Vorlesung: Statistik I für Studierende der Statistik, Mathematik & Informatik

Dozent: Fabian Scheipl
Material: H. Küchenhoff
LMU München

Deskriptive Statistik

"Data is merely the raw material of knowledge."
Charles Wheelan

Ziel: Beschreibung von Daten mit möglichst geringem Informationsverlust

- Eigenschaften und Strukturen sichtbar machen
- Graphisch und durch Kennwerte
- Eindimensional und mehrdimensional
- Zunächst keine Schlüsse auf die Grundgesamtheit oder allgemeine Phänomene (Deskription, nicht Inferenz)

Univariate deskriptive Statistik

Rohdaten und Datenmatrix

Daten liegen in der Regel als Datenmatrix bzw. Tabelle vor:

Organisationsprinzip: **tidy data** (Wickham, 2014):

- Zeilen entsprechen Untersuchungseinheiten
- Spalten entsprechen Merkmalen
- Elemente der Matrix sind die Merkmalsausprägungen
- Fragen mit Mehrfachnennungen als einzelne binäre Merkmale definieren

Hinweise zur korrekten Eingabe u.a. auf der Stablab-Homepage

Beispiel: Befragung von Redakteuren

Bitte füllen Sie diesen Fragebogen nur aus, wenn Sie Chefredakteur bzw. Redaktionsleiter einer Print-Zeitung sind

Sehr geehrter Teilnehmer,
zunächst haben wir einige allgemeine Fragen zur Organisation Ihrer Redaktion:

1. Die Redaktionen von Print-Zeitungen in Deutschland sind unterschiedlich groß. Wie viele Journalisten (festangestellte und freie) arbeiten in der Stammredaktion Ihrer Print-Tageszeitung?

_____ festangestellte Redakteure und _____ freie Mitarbeiter.

2. In jeder Redaktion gibt es verschiedene Positionen zu besetzen. Bitte geben Sie an, welche der folgenden Positionen es in Ihrer Print-Redaktion gibt und wie oft sie gegebenenfalls besetzt sind (also z.B. „2" wenn es zwei Chefs vom Dienst gibt).

Es gibt....

___ (Anzahl)	Chefredakteur(e)
___ (Anzahl)	Stellvertretende(n) Chefredakteur(e)
___ (Anzahl)	Chef(s) vom Dienst
___ (Anzahl)	Ressortleiter
___ (Anzahl)	Leitende(n) Redakteur(e)
___ (Anzahl)	weitere Position und zwar _____

3. Der Alltag von Journalisten wird durch verschiedene Tätigkeiten bestimmt. Bitte geben Sie an, wie intensiv die Print-Redakteure die folgenden Tätigkeiten im Redaktionsalltag ausüben.

	täglich	mehrmals pro Woche	einmal pro Woche	mehrmals pro Monat	einmal pro Monat	seltener als einmal pro Monat	nie
Verfassen eigener Artikel	☐	☐	☐	☐	☐	☐	☐
Redigieren von Agenturmeldungen/ Pressemitteilungen	☐	☐	☐	☐	☐	☐	☐
Redigieren von Beiträgen anderer Autoren	☐	☐	☐	☐	☐	☐	☐
Recherche vor Ort	☐	☐	☐	☐	☐	☐	☐
Recherche vom Schreibtisch aus	☐	☐	☐	☐	☐	☐	☐
Bearbeiten von Fotos	☐	☐	☐	☐	☐	☐	☐
Technische Produktion/ Layout der Beiträge	☐	☐	☐	☐	☐	☐	☐

Eindimensionale Häufigkeitsverteilung

- Ordnen der Daten nach einem Merkmal
- Auszählen der Häufigkeiten der einzelnen Merkmalsausprägungen
- Relative Häufigkeiten = Häufigkeit/Anzahl der Untersuchungseinheiten
- Kumulative Häufigkeiten bei ordinal oder metrisch skalierten Merkmalen sinnvoll:
  $F(x) :=$(Summe der relativen Häufigkeiten $\leq$ x) **empirische Verteilungsfunktion**

Häufigkeitsverteilung: Notation

Im Weiteren:

- $X, Y, \ldots$ Bezeichnung für **Merkmal**
- n **Untersuchungseinheiten**
- $x_i, i \in \{1, \ldots, n\}$: **beobachteter** Wert bzw. **Merkmalsausprägung** von X für i-te Beobachtung
- $x_1, \ldots, x_n$ **Rohdaten**, **Urliste**

Häufigkeiten I

$a_1 < a_2 < \ldots < a_k$, $k \leq n$ der Größe nach geordnete, *verschiedene* Werte der Urliste $x_1, \ldots, x_n$

Beispiel: Absolventenstudie

Für die Variable D "Ausrichtung der Diplomarbeit" ist die Urliste durch die folgende Tabelle gegeben:

Person i	1	2	3	4	5	6	7	8	9	10	11	12
Variable D	3	4	4	3	4	1	3	4	3	4	4	3
Person i	13	14	15	16	17	18	19	20	21	22	23	24
Variable D	2	3	4	3	4	4	2	3	4	3	4	2
Person i	25	26	27	28	29	30	31	32	33	34	35	36
Variable D	4	4	3	4	3	3	4	2	1	4	4	4

Häufigkeiten II

Häufigkeitstabelle für die Variable D "Ausrichtung der Diplomarbeit":

Ausprägung a_j	**absolute Häufigkeit** h_j	**relative Häufigkeit** f_j
1	2	$2/36 = 0.056$
2	4	$4/36 = 0.111$
3	12	$12/36 = 0.333$
4	18	$18/36 = 0.500$

Bemerkungen:

- Für Nominalskalen hat die Anordnung "$<$" keine inhaltliche Bedeutung.
- Bei kategorialen Merkmalen k = Anzahl der Kategorien
- Bei stetigen Merkmalen k oft nicht oder kaum kleiner als n.

Absolute und relative Häufigkeiten

$h(a_j) = h_j$ **absolute Häufigkeit** der Ausprägung a_j,

d.h. Anzahl der x_i aus $x_1, \ldots x_n$ mit $x_i = a_j$

$f(a_j) = f_j = h_j/n$ **relative Häufigkeit** von a_j

$h_1, \ldots, h_k$ **absolute Häufigkeitsverteilung**

$f_1, \ldots, f_k$ **relative Häufigkeitsverteilung**

Bemerkungen:

- Wenn statt der Urliste bereits die Ausprägungen $a_1, \ldots, a_k$ und ihre Häufigkeiten $f_1, \ldots, f_k$ bzw. $h_1, \ldots, h_k$ vorliegen, sprechen wir von **Häufigkeitsdaten**.
- **Klassenbildung**, **gruppierte Daten**: Bei metrischen, stetigen (oder quasi-stetigen) Merkmalen oft Gruppierung der Urliste durch Bildung geeigneter Klassen

Beispiel Nettomieten I

Wir greifen aus dem gesamten Datensatz des Münchner Mietspiegels die Wohnungen ohne zentrale Warmwasserversorgung (`zh0 == 1`) und mit einer Wohnfläche kleiner als $60m^2$ (`wfl < 60`) heraus.

Die folgende Urliste zeigt, bereits der Größe nach geordnet, die Nettomieten dieser $n = 61$ Wohnungen:

```
url <- "http://www.statistik.lmu.de/service/datenarchiv/miete/miete03.asc"
mietspiegel <- read.table(file = url, header = TRUE)
klein_und_kalt <- subset(mietspiegel, zh0 == 1 & wfl < 60)
sort(klein_und_kalt[, "nm"])
```

```
##  [1]  81.28  98.85 109.32 112.08 130.35 132.24 151.00 163.17 163.41 165.26
## [11] 172.23 177.36 180.40 181.98 183.09 195.72 203.75 213.01 220.81 224.61
## [21] 227.91 229.06 237.75 244.50 255.57 261.98 263.85 268.36 272.24 275.52
## [31] 279.04 279.53 281.21 287.36 304.22 304.23 309.00 311.87 314.09 329.68
## [41] 352.79 353.69 357.05 362.00 373.37 373.47 388.81 389.23 409.04 412.61
## [51] 430.06 463.40 479.46 482.96 487.16 501.07 505.38 532.56 538.07 562.43
## [61] 567.25
```

Alle Werte verschieden:
$\Rightarrow k = n$ und $\{x_1, \ldots, x_n\} = \{a_1, \ldots, a_k\}$; $f_j = \frac{1}{27} \forall j = 1, \ldots, 27.$

Beispiel Nettomieten II

Gruppiert man die Urliste in 6 Klassen mit gleicher Klassenbreite von 100€, so erhält man folgende Häufigkeitstabelle:

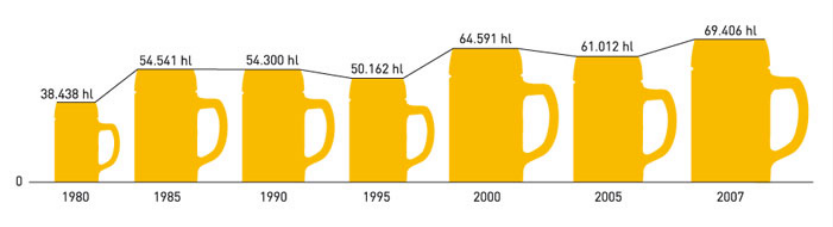
```
gruppierung <- seq(50, 650, by = 100)
klein_und_kalt[, "nm_gruppiert"] <- cut(klein_und_kalt[, "nm"], breaks = gruppierung)
table(klein_und_kalt[, "nm_gruppiert"])
```

```
##
##  (50,150] (150,250] (250,350] (350,450] (450,550] (550,650]
##         6        18        16        11         8         2
```

