

# Statistik I

# Inhalt

**Einführung**

**Kennwerte & Verteilungseigenschaften**

**Datenerhebung und Messung**

**Wichtige parametrische Verteilungen**

**Wahrscheinlichkeitsrechnung: Grundlagen  
und Definitionen**

**Schätzung & Grenzwertsätze**

**Zufallsvariablen, Verteilungen & Häufigkeiten**

**Zufallsvektoren und multivariate  
Verteilungen**

**Stochastische Unabhängigkeit und  
Zusammenhangsmaße für diskrete Merkmale**

**Zusammenhangsmaße für metrische  
Merkmale**

**Statistische Grafiken**

**Korrelation und Kausalität**

Statistik	Aufgaben	Techniken
Descriptive Statistik	Beschreibung, graphische Darstellung und Validierung von Daten. ? Keine Rückschlüsse auf Grundgesamtheit möglich.	Grafiken, Tabellen, Kennzahlen
Explorative Statistik	Suche nach Struktur in den Daten (ohne stochastische Methoden). Formulierung von Hypothesen für das den Daten zugrunde liegende stochastische Modell.	Iterative und interaktive Anwendung von Techniken aus der deskriptiven und induktiven Statistik.
Induktive Statistik	Ziehung von Schlüssen von den Daten (Stichprobe) auf Grundgesamtheit. Basierend auf stochastischen Modellen.	Statistische Modellierung, statistische Tests, Konfidenzintervalle, Schätzer

# Datenerhebung & Messung

Ein Merkmal

Matrikelnummer	Name	Vorname	Geburtsdatum	Hauptfach	Nebenfach
xxxxx 234	Muster	Peter	01.01.2001	Statistik	Informatik
xxxxx 556	Schmid	Lena	31.10.2002	Informatik	Statistik
xxxxx 123	Müller	Jonas	27.08.1999	Mathematik	NA
xxxxx 767	Nguyen	Cho	24.12.2000	Medizin	Sozialologie
xxxxx 111	Nagel	Cosima	26.10.1996	Jura	Ethik

} Alle Merkmale

Merkmaalsausprägung vom Merkmal "Nebenfach" bei der zweiten statistischen Einheit.

Eine Beobachtung

Grundgesamtheit: Studenten der LMU (über welche "Objekte" erhebe ich Daten?)

Stichprobe: z.B. Alle Statistik Studenten ! Stichproben müssen nicht per Definition zufällig gewählt sein.

Statistische Einheit / Untersuchungseinheit (UE): Ein Student bzw. ein Element der Grundgesamtheit

Merkmal: Messbare Eigenschaft einer statistischen Einheit. In der Tabelle quasi das (sinnvolle) Spaltenname. z.B. Hauptfach ist ein Merkmal

Merkmaalsausprägung: Der tatsächliche Wert des Merkmals bei einer statistischen Einheit. In der Tabelle ist das ein Wert in einer Zelle.

Beobachtung: Alle Merkmalsausprägungen einer statistischen Einheit zu einem Zeitpunkt. In der Tabelle sind das die Werte in einer Zeile.

Unterscheidung nach ...		
... Quantifizierbarkeit der Ausprägungen	Qualitative Merkmale: <ul style="list-style-type: none"> <li>nur zuordnbar (einstufig)</li> <li>Beispiele: Wohnort, Name</li> </ul>	Quantitative Merkmale: <ul style="list-style-type: none"> <li>mess- oderzählbar</li> <li>Beispiele: Alter, Körpergröße</li> </ul>
... Anzahl der Ausprägungen*	Diskrete Merkmale: <ul style="list-style-type: none"> <li>höchstens abzählbar unendlich viele mögliche Ausprägungen</li> <li>Beispiele: Gehaltsklassen, Kaufverhalten</li> </ul>	Stetige Merkmale: <ul style="list-style-type: none"> <li>überabzählbar unendlich viele mögliche Ausprägungen</li> <li>Beispiele: Geschwindigkeit, Gewicht</li> </ul>
... Direktheit der Informationsgewinnung	Beobachtbare Merkmale: <ul style="list-style-type: none"> <li>können direkt erhoben werden</li> <li>Beispiel: Abiturnote</li> </ul>	Latente Merkmale: <ul style="list-style-type: none"> <li>Operationalisierung über Indikatoren/Items notwendig</li> <li>Beispiele: Bildungsgrad, Kreativität, Nutzen</li> </ul>

(\*) Merkmale die eigentlich diskret sind, aber so viele Ausprägungen haben, dass sie wie stetige Merkmale behandelt werden können, nennt man auch quasi-stetig (z.B. Einkommen)

(\*) Stetige Merkmale können durch Klassenbildung in diskrete Merkmale umgewandelt werden.

