

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

(*) 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 Kennzahlen |
|---------------------|--|---|-----------------------|-----------------------|--------------------|-----------------------|------------------------------|
| 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 Ω wird die ein-elementige Teilmenge $\{w\} \subseteq \Omega$ als Elementarereignis bezeichnet.

Ereignisse

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

Laplace - Wahrscheinlichkeit

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

Wahrscheinlichkeitsverteilung (Axiome von Kolmogorov) (vereinfacht)

Sei Ω eine Grundmenge und P ein Funktion auf $\mathcal{P}(\Omega)$. P heißt Wahrscheinlichkeitsverteilung oder Wahrscheinlichkeitsmaß auf Ω , 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 Ω . 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

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

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

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

(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:

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

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

| | |
|----------|---------------------------------|
| 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,
 $f_Y(b_j|a_i) = \frac{f_{ij}}{f_{i \cdot}} \quad \forall i, j$

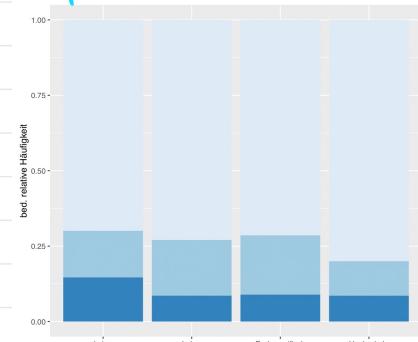
Erwartete absolute/relative Häufigkeit

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



χ^2 -Koeffizient

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

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

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

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

χ^2 gross \Leftrightarrow starker Zusammenhang

? χ^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 γ

Normierung von χ^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 | ...

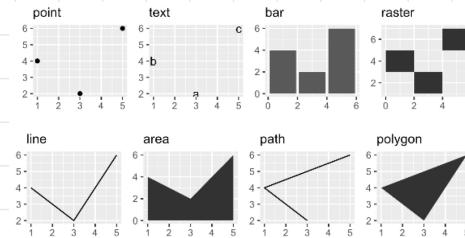
Skalen = Achsenabschnitte & Farbe & Legenden & Achsenbeschriftung & ...

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

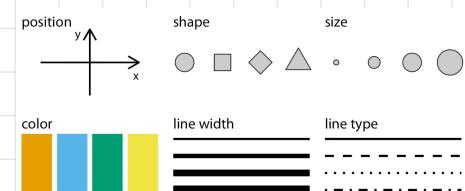
Facettierung = small multiples | lattice plot | plot | ...

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

Beispiele Geometrien

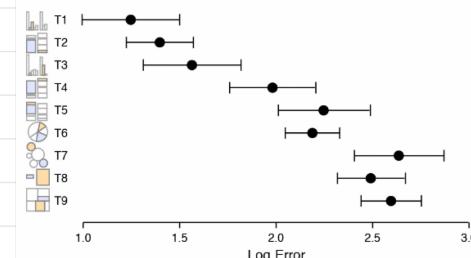


Beispiele Ästhetiken



Wahrnehmung von Grafiken

Crowdsourced Results



Hierarchie der korrekten Interpretation:

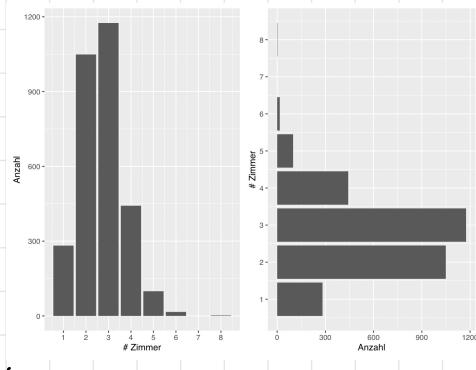
1. Position
2. Abstände / Längen
3. Steigung
4. Winkel
5. Flächen
6. Volumen
7. Farbe (Ton, Helligkeit, Sättigung)