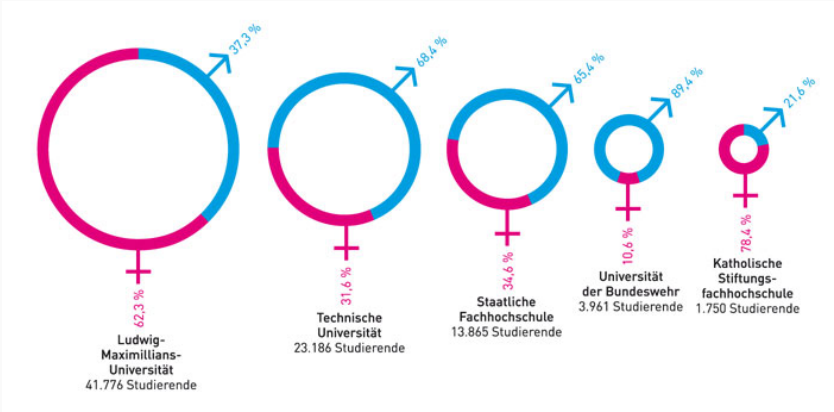
Klasse	absolute Häufigkeit	relative Häufigkeit
(50,150]	6	0.10
(150,250]	18	0.30
(250,350]	16	0.26
(350,450]	11	0.18
(450,550]	8	0.13
(550,650]	2	0.03

Grafische Darstellungen I

"Ein Bild sagt mehr als tausend Worte."

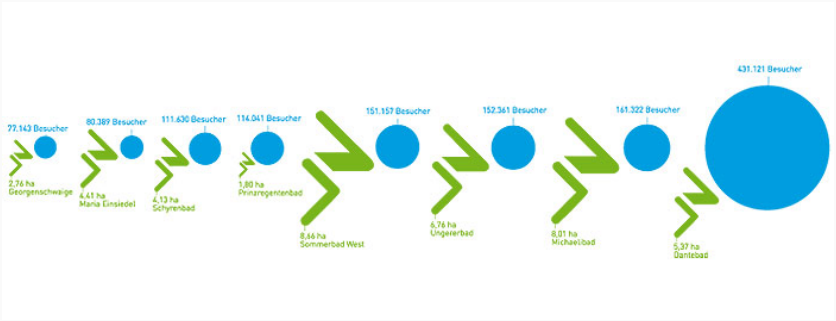


Grafische Darstellungen II



Grafische Darstellungen III

Grafische Darstellungen IV



Statistische Grafik

Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, of *statistics* and of *design*.
- Graphical excellence consists of complex ideas communicated with *clarity, precision and efficiency*.
- Graphical excellence is that which gives to the viewer the *greatest number of ideas* in the *shortest time* with the *least ink* in the *smallest space*.
- Graphical excellence is nearly always *multivariate*.
- And graphical excellence requires telling the *truth about the data*.

Tufte, E. (2001): The Visual Display of Information. Graphic Press 2nd ed.

Allgemeine Kriterien

- Wahl der Skala inkl. Bereich
- Wahl des Prinzips (Längentreue, Flächentreue)
- Einbringen von anderen Visualisierungen (Piktogramme etc.)
- Angemessene Wahl der Variablen
- Angemessene Wahl der Farben

Wahrnehmung von Grafiken

Experimente von Psychologen zeigen Hierarchie der korrekten Interpretation (Cleveland/McGill)

1. Abstände
2. Winkel
3. Flächen
4. Volumen
5. Farbton-Sattheit-Schwärzegrad

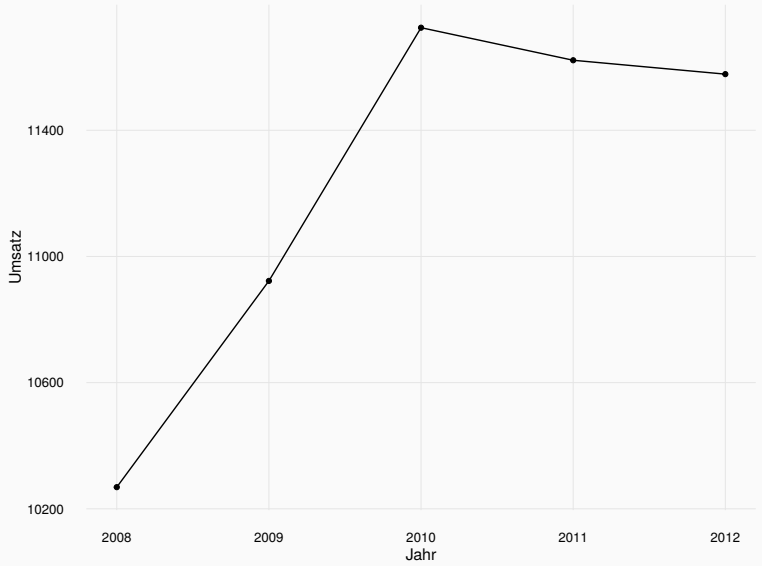
Da **Abstände** am besten wahrgenommen werden, sollten diese bevorzugt verwendet werden.

Cleveland, W.S., McGill, R. (1984): Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *JASA* 79(387), 531–554.

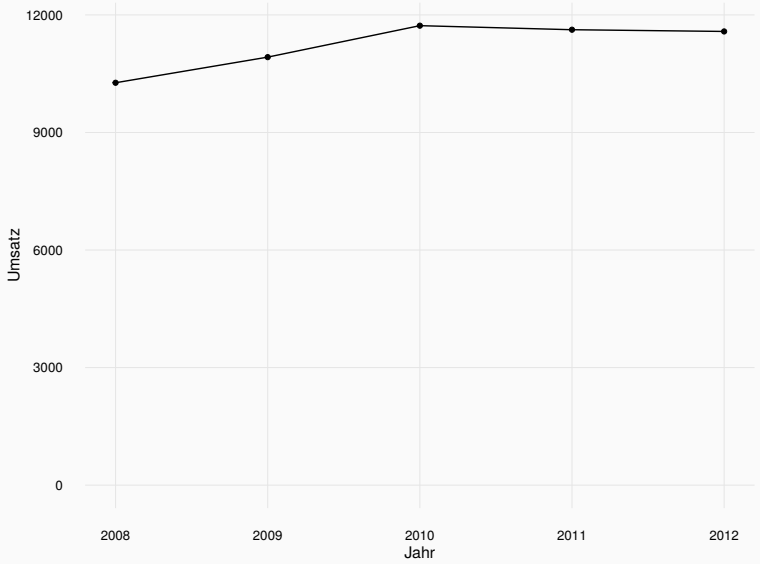
Typen von eindimensionalen Darstellungen

- Stab-, Balken- und Säulendiagramm
- Kreis (Torten)-Diagramm
- Histogramm

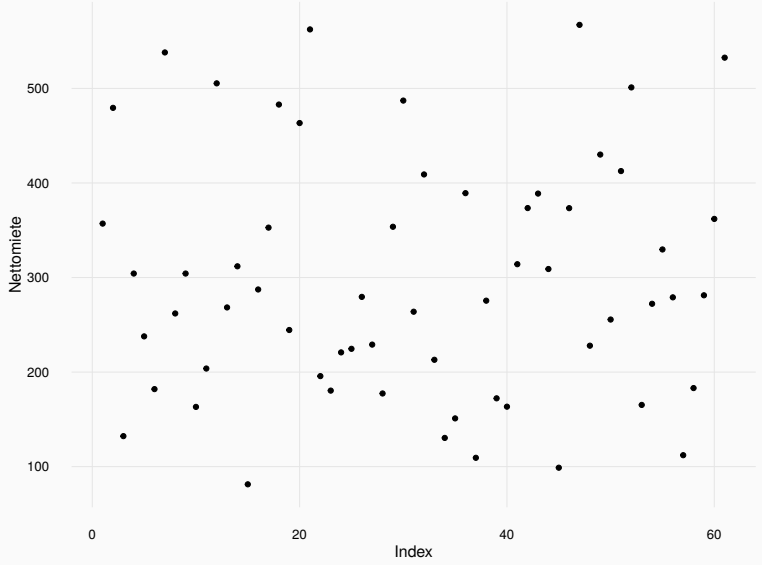
Beispiel: Liniendiagramm (??)



Beispiel: Liniendiagramm (!!)

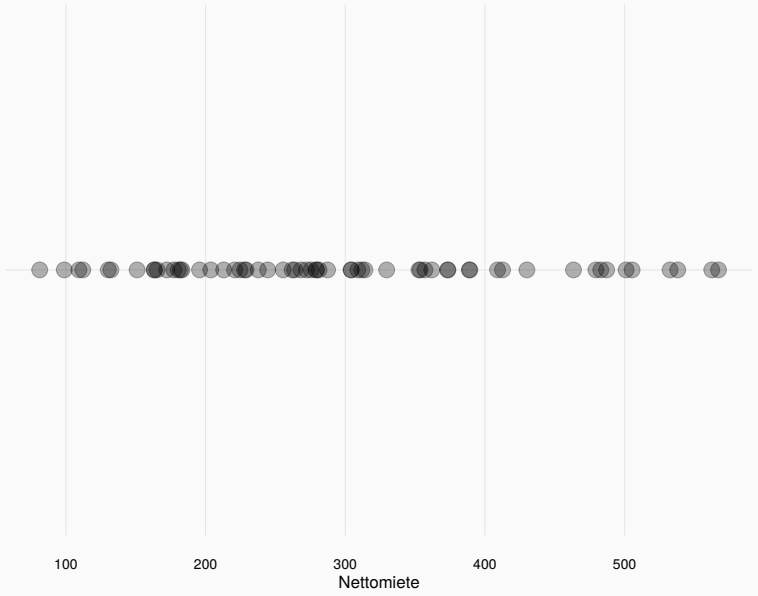


Beispiel: Streudiagramm (??)