# Skalen niveaus

Skalen niveau	Beispiele	Erlaubte Transformationen um Strukturen zu erhalten	natürliche Ordnung	sinnvolle Abstände	natürliche Null	natürliche Einheit	Berechenbare Kenntzahlen
Nominalskala	Wohnort, Farbe	Bijektionen	✗	✗	✗	✗	Mode
Ordinal - Rangskala	Noten, Michelin-Sterne Platzierung bei Sportevent	str. monoton steig. Abb.	✓	✗	✗	✗	Median
Intervallskala	Temperatur in C° Jahreszahlen	affin lin. str. mon. steig. Abb.	✓	✓	✗	✗	Arithm. Mittel
Verhältnisskala	Preis, Länge, Gewicht, Temp. in K°	lineare str. mon. steig. Abb.	✓	✓	✓	✗	Geom. Mittel Harm. Mittel
Absolutskala	Häufigkeit, Anzahl, Prozentpunkte	Identität	✓	✓	✓	✓	Alle

# Datenerhebung

Methoden:

## Beobachtung

Datengewinnung durch Erfassen von ungesteuertem Sachverhalten

## Befragung

Fragebögen für mündliche / schriftliche / online Umfrage.

## Experiment

Erzeugung der Daten durch Simulation von Situationen.

Umfang:

## Vollerhebung

Alle stat. Einheiten einer GG werden untersucht.

## Stichprobe (Teilerhebung)

Ein Teil der UE in einer GG wird untersucht.

Datenform:

## Querschnittdaten

Eine Beobachtung pro UE.

- Noten, Aktivitäten, Geschlecht, können zu bestimmtem Zeitpunkt von UE erhoben werden und z.B. mittels Regression auf Zusammenhänge untersucht werden.

## Zeitreihe

Mehrere Beobachtungen einer UE.

- Temperatur, Wind & Luftfeuchtigkeit werden in regelmäßigen Abständen gemessen um Prognosen über die zeitliche Entwicklung der UE 'Wetter' zu machen

## Längsschnittdaten

Mehrere Beobachtungen mehrerer UE.

- Kohortenstudien in Medizin
- Mikrozensus

# Wahrscheinlichkeitsrechnung

## Elementarereignisse

Für eine Grundmenge  $\Omega$  wird die ein-elementige Teilmenge  $\{w\} \subseteq \Omega$  als Elementarereignis bezeichnet

## Ereignisse

Für eine Grundmenge  $\Omega$  wird  $A \subseteq \Omega$  als Ereignis bezeichnet.

## Laplace - Wahrscheinlichkeit

Für eine abzählbare Grundmenge  $\Omega$  und ein Ereignis  $A \subseteq \Omega$  ist die Laplace - Wahrscheinlichkeit  $P(A) := \frac{|A|}{|\Omega|}$

## Wahrscheinlichkeitsverteilung (Axiome von Kolmogorov) (vereinfacht)

Sei  $\Omega$  eine Grundmenge und  $P$  ein Funktion auf  $\mathcal{P}(\Omega)$ .  $P$  heißt Wahrscheinlichkeitsverteilung oder Wahrscheinlichkeitsmaß auf  $\Omega$ , wenn sie folgende Eigenschaft erfüllt: 1)  $P(\Omega) = 1$  2)  $\forall A \subseteq \Omega : P(A) \geq 0$  3)  $\forall A, B \subseteq \Omega : A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

## Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit von  $A$  gegeben  $B$  für Ereignisse  $A, B \subseteq \Omega$  mit  $P(B) > 0$  ist  $P(A|B) := \frac{P(A \cap B)}{P(B)}$

## Folgerungen

### Korollar

- $P(\emptyset) = 0$
  - $P(\bar{A}) = 1 - P(A)$
  - $B \subseteq A \Rightarrow P(A|B) = P(A) - P(B)$
  - $B \subseteq A \Rightarrow P(B) \leq P(A)$
  - $P\left[\bigcup_{i=1}^n A_i\right] = \sum_{k=1}^n (-1)^{k+1} \cdot \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} P[A_{i_1} \cap \dots \cap A_{i_k}]$  ← Sichformel von Sylvester-Poincaré
  - $= \sum_{i=1}^n P[A_i] - \sum_{1 \leq i < j \leq n} P[A_i \cap A_j] + \sum_{1 \leq i < j < k \leq n} P[A_i \cap A_j \cap A_k] - \dots + (-1)^{n+1} \cdot P[\bigcap_{i=1}^n A_i]$
- Spezialfall:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