Goldene Regeln für Grafikgestaltung

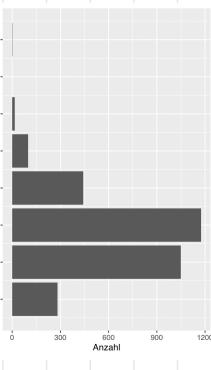
- Kommunikationsabsicht klarmachen
- Lesbarkeit maximieren

Visualisierung von Häufigkeiten & Verteilungen diskreter Merkmale

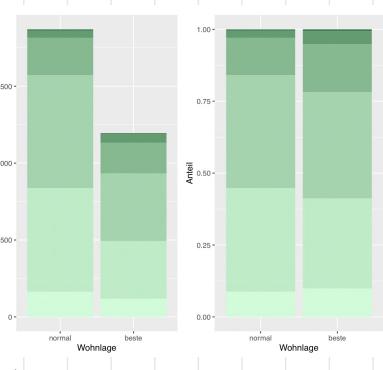
Säulendiagramm



Balkendiagramm



Stapeldiagramm (absolut & relativ)



Kreisdiagramm



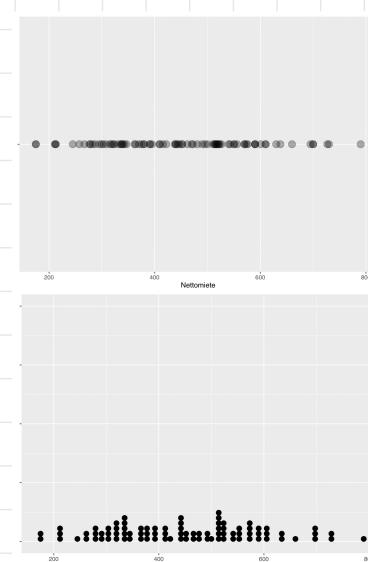
Für ordinale Merkmale, metrische Merkmale mit wenig Ausprägung und nominale Merkmale, wobei die Anordnung beliebig ist. Breite ist beliebig.

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

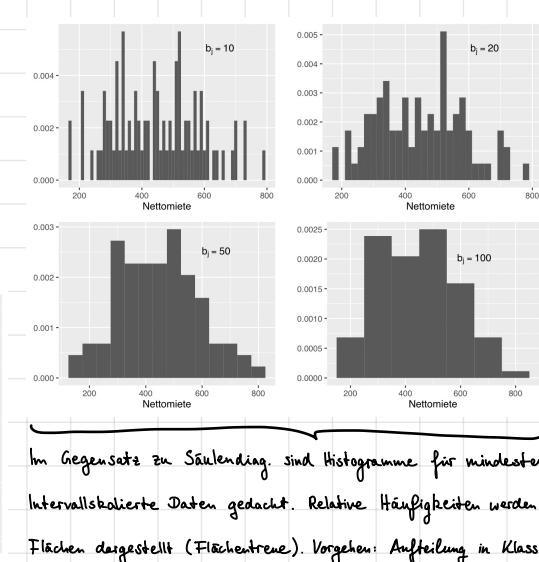
Grundsätzlich schwieriger zu interpretieren als Längendiagramme, aber enthält keine klare Ordnung.

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.

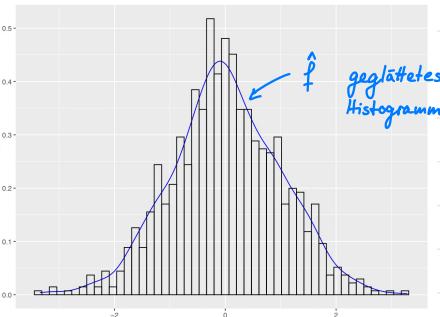
Kerndichteschätzung

Kerndichteschätzer

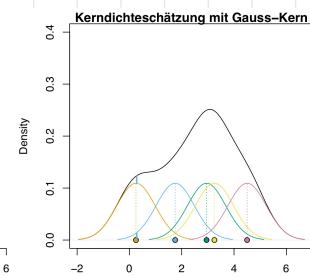
Sei K eine Kernfunktion, d.h. $\forall u: K(u) \geq 0$ und $\int_{-\infty}^{\infty} K(u) du = 1$.

