Internet?  
**Stichprobe:** freiwillige Umfrage auf 20minuten.ch  
 gut                       mässig                       schlecht
- Fragestellung:** Ist der Benzinverbrauch mit Sommer- oder mit Winterreifen höher?  
**Stichprobe:** 20 Autofahrer, die über je 4 Wochen Benzinverbrauch und Fahrleistung im Sommer und im Winter rapportieren.  
 gut                       mässig                       schlecht

### ***Beispiele zur Repräsentativität***

- 3) **Fragestellung:** Sollen in der Stadt Zürich mehr Parkplätze angeboten werden?
- Stichprobe:** Befragung von 150 Passanten in Zürich HB
- gut                       mässig                       schlecht
- 4) **Fragestellung:** Wie stark überbucht die Swiss ihre Flüge?
- Stichprobe:** Auszählung von Tickets/Sitzen in 40 zufällig ausgewählten Flügen.
- gut                       mässig                       schlecht

### ***Beispiele zur Repräsentativität***

- 5) **Fragestellung:** Welcher Teil der Bio/Chem-Studenten ist ein Jahr nach Abschluss arbeitslos?
- Stichprobe:** Vollerhebung in einem Jahrgang
- gut                       mässig                       schlecht
- 6) **Fragestellung:** Wie hoch ist der durchschnittliche Hämoglobin-Wert bei Frauen?
- Stichprobe:** Frauen, die zur Blutspende-Aktion an der Uni Zürich kommen.
- gut                       mässig                       schlecht

### Beispiele zur Repräsentativität

7) **Fragestellung:** Wie viel verdient eine in der Schweiz wohnhafte Person?

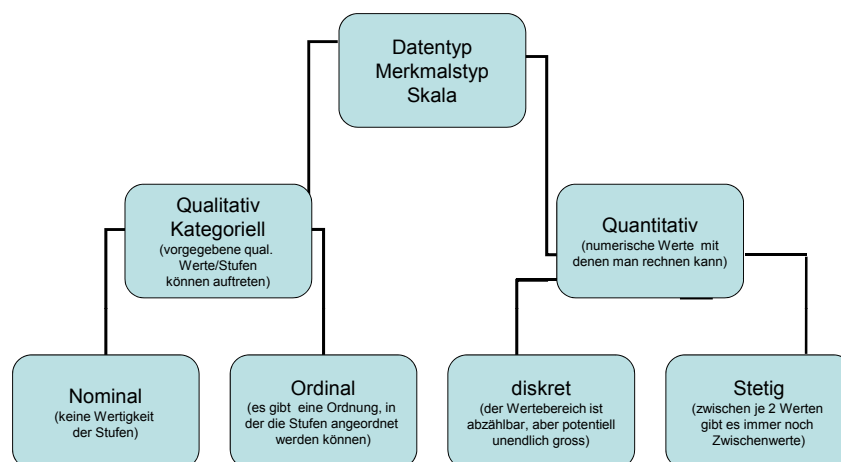
**Stichprobe:** Telefonbefragung aufgrund zufällig ausgewählter Nummern tagsüber

- gut                       mässig                       schlecht

8) **Fragestellung:** Wie viele A-Post-Briefe kommen pünktlich an?