## Satz

- $P\left[\bigcap_{i=1}^n A_i\right] = \prod_{i=1}^n P[A_i] \quad P[A_i] = P[A_i] \cdot P[A_2 | A_1] \cdot P[A_3 | A_1, A_2] \cdot \dots \cdot P[A_n | A_1, \dots, A_{n-1}]$  (Multiplikationssatz)
  - Sei  $(A_i)_{i \in I}$  eine disjunkte Zerlegung von  $\Omega$ . D.h.  $\Omega = \bigcup_{i \in I} A_i$ . Dann gilt für beliebiges  $B$
- $$P[B] = \sum_{i: P(A_i) > 0} P[B | A_i] \cdot P[A_i] \quad (\text{Satz von totaler Wahrscheinlichkeit})$$
- ! Spezialfall:  $P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

## Kombinatorik

mit Wiederholung/  
mit Zurücklegen

ohne Wiederholung/  
ohne Zurücklegen

Anzahl Kombinationen  
ohne Reihenfolge

$$\binom{n}{m}$$

$$\binom{n+m-1}{m}$$

Anzahl Kombinationen  
mit Reihenfolge

$$\frac{n!}{(n-m)!}$$

$$n^m$$

Anzahl  
Permutationen

$$n!$$

$$\frac{n!}{n_1! \cdot \dots \cdot n_k!}$$

## Stochastische Unabhängigkeit

Eine Kollektion von Ereignissen  $(A_i)_{i \in I}$  heißt (stochastisch) unabhängig, wenn für jede endliche Kollektion  $J \subseteq I$  gilt:  $P\left[\bigcap_{i \in J} A_i\right] = \prod_{i \in J} P[A_i]$

- $A \perp B \Leftrightarrow A, B$  stochastisch unabhängig
- Paarweise Unabhängig  $\Rightarrow$  Unabhängigkeit
- $A \perp B \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$

## Satz

Die Ereignisse  $(A_i)_{i \in I}$  seien unabhängig. Für jedes  $i$  sei  $B_i = A_i \vee B_i = \bar{A}_i$ . Dann sind die Ereignisse  $(B_i)_{i \in I}$  unabhängig.

## Satz von Bayes

$(A_i)_{i \in I}$  sei so, dass  $\Omega = \bigcup_{i \in I} A_i$ .  $B$  sei so, dass  $P[B] \neq 0$ .

$$\text{Dann ist } P[A_i | B] = \frac{P[B | A_i] \cdot P[A_i]}{\sum_j P[B | A_j] \cdot P[A_j]} \quad ! \text{ Spezialfall: } P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$$

## Wettverhältnis (Odds-Update)

$$\frac{P[B|A]}{P[B'|A]} = \underbrace{\frac{P[A|B]}{P[A|B']}}_{\alpha\text{-posteriori-Verhältnis}} \cdot \underbrace{\frac{P[B]}{P[B']}}_{\text{Faktor der neuen Information}} \cdot \underbrace{\frac{P[B']}{P[B' | A']}}_{\alpha\text{-priori-Verhältnis}}$$

# Zufallsvariablen, Verteilungen & Häufigkeiten

## Zufallsvariable

Eine Zufallsvariable  $X$  ist eine Abb.  $X: \Omega \rightarrow \mathbb{R}$ .

$T := X(\Omega)$  nennen wir Träger von  $X$ .

## Verteilungsfunktion (diskret)

$$F(x) = P(X \leq x) = \sum_{i: x_i \leq x} f(x_i) = \sum_{i: x_i \leq x} P(X=x_i)$$

## Verteilungsfunktion (stetig)

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

! Stringentere Definition später in Wahrscheinlichkeitstheorie (2. Semester)

## Wahrscheinlichkeitsfunktion einer diskreten ZV

$$f(x_i) = P(X=x_i) = P(\{\omega \in \Omega | X(\omega) = x_i\}) \quad \forall x_i \in \mathbb{R}$$

$f$  heißt auch **Wahrscheinlichkeitsdichte**.