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

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

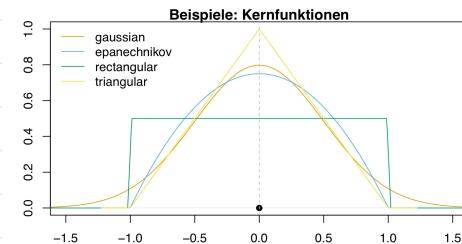


Histogram



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: $K(u) := \max\{0, 1-|u|\}$



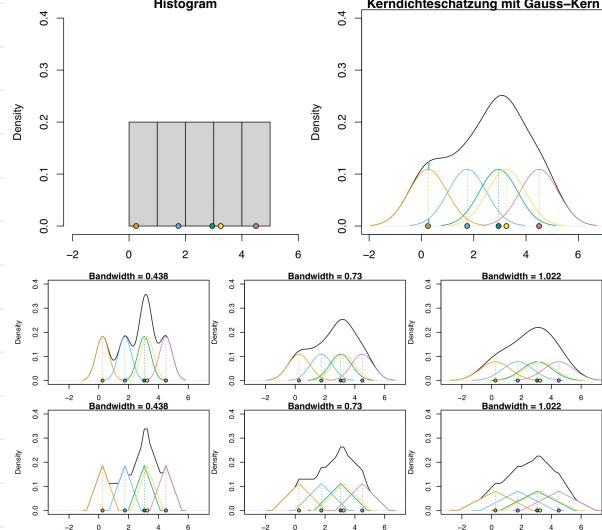
KDE = Histogramm bei größeren Datenmengen oder (quasi-)stetigen Merkmalen.

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

Nachteil:

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

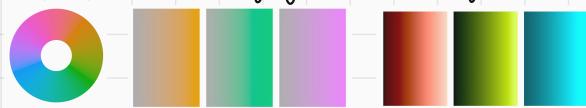
Histogram



Farbskalen

Farbraum

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



Farbraum

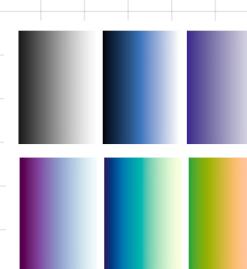
Farbskalentypen

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



2 sequentielle Farbskalen mit je konstantem Farbton kombiniert.

Beispiel divergente Farbskala



Farbe konstant, Sättigung & Helligkeit variiert.

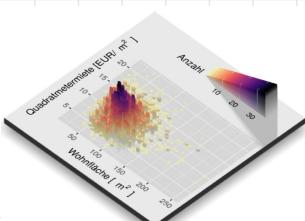
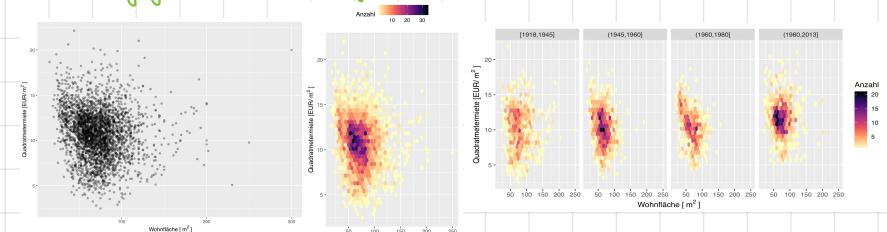
Farbe, Helligkeit und Sättigung variiert.

Beispiel sequentielle Farbskala



Sinnvoll, aber evtl. kleine Unterschiede verdeckt

Visualisierung gemeinsamer Verteilungen metrischer Merkmale



Kennwerte & Verteilungeigenschaften

Lagemaße

Modus (Emperie)

Häufigster Wert in einer Stichprobe.