**Stichprobe:** nach Bevölkerungsanzahl gewichtete Postämter aus der ganzen CH

- gut                       mässig                       schlecht



# Grundlagen der Mathematik II

## Lineare Algebra und Statistik

### FS 2010 – Woche 09

### Beispiel zu Variablentypen

id	alter	zivilstand	geschlecht	piht	kaufkraft	einkauf
1	50	Single	Mann	1	hoch	10951
2	86	Verheirater	Mann	2	sehr hoch	23091
3	NA	Single	Mann	1	hoch	NA
4	65	Single	Mann	1	sehr hoch	18866
5	NA	Single	Frau	1	hoch	NA
6	58	Single	Mann	1	mittel	2243
7	58	Konkubinat	Mann	3	hoch	7669
8	NA	Konkubinat	Mann	3	sehr hoch	NA
9	54	Verheirater	Mann	2	sehr hoch	23643
10	43	Verheirater	Mann	5	mittel	5993
11	NA	Verheirater	Mann	2	sehr hoch	NA
12	84	Konkubinat	Mann	2	sehr hoch	22379
13	52	Konkubinat	Mann	4	sehr hoch	12499
14	53	Verheirater	Mann	2	mittel	7439
15	65	Konkubinat	Mann	6	sehr hoch	11612
16	73	Verheirater	Unbekannt	1	sehr hoch	22329
17	35	Single	Unbekannt	1	mittel	5062
18	49	Verheirater	Mann	2	sehr hoch	22730
19	NA	Verheirater	Mann	2	tief	NA
20	45	Verheirater	Mann	3	hoch	7631
21	63	Konkubinat	Mann	4	hoch	9581
22	40	Verheirater	Mann	2	hoch	12114
23	NA	Single	Mann	1	mittel	NA
24	42	Single	Frau	1	tief	1020
25	42	Konkubinat	Mann	1	sehr hoch	17121
26	51	Verheirater	Mann	2	sehr hoch	21171
27	NA	Konkubinat	Mann	3	sehr hoch	NA
28	41	Verheirater	Mann	4	mittel	5197
29	45	Konkubinat	Frau	2	sehr hoch	9958

Marcel Detting, Zurich University of Applied Sciences

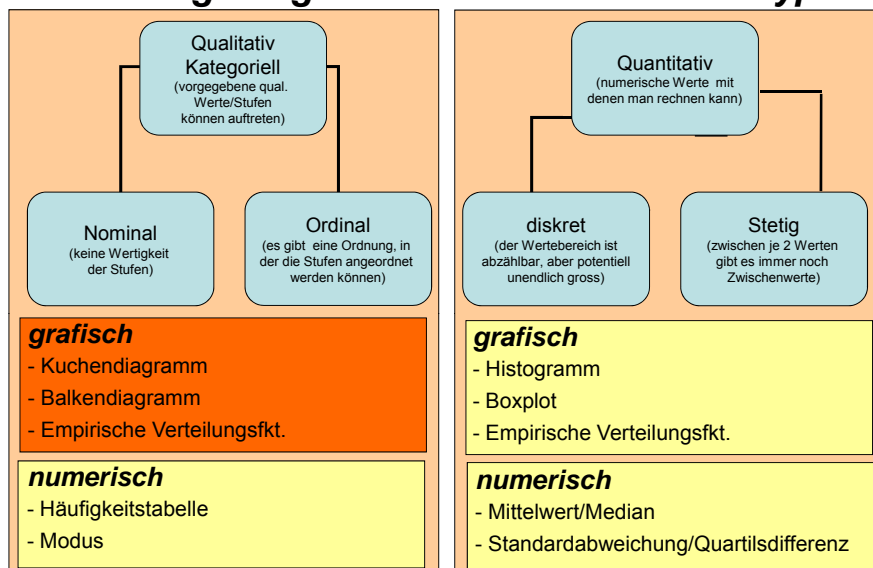
19

# Grundlagen der Mathematik II

## Lineare Algebra und Statistik

### FS 2010 – Woche 09

### Darstellungsmöglichkeiten nach Variablentyp



20

**Grundlagen der Mathematik II**  
**Lineare Algebra und Statistik**  
**FS 2010 – Woche 09**

**Häufigkeitstabelle für kategorielle Variablen**

Kaufkraft	Anzahl	Prozent
Unbekannt	967	3.96%
sehr hoch	5179	21.19%
hoch	7827	32.02%
mittel	7191	29.41%
tief	3282	13.42%
	24446	100.00%