## Satz:

Jede Verteilungsfunktion  $F$  erfüllt folgende Eigenschaften:

- i) Monotonie:  $a \leq b \Rightarrow F(a) \leq F(b)$
- ii) Rechtsseitig:  $F(a) = \lim_{h \rightarrow 0^+} F(a+h) \quad \forall a \in \mathbb{R}$
- iii) Normierung:  $\lim_{a \rightarrow -\infty} F(a) = 0$  und  $\lim_{a \rightarrow \infty} F(a) = 1$

Umgekehrt ist jede Funktion  $F$  mit diesen drei Eigenschaften eine Verteilungsfunktion einer ZV. D.h.  $\exists X: F = F_X$ .

## Verteilung von $X$

$$\begin{aligned} P(X \in B) &:= P(X^{-1}(B)) \\ &= P(\{\omega \in \Omega | X(\omega) \in B\}) \end{aligned}$$

## (diskret)

$X$  ist eine diskrete ZV, wenn der Träger von  $X$  eine abzählbare Menge ist.

## (stetig)

$X$  ist eine stetige ZV, wenn es eine Funktion  $f$  gibt, für die gilt

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x) dx = 1$
- $\forall b \in \mathbb{R}: F(b) = \int_{-\infty}^b f(x) dx$

## Indikatorfunktion

Die Indikatorfunktion dient dazu zu checken ob ein Element in einer bestimmten Menge enthalten ist.

Sei:  $A \subseteq \mathbb{R}$ .

$$I_A(x) = \begin{cases} 0, & \text{wenn } x \notin A \\ 1, & \text{wenn } x \in A \end{cases}$$

Andere Schreibweisen sind  $\mathbb{1}_A$ ,  $\mathbb{I}_A$ ,  $I(x \in A)$

## Notation & Terminologie

Bezeichnung in der Empirie	Notation/Berechnung in der Theorie
Merkmal	Zufallsvariable $X, Y, \dots$
Anzahl Untersuchungseinheiten	$n$
Merkmaalsausprägung von Merkmal $X$ der $i$ -ten UE.	$x_i, i \in \{1, \dots, n\}$
Rohdaten / Urliste	$x_1, \dots, x_n$
(evtl. geordnete) verschiedene Werte aus der Urliste	$a_1, \dots, a_k, k \leq n, a_i \in \{x_1, \dots, x_n\}$ (Eindeutige Elemente der Urliste)
relative Häufigkeit $f_j$	Wahrsch.fkt. bzw. Dichte $f_X(a_j)$
kum. rel. Häufigkeit $F(x)$	Verteilungsfunktion $F_X(x)$

## Absolute Häufigkeit

Die absolute Häufigkeit von  $a_j$  ist die Anzahl der  $x_i$  aus der Urliste mit  $x_i = a_j$   
 $h(a_j) = h_j$

## Relative Häufigkeit

Die relative Häufigkeit von  $a_j$  ist der Anteil der  $x_i$  an der Urliste für die gilt  $x_i = a_j$ .  $f(a_j) = f_j = \frac{h_j}{n}$

## Absolute / relative Häufigkeitsverteilung

$h_1, \dots, h_k$  heißt absolute Häufigkeitsverteilung  
 $f_1, \dots, f_k$  heißt relative Häufigkeitsverteilung

## Kumulative relative Häufigkeit/empirische Verteilungsfunktion

(Sinnvoll bei ordinal oder metrisch)

$$F(x) = (\text{Anteil UE mit } x_i \leq x) = \sum_{a_i \leq x} f(a_i) \quad (\text{ECDF})$$

- monoton wachsende Treppenfkt. mit Sprüngen bei  $a_1, \dots, a_k$ .
- Sprunghöhe  $f_1, \dots, f_k$
- rechtsseitig stetig
- $F(x) = 0$  für  $x < a_1$ ,  $F(x) = 1$  für  $x \geq a_k$

# Zusammenhangsmaße für diskrete ZV

## Kontingenztafel der absoluten Häufigkeiten

Seien  $X, Y$  diskrete Merkmale mit Ausprägungen  $a_1, \dots, a_k$  für  $X$  und  $b_1, \dots, b_m$  für  $Y$ . Eine  $(k \times m)$ -Kontingenztafel der absoluten Häufigkeiten besitzt die Form:

$$h_{ij} = h(a_i, b_j) = \text{Absolute Häufigkeit der Kombination } (a_i, b_j)$$

$$h_{i \cdot} = \text{Randhäufigkeit von } a_i \text{ in } X; \quad h_{\cdot j} = \text{Randhäufigkeit von } b_j \text{ in } Y$$