- oft nicht eindeutig
- nur bei gruppierten Daten oder Merkmalen mit wenig Ausprägungen sinnvoll.
- + stabil bei allen injektiven Transformationen
- + geeignet für jedes Skalenniveau.

Modus (Theorie)

Für eine ZV X ist der Modus definiert als $x_{\text{mod}} \in \mathbb{R}$ s.d. $f(x_{\text{mod}}) \geq f(x) \quad \forall x \in T_X$.

x_{mod} ist die Maximumstelle von f_X

(nicht zwingend eindeutig oder existent).

Der Mittelwert (Emperie)

Arithmetisches Mittel: $\bar{x}_{\text{arith}} := \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k h_i \cdot a_i = \sum_{i=1}^k f_i \cdot a_i$

- instabil gegenüber Ausreißern
- mindestens intervallskalierte Daten
- + bekanntestes Lagemaß und wichtiger theoretischer Nutzen.

Gewichtetes Mittel: $\bar{x}_W := \frac{1}{\sum_{i=1}^n w_i} \cdot \sum_{i=1}^n w_i \cdot x_i$, mit $w_i \geq 0 \quad \forall i$

Geometrisches Mittel: $\bar{x}_G := \sqrt[n]{\prod_{i=1}^n x_i} = \exp\left(\frac{1}{n} \sum_{i=1}^n \log(x_i)\right)$.

- Mindestens Verhältnisskala ($x_i > 0$)
- + Anwendbar auf multiplikative Faktoren

Harmonisches Mittel: $\bar{x}_H := \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{x_i}} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}\right)^{-1}$

- Mindestens Verhältnisskala ($x_i > 0$)
- + Anwendbar auf Quotienten/Verhältnisse.

Getrimmtes Mittel: Sei $\alpha \in (0,1)$, $x_{(1)}, \dots, x_{(n)}$ die geordnete Stichprobe und $r = \max\{\tilde{f} \in \mathbb{Z} | \tilde{f} \leq n \cdot \alpha\}$.

Wir definieren das α -getrimmte Mittel als $\bar{x}_\alpha := \frac{1}{n-2r} \cdot \sum_{i=r+1}^{n-r} x_{(i)}$.

$\approx \alpha$ -Anteil der extremsten Werte wird abgeschnitten

Alternativ: $\bar{x}_\alpha := \frac{1}{n} \cdot \left(\sum_{i=r+1}^{n-r} x_{(i)} + r \cdot \bar{x}_\alpha + r \cdot \bar{x}_{1-\alpha} \right)$
 α -Quantil

Lagemaßzahlen

- > Wo liegt die Masse, Mitte und Mehrzahl der Daten?
- > Welche Merkmalsausprägung ist typisch für die Verteilung?

Quantil/Percentil (Emperie)

$x_{(1)}, \dots, x_{(k)}$ sind die geordneten Werte der Stichprobe.

Das p -Quantil ($p \in [0,1]$) ist der Wert \tilde{x}_p für den gilt:

Anteil p der Daten sind $\leq \tilde{x}_p$ & Anteil $1-p$ der Daten sind $\geq \tilde{x}_p$.

$$\tilde{x}_p := \begin{cases} x_{(k)} & , \text{ falls } n \cdot p \notin \mathbb{N}_0 \text{ und } k > kp \\ \frac{1}{2} \cdot (x_{(k)} + x_{(k+1)}) & , \text{ falls } k = n \cdot p \in \mathbb{N}_0 \end{cases}$$

(Alternativ beliebigen Wert $\tilde{x}_p \in [x_{(k)}, x_{(k+1)}]$ für $k = n \cdot p \in \mathbb{N}_0$)