**Eigenschaften:**

- präzise
- langweilig, wenig übersichtlich

**Grundlagen der Mathematik II**  
**Lineare Algebra und Statistik**  
**FS 2010 – Woche 09**

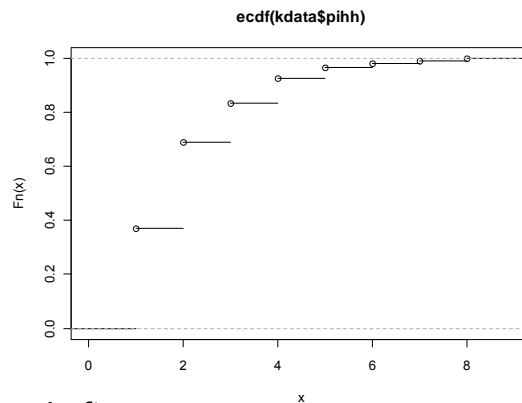
**Häufigkeitstabelle für ordinale Variablen**

PIHH	Häufigkeit	relative H.	kumulierte H.	rel. kum. H.
1	9065	37.08%	9065	37.08%
2	7789	31.86%	16854	68.94%
3	3564	14.58%	20418	83.52%
4	2253	9.22%	22671	92.74%
5	987	4.04%	23658	96.78%
6+	788	3.22%	24446	100.00%
	24446	100.00%		

Für ordinale Variablen macht es, im Gegensatz zu nominalen Variablen, auch Sinn, die kumulierten, absoluten und relativen Häufigkeiten anzugeben.

Grundlagen der Mathematik II  
Lineare Algebra und Statistik  
FS 2010 – Woche 09

**Empirische Verteilungsfunktion**



Eigenschaften:

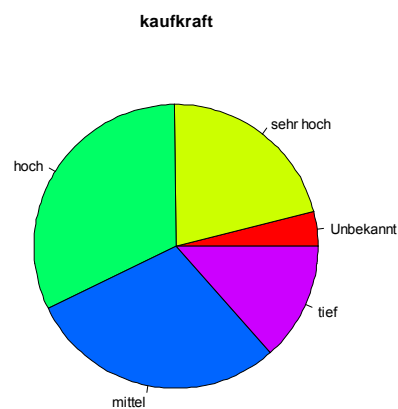
=> stellt die kumulativen, relativen Häufigkeiten dar

=> nur für ordinale (oder quantitative) Variablen sinnvoll

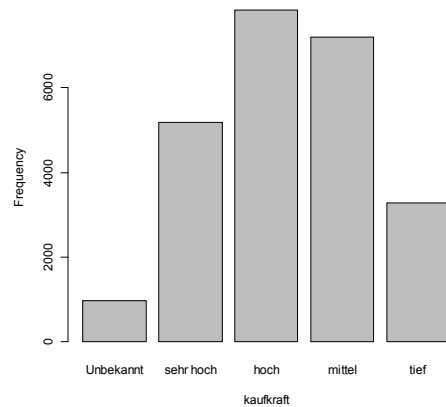
23

Grundlagen der Mathematik II  
Lineare Algebra und Statistik  
FS 2010 – Woche 09

**Kuchendiagramm für kategorielle Variablen**



**Balkendiagramm für kategorielle Variablen**



**Balken oder Kuchen???**

Kuchendiagramme sind weit verbreitet, bieten aber gegenüber Balkendiagrammen fast nur Nachteile. Der Grund:

**das menschliche Auge kann Flächen nicht gut miteinander vergleichen**

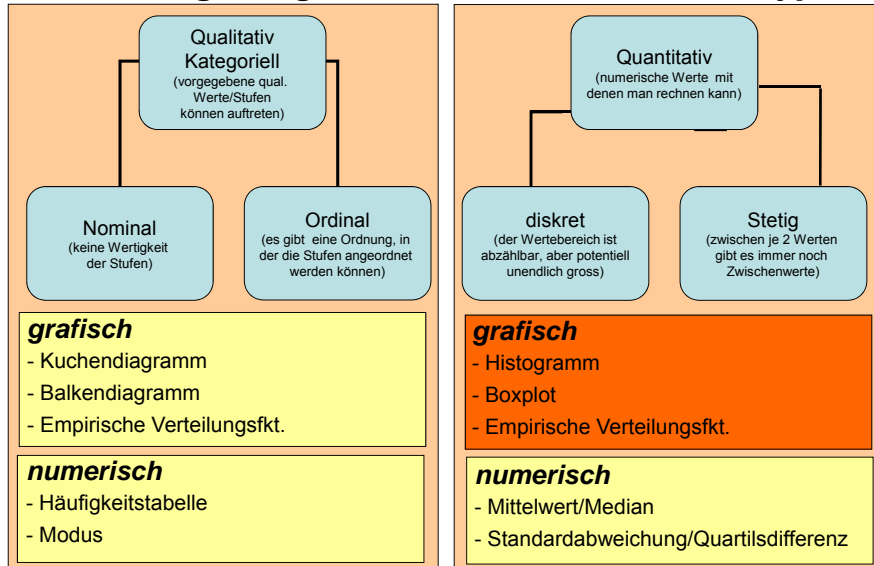
=> **es ist mit Kuchendiagrammen schwierig(er), die einzelnen Merkmale nach deren Häufigkeit zu ordnen**

=> **es sollten ausschliesslich Balkendiagramme eingesetzt werden**

## Grundlagen der Mathematik II Lineare Algebra und Statistik

FS 2010 – Woche 09

### Darstellungsmöglichkeiten nach Variablentyp



27

## Grundlagen der Mathematik II Lineare Algebra und Statistik

FS 2010 – Woche 09

### Histogramm

Departement W



5' 5'4" 5'8" 6'

Departement L



5' 5'4" 5'8" 6'

Quasi-Histogramme (es handelt sich eigentlich um „Dotplots“)

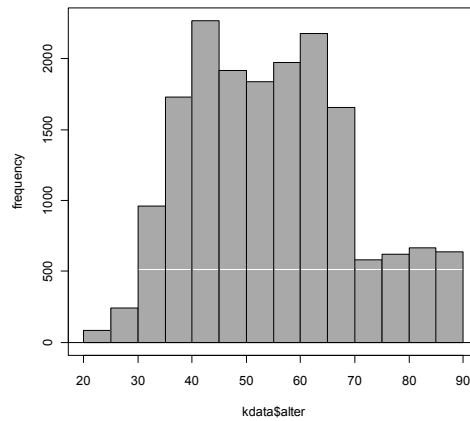
- tendieren kleiner gewachsene Personen zum Studium im Dep. L?
- welche „confounding factors“ könnte es geben, damit dieses Bild entsteht?

# Grundlagen der Mathematik II

## Lineare Algebra und Statistik

### FS 2010 – Woche 09

### Histogramm



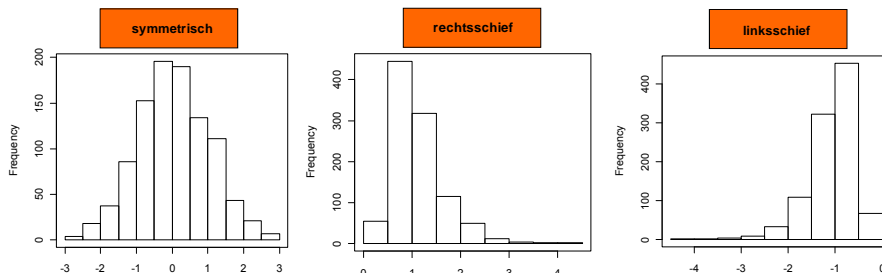
**Histogramme stellen die Verteilung eines Merkmals sehr gut dar**