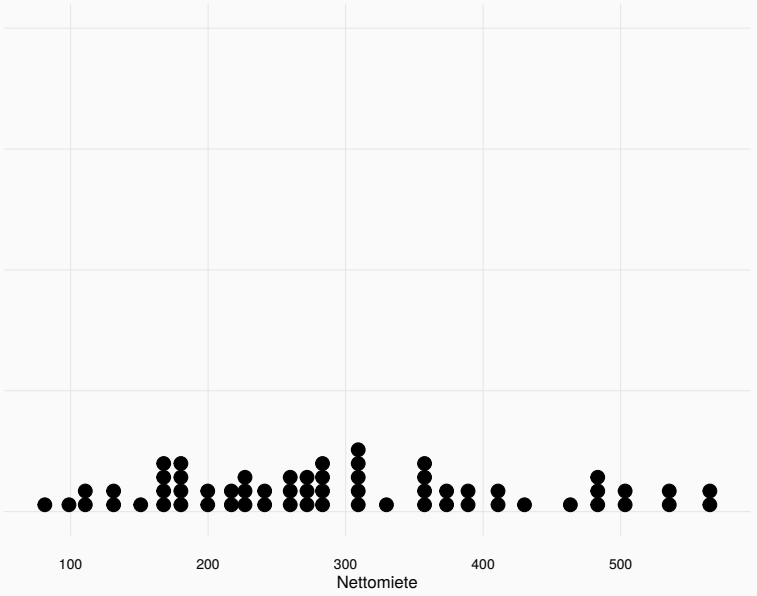


scatter plot

Beispiel: Dotplot (!?)



Beispiel: Gruppierter Dotplot



(bin width = 10)

Kreisdiagramm, Tortendiagramm

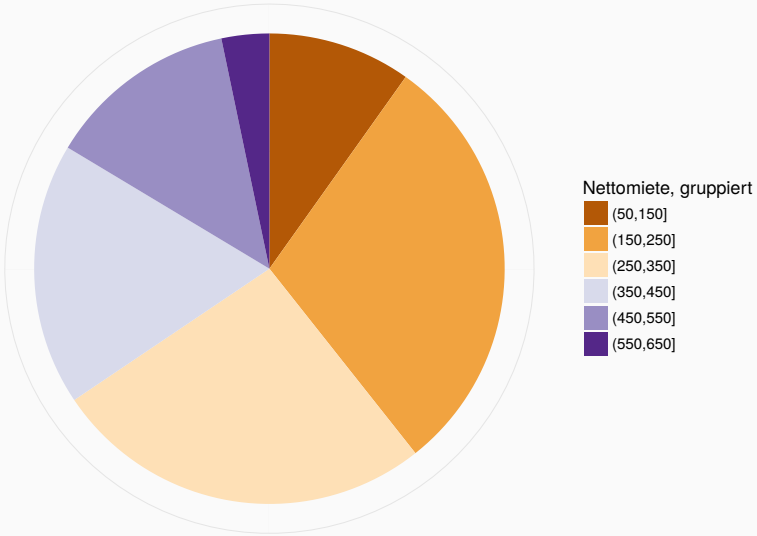
Darstellung der relativen (absoluten) Häufigkeiten als Anteile Fläche eines Kreises

Anwendung:

- Nominale Merkmale
- Ordinale Merkmale (Problem: Ordnung nicht korrekt wiedergegeben)
- Gruppierte Daten

pie chart

Tortendiagramm: Klein & Kalt



Stabdiagramm, Säulen- und Balkendiagramm

- **Stabdiagramm:** Trage über $a_1, \ldots, a_k$ jeweils einen zur x-Achse senkrechten Strich (Stab) mit Höhe $h_1, \ldots, h_k$ (oder $f_1, \ldots, f_k$) ab.
- **Säulendiagramm** wie Stabdiagramm, aber mit Rechtecken statt Strichen.
- **Balkendiagramm**: wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x-Achse.

Säulendiagramm

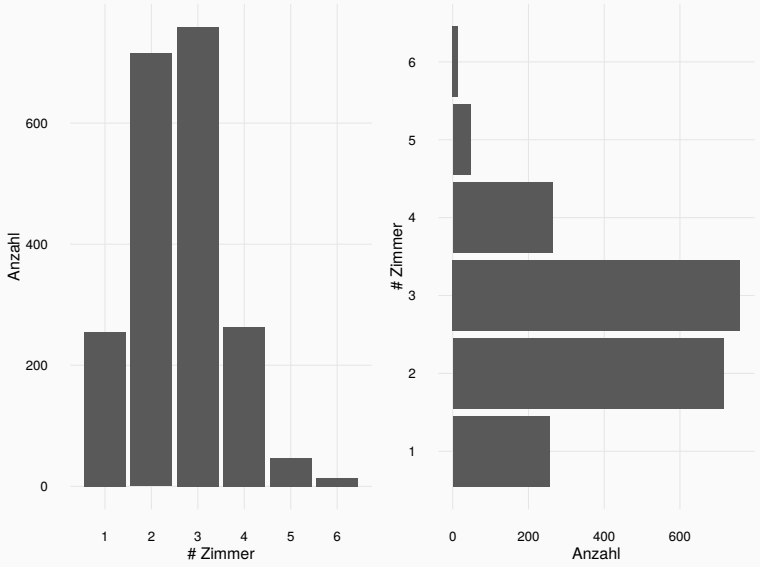
Darstellung der absoluten oder relativen Häufigkeiten als Höhen (Längen)

x-Achse: Ausprägungen des Merkmals y-Achse: absolute/ relative Häufigkeiten

Anwendungen:

- Ordinale Merkmale
- Metrische Merkmale mit wenigen Ausprägungen
- Nominale Merkmale (Problem: Ordnung nicht vorhanden)

Beispiel Mietspiegel: Säulendiagramm / Balkendiagramm



Stapeldiagramm

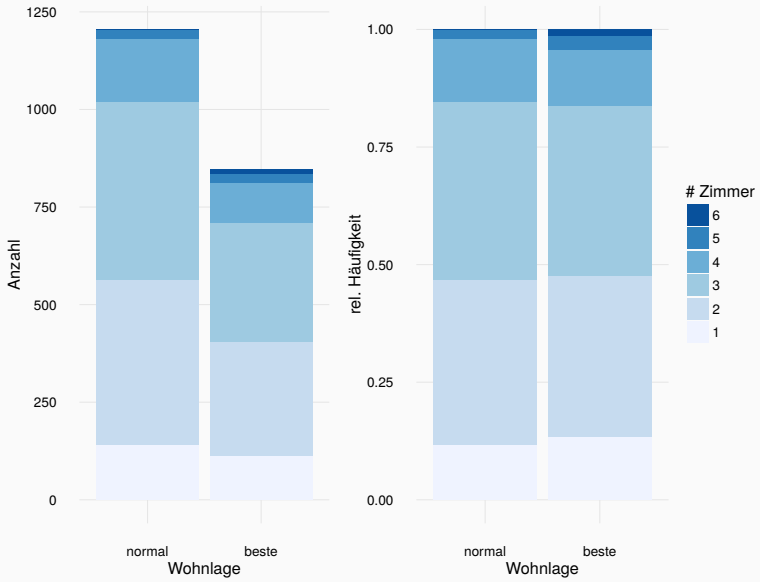
Darstellen der absoluten oder relativen Häufigkeiten als Länge. Die Abschnitte werden übereinander in verschiedenen Farben gestapelt.