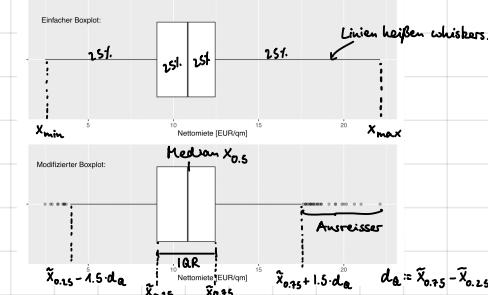
Quantil (Theorie)

Das p -Quantil x_p einer ZV X mit $p \in [0,1]$ ist definiert als $x_p := \inf\{x \in \mathbb{R} \mid F_X(x) \geq p\}$
(Wenn $f(x) > 0 \quad \forall x$ bzw. F_X str. mon. steigend, dann gilt $x_p = F_X^{-1}(p)$)

Median (Emperie)

$\tilde{x}_{\text{med}} := \tilde{x}_{0.5} := 50\%-Perzentil$

- + anschaulich
- + stabil gegenüber monotonen Transformationen
- + geeignet für mindestens ordinale Daten
- + stabil gegenüber Ausreißern



Mittelwert / Erwartungswert (Theorie)

Discrete ZV X : $E(X) := \sum_{w \in \Omega} P(w) \cdot X(w) = \sum x \cdot P(X=x) = \sum x \cdot f(x)$

Stetige ZV X : $E(X) := \int x \cdot f(x) dx$

Satz

- $E(a+bX) = a+b \cdot E(X) \quad \forall a, b \in \mathbb{R}$
- $E(X+Y) = E(X) + E(Y) \quad \text{für alle ZV } X, Y$
- $\exists c \in \mathbb{R} \forall x \in T_X : f(c-x) = f(c+x) \Rightarrow E(X) = c$.
- Für $g: \mathbb{R} \rightarrow \mathbb{R}$ und $y = g(X)$, dann gilt $E(y) = \begin{cases} \sum_{x \in T_X} g(x) f(x) & , X \text{ diskret} \\ \int g(x) f(x) dx & , X \text{ stetig} \end{cases}$
- Ist $T_X = \mathbb{N}^*$, so gilt $E(X) = \sum_{k=1}^{\infty} P(X=k)$

Jensen - Ungleichung

Für eine ZV X mit endlichem Erwartungswert und $g: \mathbb{R} \rightarrow \mathbb{R}$ konvex gilt $E[g(X)] \geq g(E(X))$

Markov - Ungleichung

Sei g eine nicht negative, monoton wachsende Funktion auf \mathbb{R} . Dann gilt $\forall c \in \mathbb{R}$ mit $g(c) > 0$

$$P[X \geq c] \leq \frac{E[g(X)]}{g(c)}$$

Streuungsmaße

Streuungsmaßzahlen

- > Über welchen Bereich erstrecken sich die Ausprägungen?
- > Wie groß ist die Schwankung der beobachteten Werte?
- > Wie eng beieinander liegen die beobachteten Werte?

Sätze

Sei $Y = a + bX$, für $a, b \in \mathbb{R}$. Dann gilt:

$$\sigma_y^2 = b^2 \cdot \sigma_x^2 ; \quad \sigma_y = |b| \cdot \sigma_x ; \quad s_y^2 = b^2 \cdot s_x^2 ; \quad s_y = |b| \cdot s_x$$

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

$$x_1, \dots, x_n, x_{n+1}, \dots, x_{n+n_2}, \dots, x_n \text{ mit } n = \sum_{j=1}^r n_j$$

$$\text{Schichtmittelwerte: } \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i ; \bar{x}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i ; \dots$$

$$\text{Schichtvarianz: } \tilde{s}_{x_j}^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_i - \bar{x}_j)^2 ; \quad \tilde{s}_x^2 = \frac{1}{n} \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2$$

$$\text{Dann gilt: } s_x^2 = \frac{1}{n} \sum_{j=1}^r n_j \cdot \tilde{s}_{x_j}^2 + \frac{1}{n} \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2$$