# Grundlagen der Mathematik II

## Lineare Algebra und Statistik

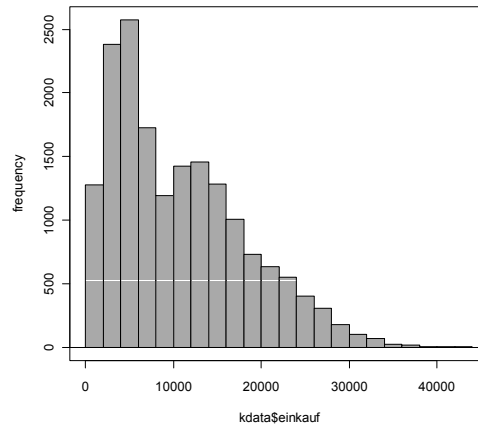
### FS 2010 – Woche 09

### Histogramm – Begriffe



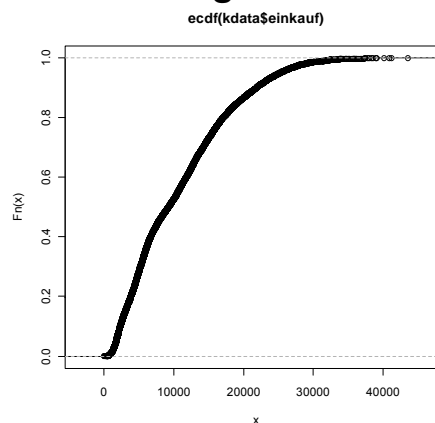
**=> Diese Histogramme sind alle unimodal!**

### Histogramm – Begriffe



=> Diese Histogramm ist (schwach ausgeprägt) bimodal!

### Empirische Verteilungsfunktion



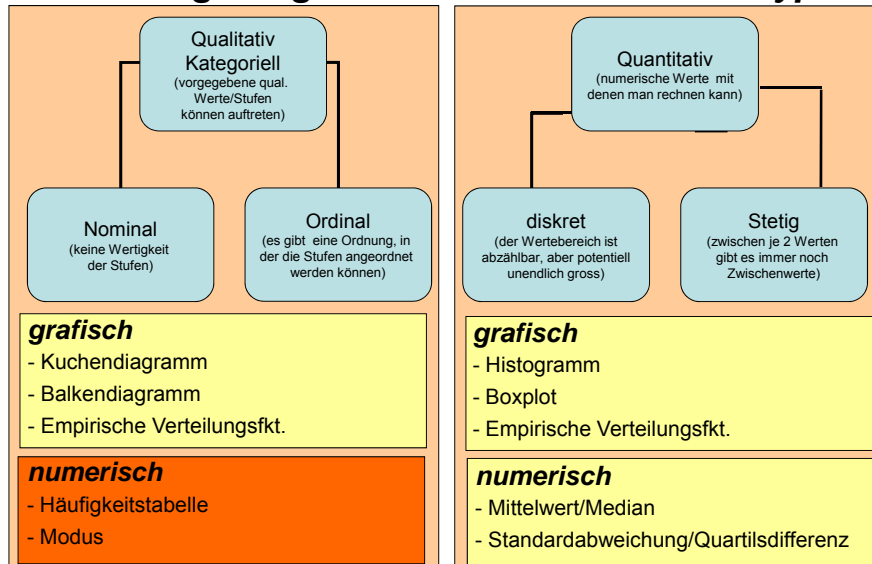
Eigenschaften:

=> stellt die kumulativen, relativen Häufigkeiten  $\leq x$  dar

=> fein unterteilte Treppenfunktion, im Grenzübergang stetig

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### Darstellungsmöglichkeiten nach Variablentyp