Anwendungen:

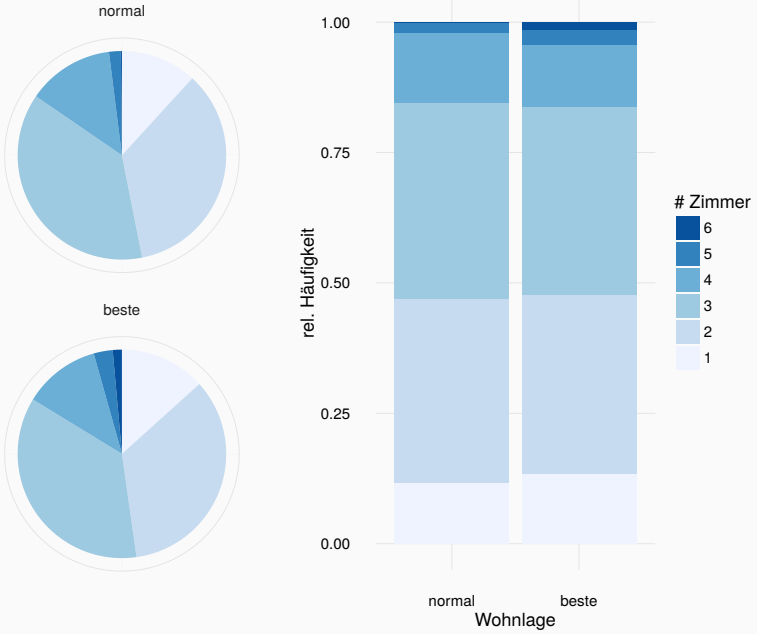
- Ordinale Daten
- Gruppierte Daten
- Metrische Daten mit wenigen Ausprägungen

Besonders geeignet für den Vergleich verschiedener Gruppen durch nebeneinander liegende Stapel. Zu beachten ist dann die Unterscheidung: relative Häufigkeit $\leftrightarrow$ absolute Häufigkeit

Beispiel Mietspiegel:: Stapeldiagramme



Beispiel Mietspiegel: Vergleich mit Kreisdiagramm



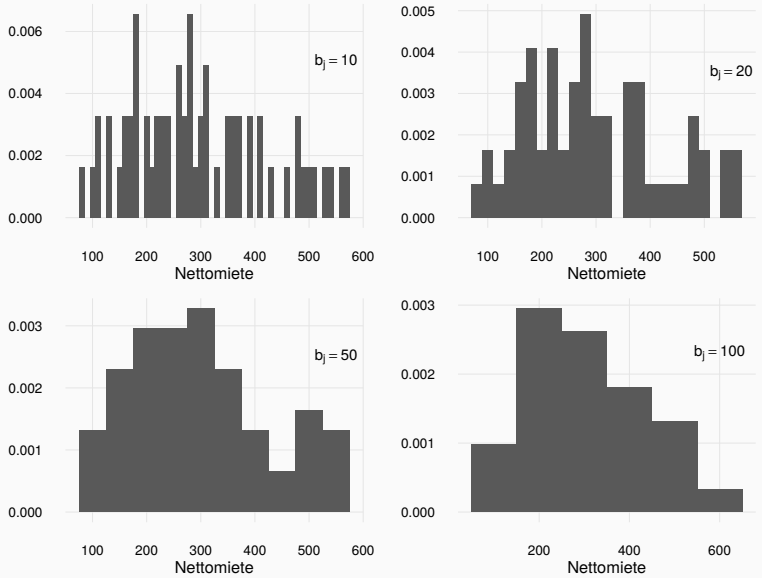
Das Histogramm

Darstellung der relativen Häufigkeiten durch Flächen (Prinzip der **Flächentreue**)

Vorgehen:

1. Aufteilung in Klassen (falls die Daten noch nicht gruppiert sind)
2. Bestimmung der relativen Häufigkeiten $f_j = \frac{n_j}{n}$
3. Bestimmung der Höhen h_j, so dass gilt $b_j \cdot h_j = f_j$ wobei b_j: Breite der Klasse j.

Beispiel: Nettomiete Klein & Kalt



Histogramm

- Anwendung bei metrischen Daten
- Beachte: Abhängigkeit von der Breite
- Klassen inhaltlich vorgeben, verschiedene Varianten ansehen
- Vorsicht bei Rändern

Stamm-Blätter-Diagramm

Spezielles Histogramm / Balkendiagramm mit

- Klassen nach Dezimalsystem
- Einzeldaten reproduzierbar

stem-and-leaf plot

Beispiel: Nettomiete Klein & Kalt

```
stem(klein_und_kalt[, "nm"], scale = 0.5)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 8
##   1 | 01133566778888
##   2 | 00122334466677888889
##   3 | 00111355667799
##   4 | 1136889
##   5 | 013467
```

```
sort(round(klein_und_kalt$nm/10) * 10)
```

```
##  [1]  80 100 110 110 130 130 150 160 160 170 170 180 180 180 180 200 200
## [18] 210 220 220 230 230 240 240 260 260 260 270 270 280 280 280 280 290
## [35] 300 300 310 310 310 330 350 350 360 360 370 370 390 390 410 410 430
## [52] 460 480 480 490 500 510 530 540 560 570
```


Empirische Verteilungsfunktion

Häufigkeitsfunktion:

$$H(x) := (\text{Anzahl der Werte} <= x)$$

Verteilungsfunktion:

$$F(x) = H(x)/n = (\text{Anteil der Werte } x_i \text{ mit } x_i \leq x)$$

bzw.

$$F(x) = f(a_1) + \ldots + f(a_j) = \sum_{i:a_i \leq x} f_i,$$

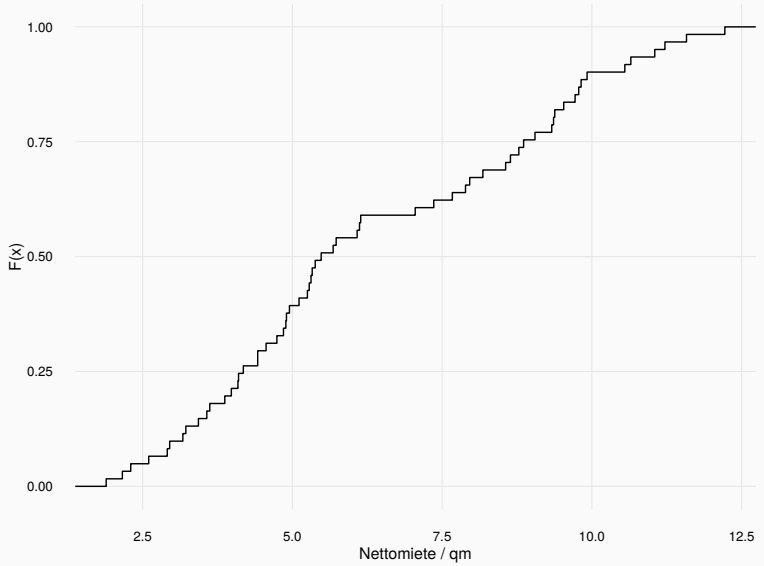
wobei $a_j \leq x$ und $a_{j+1} > x$ ist.

ECDF (empirical cumulative distribution function)

Eigenschaften von $F(x)$

- monoton wachsende Treppenfunktionen mit Sprüngen an den Ausprägungen $a_1, \ldots, a_k$
- Sprunghöhen: $f_1, \ldots, f_k$
- rechtsseitig stetig
- $F(x) = 0$ für $x < a_1$, $F(x) = 1$ für $x \geq a_k$

Beispiel: F(Quadratmetermiete) für Klein & Kalt



Eindimensionale statistische Kennwerte

Lagemaßzahlen

- Wo liegt die Masse der Daten?
- Wo liegt die Mehrzahl der Daten?
- Wo liegt die Mitte der Daten?
- Welche Merkmalsausprägung ist typisch für die Häufigkeitsverteilung?

Statistische Kennwerte

Streumaßzahlen

- Über welchen Bereich erstrecken sich die Daten?
- Wie groß ist die Schwankung der Ausprägungen?

Lagemaß: Modus

Definition: **Häufigster Wert**

Eigenschaften:

- oft nicht eindeutig
- nur bei gruppierten Daten oder bei Merkmalen mit wenigen Ausprägungen sinnvoll
- stabil bei allen *eindeutigen* Transformationen
- geeignet für alle Skalenniveaus

Lagemaß: Median

Definition: Der **Median** ($\tilde{x}_{\text{med}}$) ist der Wert für den gilt

- mindestens 50% der Daten sind kleiner oder gleich $\tilde{x}_{\text{med}}$,
- mindestens 50% der Daten sind größer oder gleich $\tilde{x}_{\text{med}}$.

$$\tilde{x}_{\text{med}} = \begin{cases} x_{(k)} & \text{falls } k = \frac{n+1}{2} \text{ ganze Zahl} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{falls } k = \frac{n}{2} \text{ ganze Zahl} \end{cases}$$

- $x_{(1)}, \ldots, x_{(n)}$ sind **geordnete Werte**
- Alternative Definition: $\tilde{x}_{\text{med}} \in [x_{(k)}, x_{(k+1)}]$ falls $k = \frac{n}{2}$ ganze Zahl.

Eigenschaften des Medians

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

Lagemaß: Quantil

Definition: Das p-**Quantil** ist der Wert $\tilde{x}_p$ für den gilt

- mindestens Anteil p der Daten sind kleiner oder gleich $\tilde{x}_p$,
- mindestens Anteil $1 - p$ der Daten sind größer oder gleich $\tilde{x}_p$.

$$\tilde{x}_p = \begin{cases} x_{(k)} & \text{falls } np \text{ keine ganze Zahl und } k \text{ kleinste Zahl} > np \\ \in \left[x_{(k)}; x_{(k+1)}\right] & \text{falls } k = np \text{ ganze Zahl} \end{cases}$$

- Es gibt weitere Definitionen von Quantilen (in R 9 Typen!), die sich aber in der Praxis kaum unterschieden.
- p-Quantil = $(100 \cdot p)$-Perzentil
- Der Median ist das 0.5-Quantil bzw. 50%-Perzentil

Boxplot

Einfacher **Boxplot**:

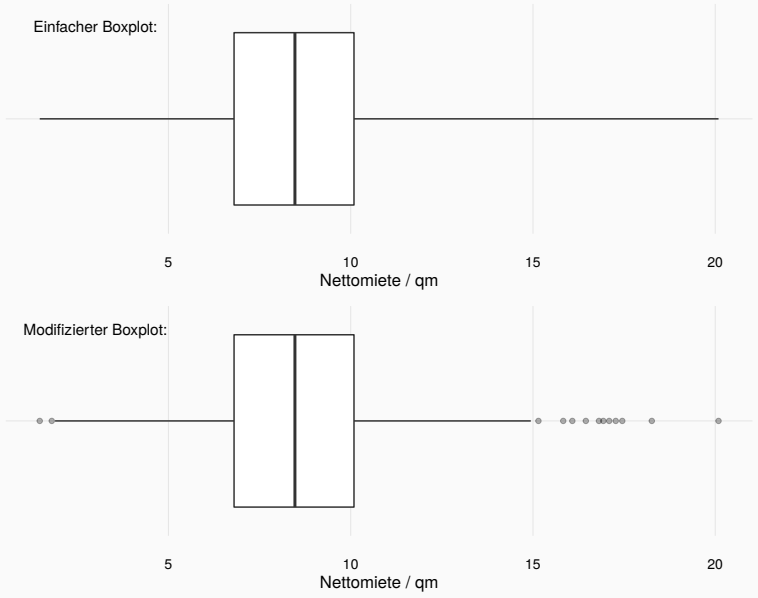
- $\tilde{x}_{0.25}$ = Anfang der Schachtel (Box) (= unteres **Quartil**)
- $\tilde{x}_{0.75}$ = Ende der Schachtel (= oberes **Quartil**)
- d_Q = Länge der Schachtel (= **Inter-Quartile-Range** (IQR))
- Der **Median** wird durch den Strich in der Box markiert
- Zwei Linien (*whiskers*) außerhalb der Box gehen bis zu x_{min} und x_{max}.

Boxplot

Modifizierter **Boxplot**:

- *whiskers* werden nur bis zu x_{min} bzw. x_{max} gezogen, falls x_{min} und x_{max} innerhalb des Bereichs $[z_u, z_o]$ der Zäune liegen.
  Üblicherweise: $z_u = \tilde{x}_{0.25} - 1.5d_Q$, $z_o = \tilde{x}_{0.75} + 1.5d_Q$
- Ansonsten gehen die Linien nur bis zum kleinsten bzw. größten Wert innerhalb der Zäune, die außerhalb liegenden Werte werden individuell eingezeichnet.

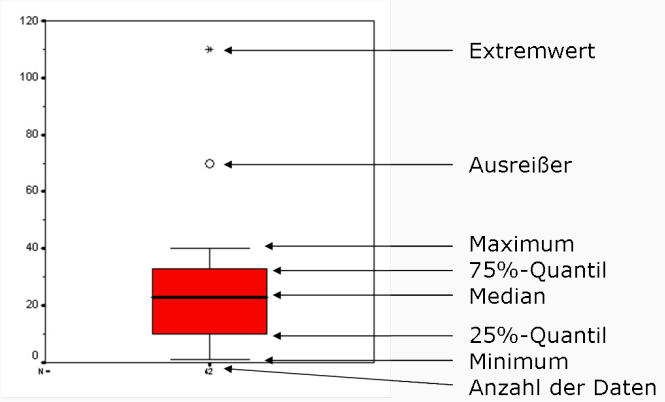
Boxplot: Beispiel Quadratmetermiete Mietspiegel



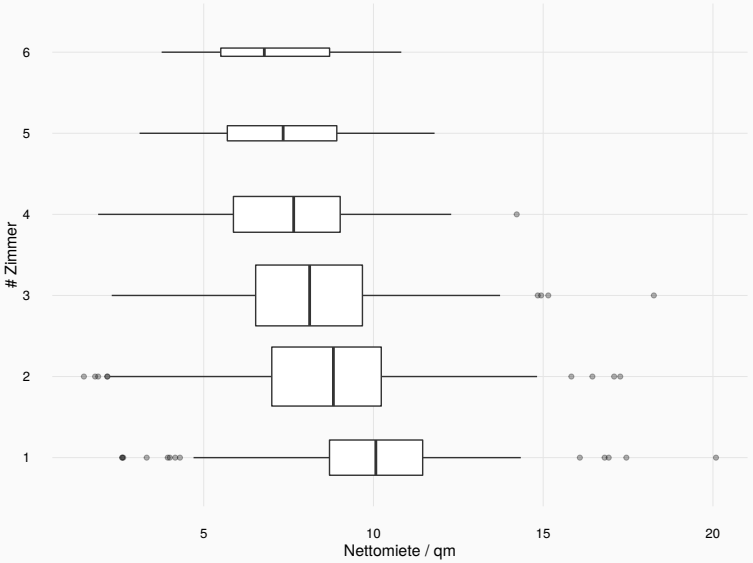
Boxplot

- Eindimensionale Darstellung auf der zugehörigen Skala
- Visualisieren der 5-Punkte-Zusammenfassung (Minimum, 25%-,50%-,75%-Perzentile, Maximum)

SPSS-Output:



Gruppierter Boxplot:



Boxplot: Vor- und Nachteile

+:

- kompakt
- geeignet für Vergleiche
- **Ausreißer** sichtbar
- **Schiefe** sichtbar

−:

- gegen Intuition (Viel Farbe – wenig Daten) da Ausreißer sehr prominent
- **Multimodale** Verteilungen nicht sichtbar
- (Breite redundant)

Der Mittelwert (arithmetisches Mittel)

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

- bekanntestes Lagemaß
- instabil gegen extreme Werte
- geeignet für intervallskalierte Daten
- "Durchschnitt"

Mittelwert bei gruppierten Daten

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\ &= \frac{1}{n}\sum_{j=1}^{k} h_j a_j\end{aligned}$$

h_j : Häufigkeit von a_j

Das geometrische Mittel

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- arithmetisches Mittel auf der log-Skala

$$x_g = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right)$$

- nur geeignet für positive Werte
- geeignet für intervallskalierte Daten
- Anwendung: Durchschnitt von Änderungsraten, z.B. durchschnittliche Verzinsung

Das harmonische Mittel

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

Das harmonische Mittel entspricht dem Mittel durch Transformation

$$t \to \frac{1}{t}: \qquad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Beispiel: $x_1, \ldots, x_n$ Geschwindigkeiten, mit denen jeweils Wegstrecke L zurückgelegt wird

Gesamt-Geschwindigkeit:

$$\frac{L \cdot n}{\frac{L}{x_1} + \ldots + \frac{L}{x_n}} = \bar{x}_H$$

Allgemeine Transformation des Mittelwerts I

Lineare Transformation:

$$\begin{aligned} g(t) &= a + bt \\ y_i &= a + bx_i \Rightarrow \bar{y} = a + b\bar{x} \end{aligned}$$

d.h.

$$\begin{aligned} \overline{a + bx} &= a + b\bar{x} \\ \overline{g(x)} &= g(\bar{x}) \end{aligned}$$

Allgemeine Transformation:

Generell ist $\overline{g(x)} \neq g(\bar{x})$

Allgemeine Transformation des Mittelwerts II

Für **konvexe** Funktionen g gilt:

$$\begin{aligned} g(\bar{x}) &\leq \overline{g(x)} \\ g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &\leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \end{aligned}$$

(Jensen-Ungleichung)

g konvex: $\Leftrightarrow$ $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$
$\forall \lambda \in [0,1], \quad x, y \in D_g$

Beispiel:
$\bar{x}^2 \leq \overline{x^2}$

Vergleich I

Es gilt allgemein für positive x_i:

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

Beweis:

1. Zeige $\bar{x}_G \leq \bar{x}$:

$g : t \to \log(t)$ konkav, da $g''(t) = -\frac{1}{t^2} < 0$
$\Rightarrow \log(\bar{x}) \geq \overline{\log(x)}$
$\Rightarrow \bar{x} \geq \exp\left(\overline{log(x)}\right) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right)$
$= \left(\prod_{i=1}^{n}\exp\left(\log(x_i)\right)\right)^{\frac{1}{n}} = \bar{x}_G$

Vergleich II

2. Zeige $\bar{x}_H \leq \bar{x}_G$:

$g_2 : t \to \frac{1}{\exp(t)}$ ist konvex, da $g_2''(t) = \frac{1}{\exp(t)} \geq 0$
Benutze transformierte Daten $\log(x_1), \ldots, \log(x_n)$

$$g_2\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\left(\exp(\log(x_i))^{-1}\right)$$

$$\Rightarrow \frac{1}{\sqrt[n]{\prod_{i=1}^{n}x_i}} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

$$\Rightarrow \underbrace{\sqrt[n]{\prod_{i=1}^{n}x_i}}_{\bar{x}_G} \geq \underbrace{\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}}_{\bar{x}_H}$$

Getrimmtes Mittel

Um die Ausreißerempfindlichkeit von $\bar{x}$ abzuschwächen definiert man

$$\bar{x}_\alpha = \frac{1}{n-2r}\sum_{i=r+1}^{n-r}x_{(i)}$$

- $x_{(i)}$: geordnete x-Werte
- r ist die größte ganze Zahl mit $r \leq n\alpha$

Es wird also der Anteil α der extremsten Werte abgeschnitten:
α-getrimmtes Mittel

Winsorisiertes Mittel (**gestutztes** Mittel):
Der Anteil α der extremsten Werte wird **durch das entsprechende Quantil ersetzt**.

Maße für die Streuung

- Spannweite
- Interquartilsabstand
- Standardabweichung und Varianz
- Variationskoeffizient

Die Spannweite (Range)

Definition:

$$q = x_{\max} - x_{\min}$$

- Bereich in dem die Daten liegen
- Wichtig für Datenkontrolle: Plausibilität, Eingabe-/Codierungsfehler, etc.