Gesamtstreuung = Streuung innerhalb der Schichten + Streuung zwischen den Schichten

Verschiebungssatz

$$\forall c \in \mathbb{R}: \sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n \cdot (\bar{x} - c)^2$$

$$\stackrel{!}{=} \Rightarrow \text{Var}(x) = E(x^2) - E(x)^2$$

Streuungszerlegung der Netto-Quadratmetermiete bezüglich Zimmerzahl:

| Zimmer | n_j | \bar{x}_j | $\tilde{s}_{x_j}^2$ |
|--------|-------|-------------|---------------------|
| 8 | 2 | 6.2 | 1.2 |
| 6 | 16 | 10.0 | 13.2 |
| 5 | 99 | 9.9 | 7.3 |
| 4 | 442 | 10.2 | 7.0 |
| 3 | 1175 | 10.4 | 7.1 |
| 2 | 1049 | 10.9 | 6.1 |
| 1 | 282 | 12.6 | 6.2 |

Gesamtvarianz: $\tilde{s}_x^2 \approx 7.15$

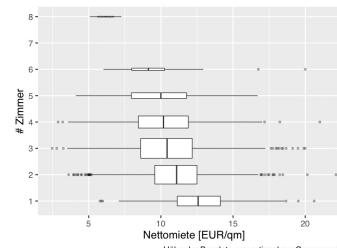
Innerhalb: $\frac{1}{n} \sum_{j=1}^r n_j \tilde{s}_{x_j}^2 = 6.7$

Zwischen: $\frac{1}{n} \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2 = 0.45$

Variationskoeffizient

$$v = \frac{\sigma_x}{\bar{x}} \quad \text{mit } \bar{x} > 0.$$

Skalierungsunabhängige Maßzahl für relative Schwankungen um \bar{x} .



⇒ nur $\frac{0.45}{7.15} = 6.25\%$ der Gesamtvarianz der Quadratmetermiete entfallen auf Unterschiede zwischen Wohnungen mit unterschiedlicher Zimmerzahl.

Standardabweichung einer ZV

Die Standardabweichung $\sigma(x)$ einer ZV X ist definiert als $\sigma(x) = \sqrt{\text{Var}(x)}$ bzw. $s_x = \sqrt{s_x^2}$ für Stichproben

Mittlere absolute Abweichung (L1-Loss)

$$\text{MAD}_X := \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|. \quad \text{Es gilt } \text{MAD}_X \leq s_x$$

$$\text{MedAD}_X := \text{Median}(|x_i - \bar{x}_{\text{med}}|)$$

+ einfacher zu interpretieren als s_x .

+ Weniger Ausreißer-empfindlich.

- Weniger schöne theoretische Eigenschaften als s_x .

Varianz einer ZV

$$! E(|x|) < \infty$$

Die Varianz $\text{Var}(X)$ einer ZV X ist definiert als:

$$\text{Var}(X) := E[(X - E(X))^2] = \begin{cases} \sum_{x \in T_X} (x - E(X))^2 p(x=x), & X \text{ diskret} \\ \int_{\mathbb{R}} (x - E(X))^2 f_X(x) dx, & X \text{ stetig} \end{cases}$$

Beispiel (Pareto Verteilung)

$$f_X(x) = \begin{cases} \alpha \cdot \frac{x^\alpha}{x^{\alpha+1}}, & x > x_0 \\ 0, & x \leq x_0 \end{cases} \quad \text{finite } E(x), \text{ but infinite } \text{Var}(x) \text{ for } \alpha \in (1, 2]$$

$$E(X) = \begin{cases} \infty, & \alpha \leq 1 \\ \frac{\alpha \cdot x_0}{\alpha - 1}, & \alpha > 1 \end{cases}$$

$$\text{Var}(X) = \begin{cases} \infty, & \alpha \leq 2 \\ \frac{\alpha x_0^2}{(\alpha - 1)^2 \cdot (\alpha - 2)}, & \alpha > 2 \end{cases}$$