33

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

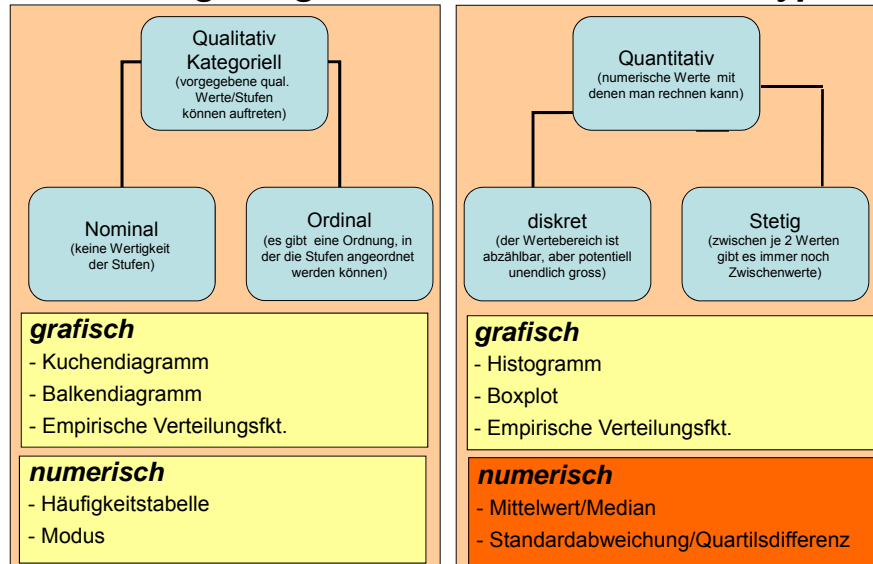
### Modus

Der Modus ist derjenige Wert, der am häufigsten vorkommt. Dieses Mass ist demnach nur bei qualitativen, oder allenfalls diskreten Daten sinnvoll.

Kaufkraft	Anzahl	Prozent
Unbekannt	967	3.96%
sehr hoch	5179	21.19%
hoch	7827	32.02%
mittel	7191	29.41%
tief	3282	13.42%
	24446	100.00%

=> Der Modus der Variable Kaufkraft ist die Ausprägung „hoch“

### Darstellungsmöglichkeiten nach Variablentyp



35

### Kennzahlen, allgemeines

Daten, d.h. ein Zahlenhaufen, können nicht nur grafisch, sondern auch durch Kennzahlen (Masszahlen) charakterisiert werden. Wichtig sind vor allem jene Grössen, die **Lage** oder **Streuung** der Daten beschreiben.

#### Lagemasse

Eine Zahl, die den „mittleren“ Wert charakterisiert

#### Streuungsmasse

Eine Zahl, die angibt, wie stark die Werte um den „mittleren“ Wert streuen

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Mittelwert**

Dieses Mass ist allgemein bekannt und einfach zu berechnen. Man bildet die Summe aller Werte und teilt durch die Anzahl Werte.

### **Median**

Der Median ist so definiert, dass 50% aller Werte kleiner und 50% aller Werte grösser sind. Wenn eine ungerade Anzahl von Daten vorliegen, ist der Median gleich dem mittleren Wert der nach der Grösse geordneten Daten. Bei einer geraden Anzahl von Daten ist der Median das Mittel der zwei mittleren Werte der geordneten Daten

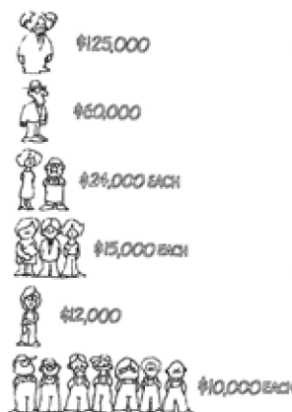
## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### **Gedanken zu den Lagemassen**

#### Mittelwert

$$\begin{aligned}
 &125,000 \\
 &+ 60,000 \\
 &+ 2 \times 24,000 \\
 &+ 3 \times 15,000 \\
 &+ 12,000 \\
 &+ 7 \times 10,000 \\
 \hline
 &= 360,000
 \end{aligned}$$

$$\frac{\$360,000}{15} = \boxed{\$24,000}$$



#### Median

Ordne die 15 Angestellte nach deren Lohn  
~> der Median des Lohns ist dann derjenige der 8. Person

Es ist dann so, dass die Hälfte aller Angestellten einen höheren Lohn als den Median bekommt, und die andere Hälfte der Angestellten einen niedrigeren.