$a_1$	$b_1 \dots b_m$
$a_2$	$h_{11} \dots h_{1m}$
$\vdots$	$\vdots$
$a_k$	$h_{k1} \dots h_{km}$
	$h_{\cdot 1} \dots h_{\cdot m}$
	$n$

## Kontingenztafel der relativen Häufigkeiten

Seien  $X, Y$  diskrete Merkmale mit Ausprägungen  $a_1, \dots, a_k$  für  $X$  und  $b_1, \dots, b_m$  für  $Y$ . Eine  $(k \times m)$ -Kontingenztafel der relativen Häufigkeiten besitzt die Form:

$$f_{ij} = \frac{h_{ij}}{n} = \text{Relative Häufigkeit der Kombination } (a_i, b_j)$$

$$f_{i \cdot} = \frac{h_{i \cdot}}{n} = \text{relative Randhäufigkeit von } a_i \text{ in } X;$$

$$f_{\cdot j} = \frac{h_{\cdot j}}{n} = \text{relative Randhäufigkeit von } b_j \text{ in } Y$$

$a_1$	$b_1 \dots b_m$	$f_{11} \dots f_{1m}$	$f_{1 \cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_k$	$b_k \dots b_m$	$f_{k1} \dots f_{km}$	$f_{k \cdot}$
	$f_{\cdot 1} \dots f_{\cdot m}$		$1$

## Bedingte Häufigkeitsverteilung

Die bedingte Häufigkeitsverteilung von  $Y$  unter der Bedingung  $X=a_i$ ,  $(Y|X=a_i)$  ist definiert als  $f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i \cdot}}$ ,  $\dots, f_Y(b_m|a_i) = \frac{h_{im}}{h_{i \cdot}}$

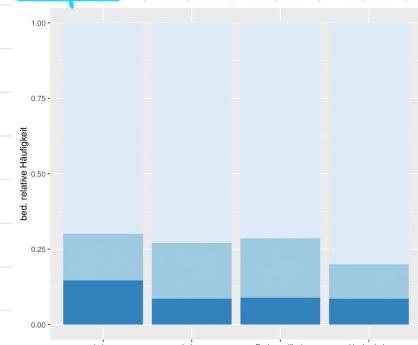
## Satz

Wegen  $\frac{h_{ij}}{h_{i \cdot}} = \frac{h_{ij}/n}{h_{i \cdot}/n} = \frac{f_{ij}}{f_{i \cdot}}$  gilt,  
 $f_Y(b_j|a_i) = \frac{f_{ij}}{f_{i \cdot}} \quad \forall i, j$

## Erwartete absolute/relative Häufigkeit

Unter empirischer Unabhängigkeit wird erwartet, dass  $f_Y(b_j|a_i) = f_Y(b_j) \quad \forall i, j$  gilt.  
Daher gilt für die erwartete absolute Häufigkeit  $\tilde{h}_{ij}$ :  $\tilde{h}_{ij} = \frac{h_{i \cdot} \cdot h_{\cdot j}}{n}$  (nicht immer  $\in \mathbb{N}$ )  
Und für die erwartete relative Häufigkeit  $\tilde{f}_{ij}$ :  $\tilde{f}_{ij} = f_{i \cdot} \cdot f_{\cdot j}$

## Beispiel



## $\chi^2$ -Koeffizient

Zusammenhangsmaß zum Quantifizieren vom "Abstand" zwischen beobachteten und erwarteten gemeinsamen Häufigkeiten.

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = n \cdot \sum_{i=1}^k \sum_{j=1}^m \frac{(f_{ij} - \tilde{f}_{ij})^2}{\tilde{f}_{ij}}$$

$$\chi^2 \in [0, n \cdot \min(k, m) - 1]$$

$$\chi^2 = 0 \Leftrightarrow X, Y \text{ empirisch unabhängig}$$

$\chi^2$  gross  $\Leftrightarrow$  starker Zusammenhang

?  $\chi^2$  hängt von  $n, k$  und  $m$  ab

$\hookrightarrow$  schwer zu interpretieren.