Der Quartilsabstand

Definition:

$$d_Q = x_{0.75} - x_{0.25}$$

- Größe des Bereichs in dem die "mittlere Hälfte" der Daten liegt
  - Länge der Box des Boxplots
- Bei ordinal skalierten Daten Angabe von $x_{0.75}$ und $x_{0.25}$:
  - Zentraler 50%-Bereich
- Robust gegen Ausreißer

Standardabweichung und Varianz

Definition:

$$\begin{aligned} \textbf{Varianz } S_x^2 &:= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ \textbf{Standardabweichung } S_x &:= \sqrt{S_x^2} \end{aligned}$$

- S_x = "Mittlere Abweichung vom Mittelwert"
- Mindestens Intervallskala
- empfindlich gegen Ausreißer
- Verwende $\widetilde{S}_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ für Vollerhebungen, Division durch $n-1$ nur bei Stichproben sinnvoll

Transformationsregel

$y_i = a + bx_i$

$$\begin{aligned} \Rightarrow \widetilde{S}_y^2 &= b^2\widetilde{S}_x^2 \\ \widetilde{S}_y &= |b|\widetilde{S}_x \end{aligned}$$

(Analog für S_x, S_y)

Varianz und Standardabweichung sind stabil mit linearen Transformationen verträglich.

Verschiebungssatz:

Für jedes $c \in \mathbb{R}$ gilt:

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

$$\begin{aligned} c = 0 \Rightarrow \widetilde{S}_x^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \\ \widetilde{S}_x^2 &= \overline{x^2} - \bar{x}^2 \end{aligned}$$

Beachte:
Verschiebungssatz für numerische Berechnung mit Computer **nicht geeignet**.

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

$$x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, \ldots, x_{n_r}$$

Schichtmittelwerte:

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i,\ \bar{x}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} x_i,\ \text{usw.}$$

Schichtvarianzen:

$$\widetilde{S}_{x1}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2,$$
$$\widetilde{S}_{x2}^2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(x_i - \bar{x}_2)^2 = \frac{1}{n_1}\sum_{i=n_1+1}^{n_1+n_2}(x_i - \bar{x}_2)^2,\ \text{usw.}$$

Streuungszerlegung II

Dann gilt:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{j=1}^{r} n_j \bar{x}_j \\ \widetilde{S}_x^2 &= \frac{1}{n}\sum_{j=1}^{r} n_j \widetilde{S}_{xj}^2 + \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x})^2 \end{aligned}$$

$$\text{Gesamtstreuung} = \begin{matrix}\text{Streuung}\\ \text{innerhalb}\\ \text{der Schicht}\end{matrix} + \begin{matrix}\text{Streuung}\\ \text{zwischen}\\ \text{den Schichten}\end{matrix}$$

Variationskoeffizient

Das Verhältnis von Standardabweichung und Mittelwert ist gegeben durch

$$v = \frac{\widetilde{S}_x}{\bar{x}} \text{ mit } \bar{x} > 0$$

Der Variationskoeffizient hat keine Einheit und ist skalenunabhängig.

Er ist eine Maßzahl für die relative Schwankung um den Mittelwert.

Mittlere absolute Abweichung (MAD)

Die mittlere absolute Abweichung ist definiert als

$$\text{MAD}_x = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{MedAD}_x := \textit{median}(|x_i - x_{med}|)$$

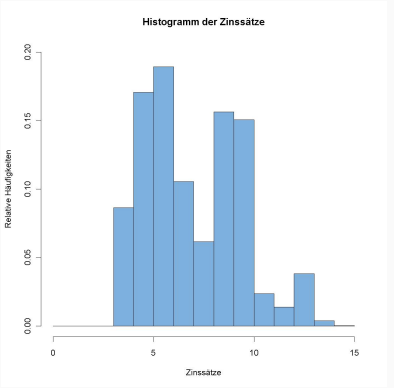
Wegen der Jensen-Ungleichung gilt: $\text{MAD}_x \leq \widetilde{S}$

$\text{MAD}_x, \text{MedAD}_x$

- nicht so "schöne" theoretische Eigenschaften wie S_x,
- klarer interpretierbar als S_x
- weniger Ausreißer-empfindlich

Uni- und multimodale Verteilungen

unimodal = eingipflig, **multimodal** = mehrgipflig

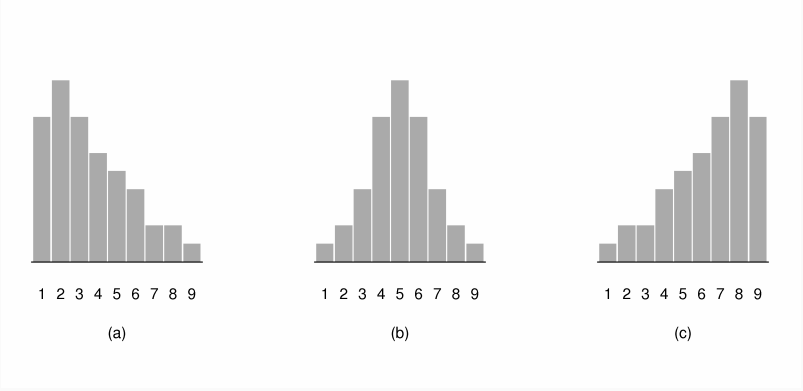


Das Histogramm der Zinssätze zeigt eine bimodale Verteilung.

Symmetrie und Schiefe I

symmetrisch	$\Leftrightarrow$	Rechte und linke Häfte der Verteilung sind annähernd zueinander spiegelbildlich
linkssteil (rechtsschief)	$\Leftrightarrow$	Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab
rechtssteil (linksschief)	$\Leftrightarrow$	Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab

Symmetrie und Schiefe II



Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)

Lageregeln

- Symmetrische und unimodale Verteilung: $\bar{x} \approx x_{med} \approx x_{mod}$
- Linkssteile Verteilung: $\bar{x} > x_{med} > x_{mod}$
- Rechtssteile Verteilung: $\bar{x} < x_{med} < x_{mod}$
- Bei gruppierten Daten: Auch für Histogramme gültig

Beachte:

Form der Verteilung bleibt bei linearen Transformationen gleich. Änderung bei nichtlinearen Transformationen.

Maßzahlen für die Schiefe I

Quantilskoeffizient:

$$g_p = \frac{(x_{1-p} - x_{med}) - (x_{med} - x_p)}{x_{1-p} - x_p}$$

$p = 0.25$ **Quartilskoeffizient**

Werte des Quantilskoeffizienten:

$g_p = 0$ für symmetrische Verteilungen
$g_p > 0$ für linkssteile Verteilungen
$g_p < 0$ für rechtssteile Verteilungen

Maßzahlen für die Schiefe II

Momentenkoeffizient der Schiefe:

$$g_m = \frac{m_3}{\tilde{s}^3} \quad \text{mit} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Werte des Momentenkoeffizienten:

$g_m = 0$ für symmetrische Verteilungen
$g_m > 0$ für linkssteile Verteilungen
$g_m < 0$ für rechtssteile Verteilungen

Dichtefunktion I

Histogramm: - Anteil der Beob. an den Daten = Fläche unter der Kurve - Histogramm ist stückweise konstante Funktion - Problematisch: Abhängigkeit von der Wahl der Klassengrenzen - Ersetze Histogramm durch glatte Funktion f

Dichtefunktion II

Eine positive stetige Funktion heißt **Dichte**(-funktion) (*density*), wenn - $f(x) \geq 0$ und - $\int_{-\infty}^{\infty} f(x)dx = 1$

Die Fläche unter der Dichte soll in etwa den relativen Häufigkeiten entsprechen, d.h.

$$\int_a^b f(x)dx \approx \frac{1}{n}\#\{x_i | a < x_i \leq b\}$$

Dichte und Verteilung

Verteilung, Verteilungsfunktion (*distribution*):

$$F(x_0) = \frac{1}{n}\#\{x_i | x_i \leq x_0\} \approx \int_{-\infty}^{x_0} f(x)dx$$

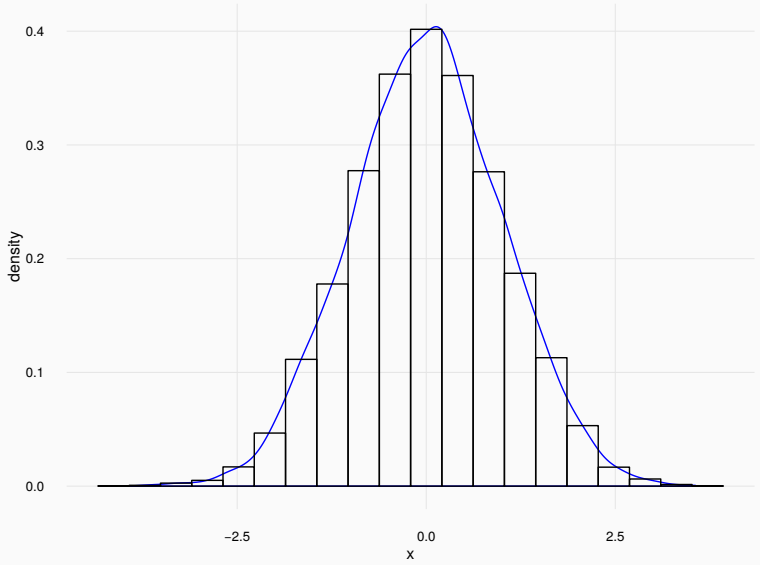
$$\hat{F}(x_0) = \int_{-\infty}^{x_0} \hat{f}(x)dx$$