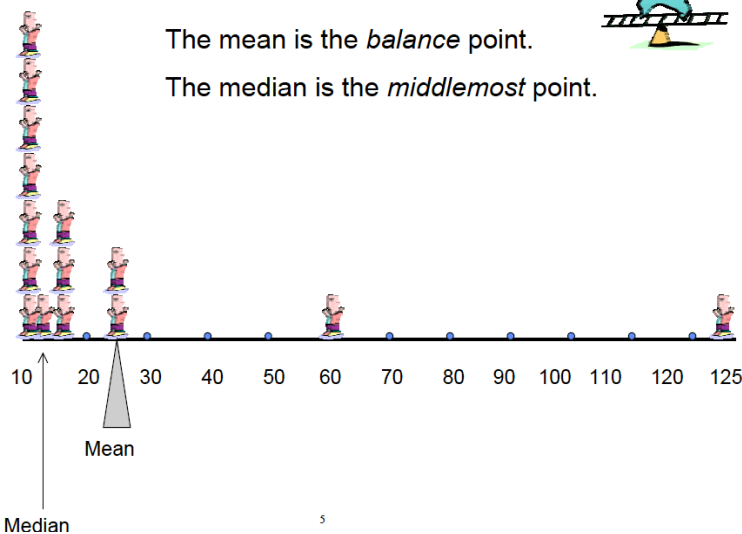
$$\Rightarrow \text{Median} = \$12,000$$

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09

### ***Wann Mittelwert, wann Median?***

- Bei rechts-/linksschiefen Verteilungen liefert der Mittelwert häufig einen Wert, der für den Grossteil der Beobachtungen zu hoch/tief ist. Der Median liegt meist eher dort, wo die Mehrheit der Beobachtungen sich befindet.
- für symmetrische Verteilungen ohne Ausreisser ist der Mittelwert ein gut geeignetes Lagemass. Dabei aber bedenken, dass:
- der Mittelwert ist nicht robust! D.h., eine einzige, falsche Beobachtung kann ausreichen, um ihn grob zu verfälschen!

## Grundlagen der Mathematik II Lineare Algebra und Statistik FS 2010 – Woche 09



### **Quartile, Perzentile**

Die drei **Quartile** Q1, Q2 und Q3 sind diejenigen Zahlen, welche die der Grösse nach geordneten Daten in vier Gruppen mit gleichvielen Werten unterteilen. Das erste Quartil Q1 teilt die Daten im Verhältnis 25:75. Q2 ist gleich dem Median, und Q3 wird so berechnet, dass 75% aller Werte kleiner und 25% aller Werte grösser als Q3 sind.

Es können auch beliebige a%-**Perzentile** berechnet werden. Diese teilen die Daten im Verhältnis a:(1-a).

### **Interquartile Range (IQR)**

Dieses robuste Streuungsmass deckt 50% aller Daten ab und lässt sich einfach und rasch berechnen:

$$\text{IQR} = \text{Q3} - \text{Q1}$$

### **Standardabweichung/Varianz**

Die Varianz erhält man, indem man alle quadrierten Differenzen zwischen Mittelwert und den Datenpunkten aufsummiert. Die Summe wird danach durch die (# Beobachtungen – 1) geteilt. Die Wurzel aus diesem Wert entspricht der Standardabweichung.

***Wann IQR, wann Standardabweichung?***

- Bei rechts-/linksschiefen Verteilungen liefert die Standardabweichung einen Wert, der für den Grossteil der Daten nicht repräsentativ, d.h. zu hoch ist. Die IQR liefert meist eine klar bessere Vorstellung der Streuung.
- für symmetrische Verteilungen ohne Ausreisser ist die Standardabweichung ein gut geeignetes Streuungsmass. Dabei aber bedenken, dass:
- die Standardabweichung ist nicht robust! Eine einzige, falsche Beobachtung kann ausreichen, um sie grob zu verfälschen!