## Satz

Für eine Kontingenztafel der Form

$a$	$b$
$c$	$d$

gilt

$$\chi^2 = \frac{n \cdot (a \cdot d - c \cdot b)^2}{(a+b)(a+c)(c+d)(b+d)}$$

## (Korrigierter) Kontingenzkoeffizient

! Misst nur Stärke des Zusammenhangs,

Nicht die Richtung, wie bei  $\gamma$

Normierung von  $\chi^2$ : Kontingenzkoeffizient  $K := \sqrt{\frac{\chi^2}{n + \chi^2}}$ ,  $K \in [0, \sqrt{\frac{\min(k, m) - 1}{\min(k, m)}}]$

Korrigierter Kontingenzkoeffizient  $K^* := \frac{K}{\sqrt{\frac{\min(k, m) - 1}{\min(k, m)}}}$ ,  $K^* \in [0, 1]$

## Bedingte Odds & Odds ratio

Für festes  $X=a_i$  bezeichnen wir  $\gamma(1,2|X=a_i) = \frac{h_{11}}{h_{21}}$  als bedingte Odds.

Als relative Chancen (Odds ratio) bezeichnen wir  $\gamma(1,2|X=1, X=2) = \frac{\gamma(1,2|X=1)}{\gamma(1,2|X=2)} = \frac{h_{11} \cdot h_{22}}{h_{21} \cdot h_{12}}$

Odds ratio (Kreuzproduktverhältnis) ist symmetrisch bezüglich der Wahl von  $X, Y$ .

$$\gamma(Y=1, Y=2 | X=1, X=2) = \gamma(X=1, X=2 | Y=1, Y=2)$$

! Symmetrisches Maß und kann als Risikofaktor interpretiert werden

•  $\gamma=1$ : Odds in beiden Populationen gleich.

•  $\gamma > 1$ : Odds in  $X=1$  höher als in  $X=2$

•  $\gamma < 1$ : Odds in  $X=1$  niedriger als in  $X=2$ .

# Statistische Grafiken

## Grammatik von Grafiken

Grafik = Daten + geometrische Elemente + Ästhetische Zuordnung  
 + Datentransformationen + Skalen + Koordinatensysteme  
 + Facettierung + Theme + {Grafik}

Geometrische Elemente = Punkte | Linien | Rechtecke | Boxplots | Dichtefkt. | ...

Ästhetische Zuordnung = Position & Farbe & Größe & Form

Datentransformationen = id | Mittelwerte | Anteile | ...

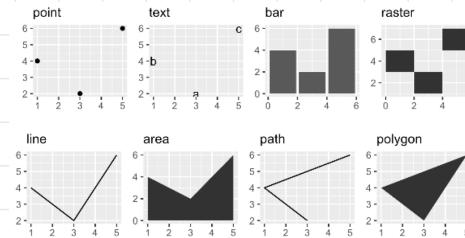
Skalen = Achsenabschnitte & Farbe & Legenden & Achsenbeschriftung & ...

Koordinatensysteme = kartesisch | logarithmisch | Polarkoord. | Kartenproj. | ...

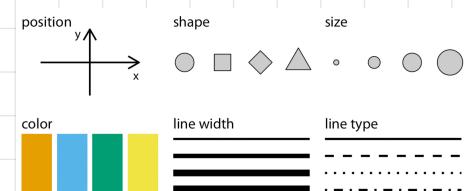
Facettierung = small multiples | lattice plot | plot | ...

Theme = Font & Gitterlinien & Hintergrundfarben & Layout von Text & ...

## Beispiele Geometrien

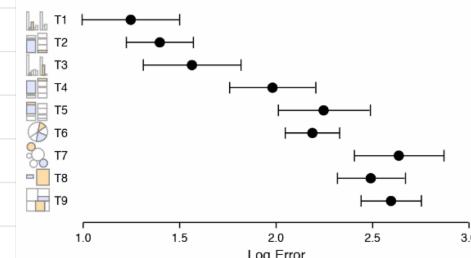


## Beispiele Ästhetiken



## Wahrnehmung von Grafiken

### Crowdsourced Results



Hierarchie der korrekten Interpretation:

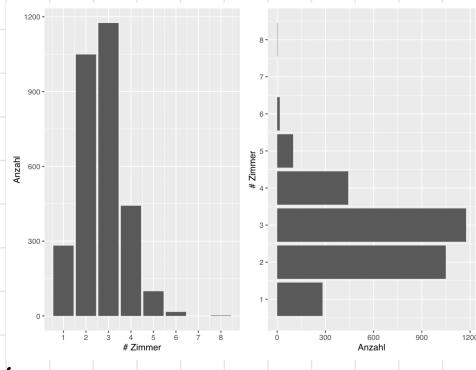
1. Position
2. Abstände / Längen
3. Steigung
4. Winkel
5. Flächen
6. Volumen
7. Farbe (Ton, Helligkeit, Sättigung)