Quantile:
Für $0 < p < 1$ ist das p-Quantil x_p der Wert auf der x-Achse, für den gilt:

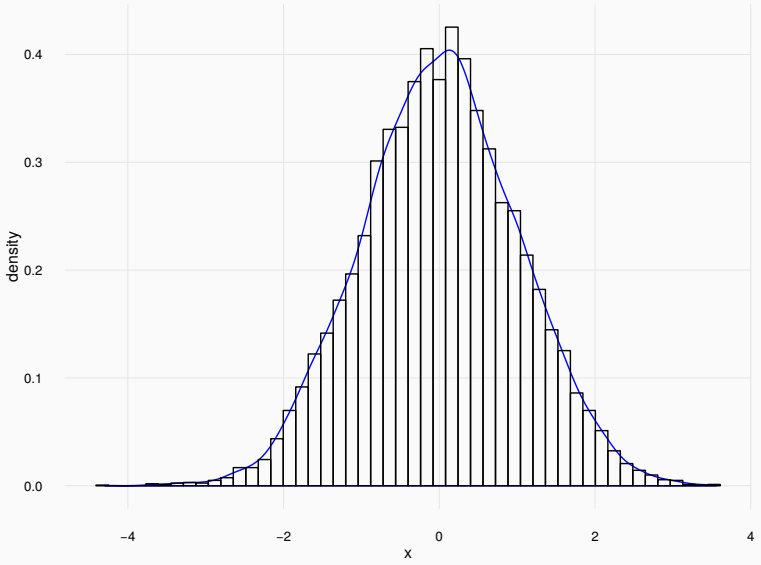
$$\int_{-\infty}^{x_p} f(x)dx = p$$

Der Median teilt die Fläche unter der Dichte in 2 gleich große Teile.

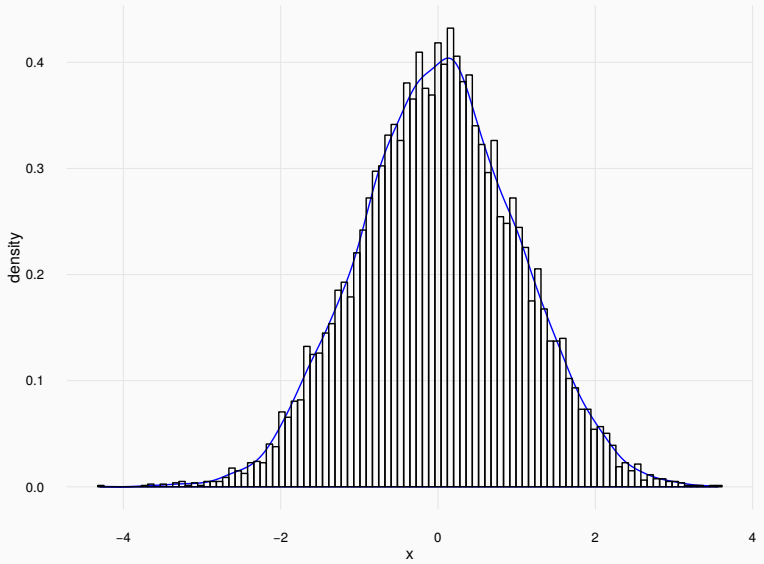
Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Berechnung von Dichte-Kurven

$$\hat{f}(x) = \frac{\frac{1}{n}\#\{x_i | x_i \in [x-h, x+h)\}}{2h}$$

$\Rightarrow$ "'Gleitendes Histogramm"'

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

mit $K(u) = \begin{cases} \frac{1}{2} \text{ für } -1 \leq u < 1 \\ 0 \text{ sonst} \end{cases}$

K : **Kernfunktion**

Kern-Dichteschätzer

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

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

heißt Kern-Dichteschätzer

Kerne:

Epanechnikov-Kern	$K(u) = \frac{3}{4}(1-u^2)$	für $-1 \leq u < 1$, 0 sonst.
Bisquare-Kern	$K(u) = \frac{15}{16}(1-u^2)^2$	für $-1 \leq u < 1$, 0 sonst.
Gauß-Kern	$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right)$	für $u \in \mathbb{R}$

Bemerkungen zur Dichteschätzung

- Abhängigkeit von der Bandweite $h \rightarrow$ Verfahren zur Bestimmung von h aus den Daten
- Abhängigkeit von der Wahl des Kerns eher unbedeutend
- Kerndichteschätzungen sind insbesondere bei größeren Datenmengen Histogrammen vorzuziehen

Normalverteilung

Für $x \in \mathbb{R}$ heißt

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Normalverteilungsdichte mit Mittelwert μ und Standardabweichung σ

Für $\mu = 0$ und $\sigma = 1$ erhält man

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}x^2\right)$$

Quantile I

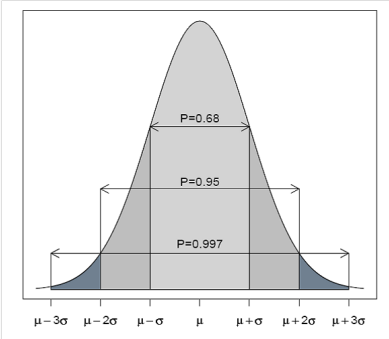
Quantile der Standardnormalverteilung:

p	50%	75%	90%	95%	97.5%	99%
x_p	0.0 (Median)	0.67	1.28	1.64	1.96	2.33

Quantile II

68-95-99.7-Prozent-Regel:

68%	der Beob. liegen im Interv.	$\mu \pm \sigma$
95%	der Beob. liegen im Interv.	$\mu \pm 2\sigma$
99.7%	der Beob. liegen im Interv.	$\mu \pm 3\sigma$

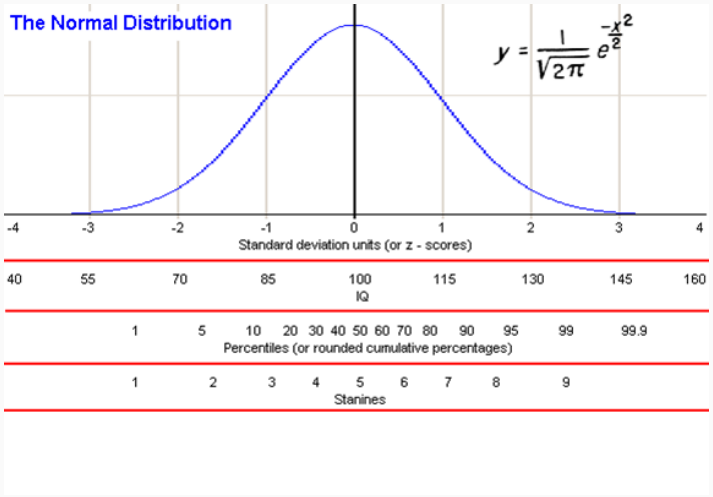


Normalverteilung in der Psychologie

Skalen werden so konstruiert, dass die Verteilung in der Population einer Normalverteilung genügt.

IQ :	Mittelwert = 100, Standardabweichung 15
T- Werte:	Mittelwert 50, Standardabweichung 10
Sta(ndard)nine	Mittelwert 5 Standardabweichung 2

Normalverteilung in der Psychologie II



Strategie zur Skalenbildung

- Ziehe grosse Stichprobe aus der Population
- Ordne Ergebnisse
- Zuordnung von Stanine

Schema:

4%	7%	12%	17%	20%	17%	12%	7%	4%
1	2	3	4	5	6	7	8	9

Strategie zur Skalenbildung II

- Ziehe grosse Stichprobe aus der Population
- **Standardisierung** (**Z- Werte**):

$$z_i = (x_i - \bar{x})/S_x$$

- Reskalieren durch Multiplikation mit gewünschter Standardabweichung und Addition des gewünschten Mittelwertes.

Normalverteilung in der technischen Statistik

- Größen in der Produktion (Längen etc.)
- Messfehler
- Größen nach geeigneter Transformation
- 6-Sigma

Überprüfung der Annahme der Normalverteilung

Fragestellung: Passen die Daten zu einer Normalverteilung?

1. Vergleiche Histogramm, Dichteschätzer mit NV-Dichte $(\mu = \bar{x}, \sigma = S)$
2. Vergleiche Verteilungsfunktion, d.h. empirische Verteilungsfunktion $F(x)$ mit $\Phi(t) = \int_{-\infty}^{t} f(\mu, \sigma, t)dt$
3. Prüfe Schiefe $= 0$
4. **Q-Q-Plot**

Normal-Quantil-Plot

Sei $x_{(1)}, \ldots, x_{(n)}$ die geordneten Daten. Für $i = 1, \ldots, n$ werden die $(i - 0.5)/n$-Quantile $z_{(i)}$ der Standardnormalverteilung berechnet.