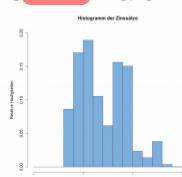
Chebyshew - Ungleichung

Für $Y = |X - E[X]|$ und $g(x) = (\max\{x, 0\})^2$ folgt durch die Markov Ungleichung

$$P[|X - E[X]| > c] \leq \frac{\text{Var}[X]}{c^2}$$

Verteilungseigenschaften

unimodal = eingespielt, multimodal = mehrspiegelig



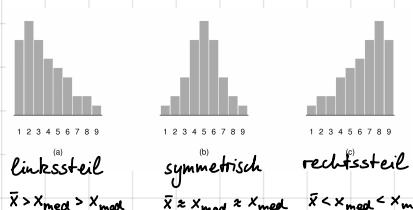
Das Histogramm der Zinssätze zeigt eine bimodale (trimodale...?) Verteilung.

Symmetrie & Schiefe

Symmetrisch ⇔ Annähernd spiegelsymm.

linkssteil ⇔ Verteilung fällt nach links (rechtsschief) deutlich steiler und nach rechts langsamer ab.

rechtssteil ⇔ Verteilung fällt nach rechts (linksschief) deutlich steiler und nach links langsamer ab.



Quartilskoeffizient

$$g_p := \frac{(\bar{x}_{1-p} - \bar{x}_{\text{med}}) - (\bar{x}_{\text{med}} - \bar{x}_p)}{\bar{x}_{1-p} - \bar{x}_p}, \quad p \in (0, 1)$$

$g_{0.25}$ nennt man auch Quartilskoeffizient.

Symmetrisch: $g_p = 0$

linkssteil: $g_p > 0$

rechtssteil: $g_p < 0$

(Zentrierter) Moment einer ZV

Für $p \in \mathbb{N}$ heißt

- p -tes Moment: $E[X^p]$

- p -tes absolutes Moment: $E[|X|^p]$

- p -tes zentriertes Moment: $E[(X - E[X])^p]$

Fisher's Momentkoeffizient für Schiefe (3. normiertes Moment)

$$g_{m_3} = E \left[\left(\frac{X - E(X)}{\sigma_X} \right)^3 \right] = \frac{E[(X - E(X))^3]}{E[(X - E(X))^2]^{\frac{3}{2}}} = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^3}{\sigma_X^3} \quad \text{wenn } E(X) < \infty$$

symmetrisch: $g_{m_3} = 0$

linkssteil: $g_{m_3} > 0$

rechtssteil: $g_{m_3} < 0$

$$\text{Corrected skew: } \tilde{g}_m = \frac{n \cdot \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2) \cdot \hat{s}_X^3}$$

Kurtosis (4. normiertes Moment)

$$k_X := \text{kurt}(X) := E \left[\left(\frac{X - E(X)}{\sigma_X} \right)^4 \right] = \frac{E[(X - E(X))^4]}{E[(X - E(X))^2]^2} = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^4}{\sigma_X^4}$$

Exzess-Kurtosis: $k_X^* = k_X - 3$

mesokurtisch: $k_X^* \approx 0$

leptokurtisch: $k_X^* > 0$. Viele extreme Werte

platykurtisch: $k_X^* < 0$. Wenig extreme Werte

$$\text{Sample kurtosis: } \tilde{k} = \frac{n \cdot (n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\hat{s}_X^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Konzentrationsmaße

Lorenzkurve

Das Merkmal darf nur positive Werte annehmen.

$x_{(1)}, \dots, x_{(n)}$ sei die geordnete Stichprobe.

Die Lorenzkurve verbindet Punktpaare bestehend aus den Teilsummen von $x_{(1)}$ (d.h. $\sum_{i=0}^k x_{(i)}$) und dem relativen Anteil an Individuen, die diese Teilsumme besitzen.