## Goldene Regeln für Grafikgestaltung

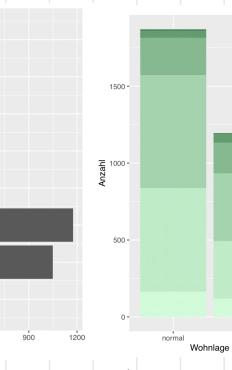
- Kommunikationsabsicht klarmachen
- Lesbarkeit maximieren

## Visualisierung von Häufigkeiten & Verteilungen diskreter Merkmale

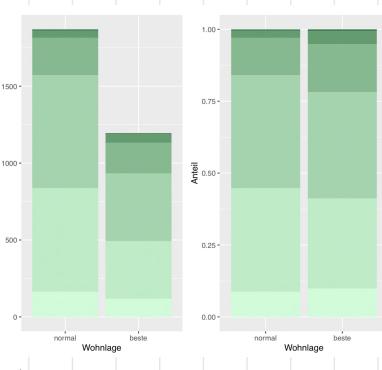
### Säulendiagramm



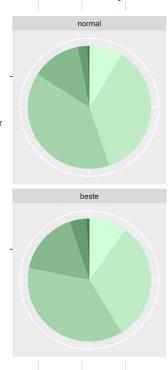
### Balkendiagramm



### Stapeldiagramm (absolut & relativ)



### Kreisdiagramm



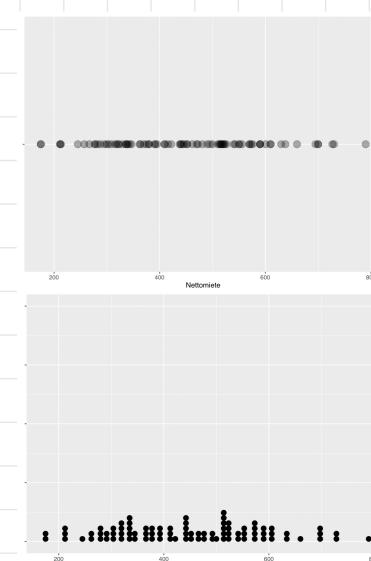
Für ordinale Merkmale, metrische Merkmale mit wenig Ausprägung und nominale Merkmale, wobei die Anordnung beliebig ist. Breite ist beliebig.

Anwendbar für die gleichen Merkmale wie zuvor.  
 Besonders geeignet für den Vergleich verschiedener Gruppen (bedingte Häufigkeiten).

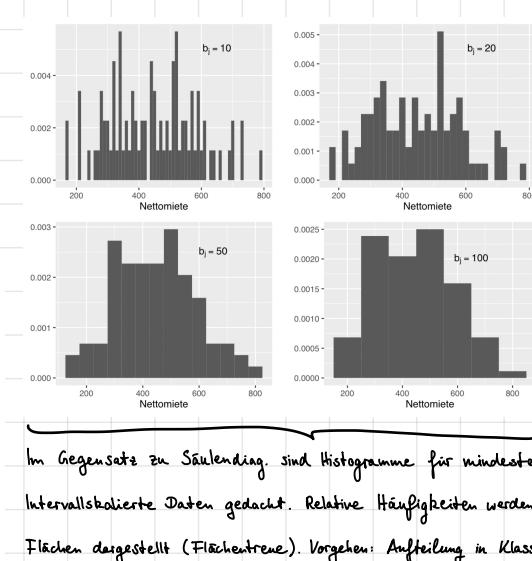
Grundsätzlich schwieriger zu interpretieren als Längendiagramme, aber enthält keine klare Ordnung.

## Visualisierung von Häufigkeiten & Verteilungen metrischer Merkmale

### Dotplots



### Histogramme



Im Gegensatz zu Säulendiag. sind Histogramme für mindestens intervallskalierte Daten gedacht. Relative Häufigkeiten werden durch Flächen dargestellt (Flächentrenne). Vorgehen: Aufteilung in Klassen und Bestimmung der relativen Häufigkeiten  $f_i = \frac{n_i}{n}$ . Höhe  $y_i$  des Balkens bestimmen mittels  $b_i \cdot y_i = f_i$ , wobei  $b_i$  die Breite der Klasse  $i$  ist.

Nachteil: Interpretation der Höhe bei unterschiedlichen Breiten nicht sinnvoll.  
 • Visueller Eindruck hängt von Klassenbreiten ab.  
 • Vorsicht bei Rändern.

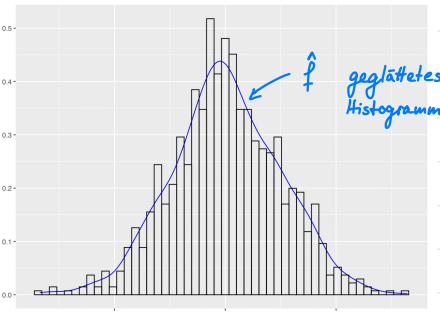
## Kerndichteschätzung

### Kerndichteschätzer

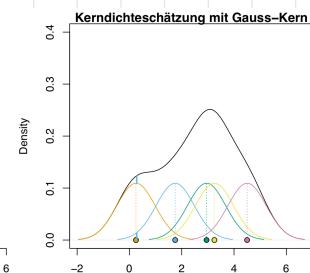
Sei  $K$  eine Kernfunktion, d.h.  $\forall u: K(u) \geq 0$  und  $\int_{-\infty}^{\infty} K(u) du = 1$ .

Dann ist der Kerndichteschätzer (KDE: kernel density estimator) definiert als

$$\hat{f}(x) := \frac{1}{n \cdot h} \cdot \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)$$

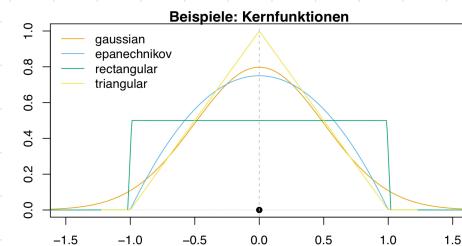


Histogram



### Beispiele für Kernfunktionen

- Gauß-Kern:  $K(u) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{1}{2}u^2)$
- Epanechnikov-Kern:  $K(u) := \max\{0, \frac{3}{4} \cdot (1-u^2)\}$
- Dreieck-Kern:  $K(u) := \max\{0, 1 - |u|\}$

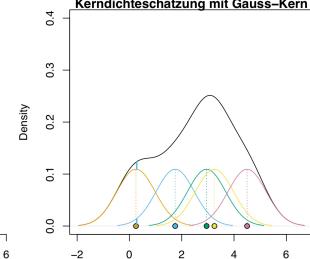
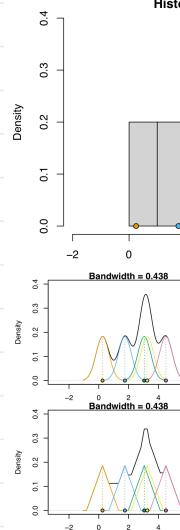


KDE = Histogramm bei größeren Datenmengen oder (quasi-)stetigen Merkmalen.

Vorteil:  
Kerndichteschätzungen berücksichtigen Entfernung der benachbarten Punkte mit abnehmender Gewichtung über Distanz.

Nachteil:

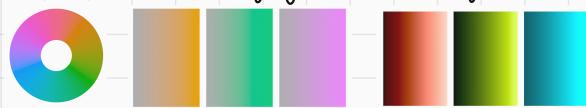
Abhängigkeit von Bandbreite  $h \rightarrow$  Wird aus den Daten bestimmt.



## Farbskalen

### Farbraum

Der Farbraum ist definiert durch die Wahlmöglichkeiten für Farbton, Farbsättigung und Helligkeit.



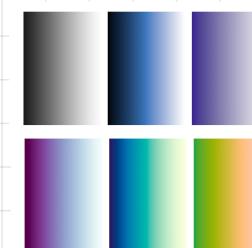
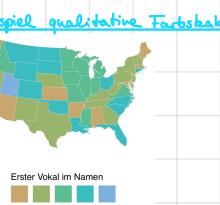
### Farbskalentypen

#### Farbskalentypen

- Qualitativ: (eher) nur für nominales Skalenniveau.
- Sequential: mindestens ordinates Skalenniveau.  $\rightarrow$
- Divergent: mindestens ordinates Skalenniveau mit "neutralen" mittlerem Wert



2 sequentielle Farbskalen mit je konstantem Farbton kombiniert.



Farbe konstant, Sättigung & Helligkeit variert.

Farbe, Helligkeit und Sättigung variert.



Electoral College Votes

10 20 30 40 50

Sinnvoll, aber evtl. kleine Unterschiede verdeckt

### Visualisierung gemeinsamer Verteilungen metrischer Merkmale