Berechnung

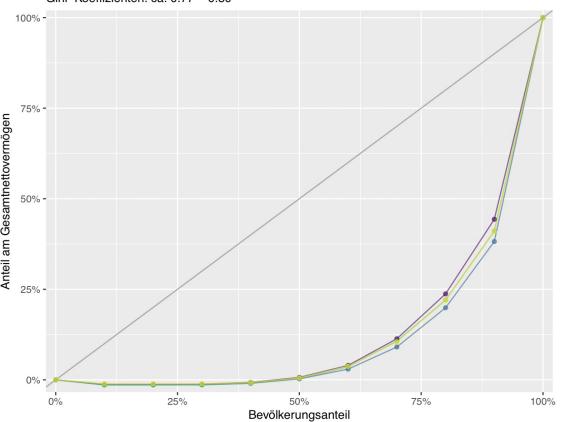
$$u_{(0)} = 0, \quad v_{(0)} = 0, \quad j = 1, \dots, n.$$

$$u_{(j)} := \frac{j}{n} \quad (\text{Aufteilung der } x\text{-Achse})$$

$$v_{(j)} := \frac{\sum_{i=1}^j x_{(i)}}{\sum_{i=1}^n x_{(i)}} \quad (\text{y-Werte})$$

$v_{(j)}$ ist monoton steigend

Lorenzkurve der individuellen Nettovermögen in Deutschland
Gini-Koeffizienten: ca. 0.77 - 0.80



Gini-Koeffizient / Lorenz'sches Konzentrationsmaß

Der Gini-Koeffizient ist eine Maßzahl, die das Ausmaß der Konzentration beschreibt. Er ist definiert als

$$G = 2 \cdot F \quad ; \quad G \in [0, \frac{n-1}{n}]$$

wobei F die Fläche zwischen $y=x$ und der Lorenzkurve ist.

$$G = \frac{2 \cdot \sum_{i=1}^n i \cdot x_{(i)} - (n+1) \cdot \sum_{i=1}^n x_{(i)}}{n \cdot \sum_{i=1}^n x_{(i)}} \quad \text{oder} \quad G = 1 - \frac{1}{n} \sum_{i=1}^n (v_{(i-1)} + v_{(i)})$$

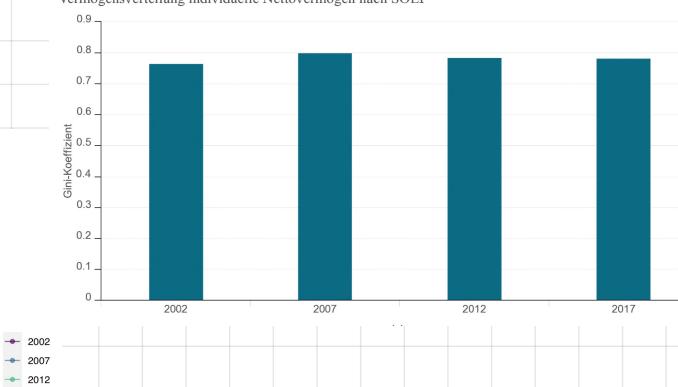
$$= \frac{n-1}{n} - \frac{2}{n} \cdot \sum_{i=1}^{n-1} v_{(i)}$$

$$\text{Normierter Gini-Koeffizient: } G^+ = \frac{n}{n-1} \cdot G \quad G^+ \in [0, 1]$$

$G^+ = 0$ bedeutet keine Konzentration (Gleichverteilung)

$G^+ = 1$ bedeutet volle Konzentration (Monopol)

Vermögensverteilung individuelle Nettovermögen nach SOEP



Herfindahl-Index

Seien x_1, \dots, x_n Daten mit $x_i \geq 0$. $p_i := \frac{x_i}{\sum_j x_j}$

Der Herfindahl-Index ist

$$H := \sum_{i=1}^n p_i^2 \in [\frac{1}{n}, 1]$$