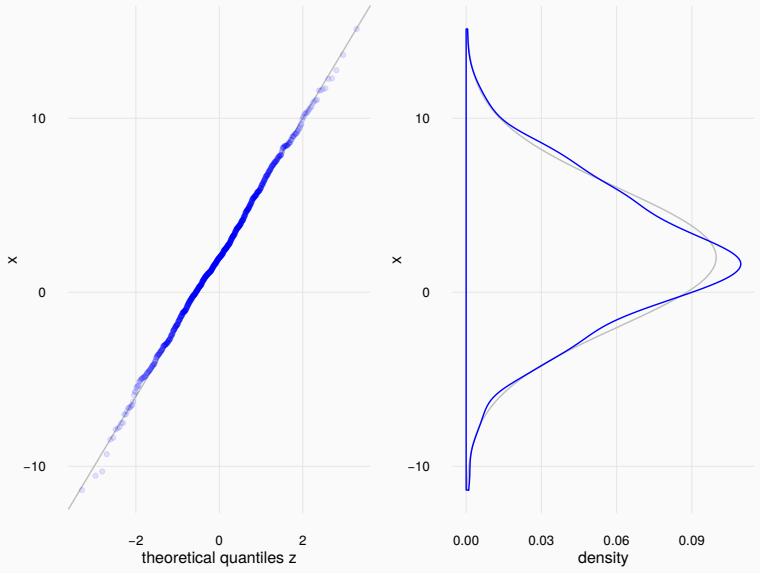
Der **Normal-Quantil-Plot** (Normal-Q-Q-Plot) besteht aus den Punkten

$$(z_{(1)}, x_{(1)}), \ldots, (z_{(n)}, x_{(n)})$$

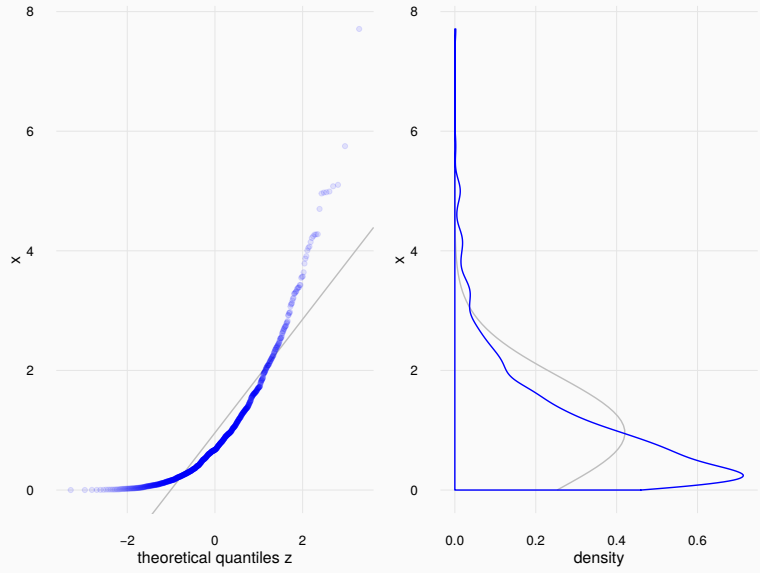
im $z - x-$Koordinatensystem.

Liegen die Punkte auf einer Geraden, so passt die Normalverteilung gut zu den Daten

Q-Q-Plot I



Q-Q-Plot II



Konzentrationsmaße

Motivation:

Existiert eine Menge, die auf viele Individuen verteilt ist, kann es hilfreich sein zu wissen, wie diese Menge verteilt ist.

Beispiele:

- Vermögensverteilung in einem Staat
- Marktanteile von Firmen in einem Segment

Lorenzkurve

Grundidee:

Es sollen folgende Aussagen grafisch dargestellt werden:

- Die "Ärmsten"/"Kleinsten" x% besitzen einen Anteil von y%.
- Die "Reichsten"/"Größten" x% besitzen einen Anteil von y%.

Lorenzkurve

Definition:

- Das Merkmal darf nur *positive* Ausprägungen annehmen
- Die Gesamtsumme aller Merkmalswerte ist $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_{(j)}$
- Die Lorenzkurve verbindet Punktepaare bestehend aus den *kumulierten Summen* der nach Größe geordneten Beobachtungswerte $0 \leq x_{(1)} \leq \ldots \leq x_{(n)}$ und dem *relativen Anteil* der Individuen, die diese kumulierte Summe besitzen.

Lorenzkurve

Gestaltung:

- Es wird festgelegt: $u_{(0)} = 0$ und $v_{(0)} = 0$
- Die x-Achse wird in *gleiche Längen* aufgeteilt, deren Anzahl der der Individuen (Merkmalsausprägungen) entspricht:

$$u_i = \frac{i}{n}, \quad i = 1, ..., n$$

- Die y- Werte werden wie folgt berechnet:

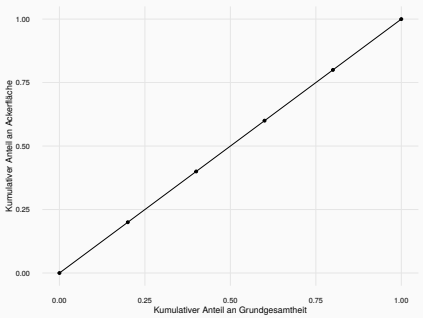
$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \quad i = 1, ..., n,$$

  also dem Quotienten aus der kumulierten Summe und der Gesamtsumme.
- Die so errechneten Koordinatenpunkte (u_i, v_i) werden in den Graphen eingetragen und mit Geraden verbunden.

Lorenzkurve

Beispiel: 5 Bauern teilen sich eine Ackerfläche von 100ha zu je 20ha.

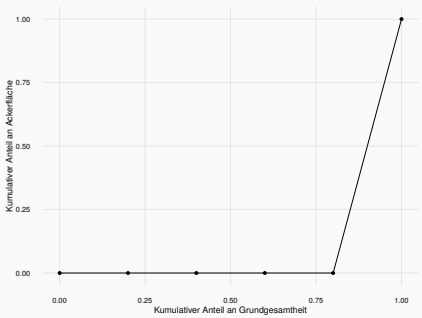
i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	20	$\frac{1}{5}$	$\frac{20}{100}$
2	20	$\frac{2}{5}$	$\frac{40}{100}$
3	20	$\frac{3}{5}$	$\frac{60}{100}$
4	20	$\frac{4}{5}$	$\frac{80}{100}$
5	20	$\frac{5}{5}$	$\frac{100}{100}$



Lorenzkurve

Beispiel: 4 Bauern besitzen nichts, 1 besitzt alles:

i	$x_{(i)}$	u_i	v_i
0	-	0	0
1	0	$\frac{1}{5}$	$\frac{0}{100}$
2	0	$\frac{2}{5}$	$\frac{0}{100}$
3	0	$\frac{3}{5}$	$\frac{0}{100}$
4	0	$\frac{4}{5}$	$\frac{0}{100}$
5	100	$\frac{5}{5}$	$\frac{100}{100}$

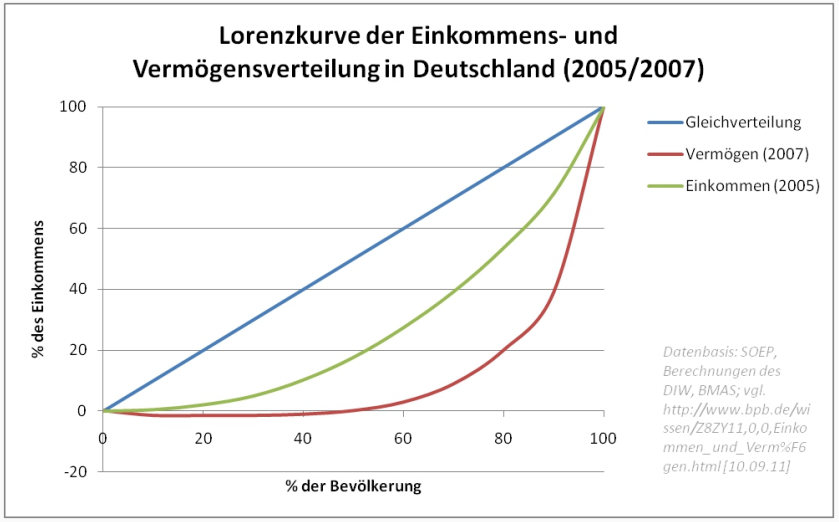


Lorenzkurve

Erscheinungsbild von Lorenzkurven:

- Die Koordinate $(u_0; v_0)$ ist *immer* $(0; 0)$.
- Die Koordinate $(u_n; v_n)$ ist *immer* $(1; 1)$.
- Der konstruierte Polygonzug verläuft *immer unterhalb* (im Grenzfall auf) der Winkelhalbierenden.
- Der konstruierte Polygonzug ist *(streng) monoton steigend*.
- Die Steigung des nächsten Polygonsegments ist entweder *gleich groß* oder *größer* als die Steigung des letzten Polygonsegments.

Lorenzkurve

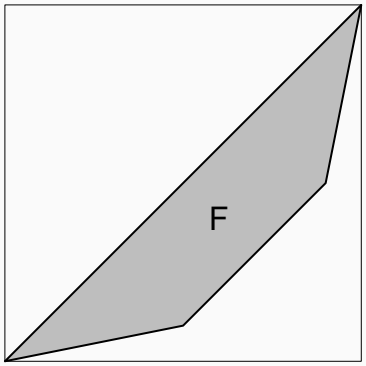


Gini-Koeffizient

Der **Gini-Koeffizient** bzw. das **Lorenz'sche Konzentrationsmaß** ist eine Maßzahl, die das *Ausmaß* der Konzentration beschreibt. Er ist definiert als

$$G = 2 \cdot F,$$

wobei F die Fläche zwischen der Diagonalen und der Lorenzkurve ist.



Gini-Koeffizient

Berechnung:

Für die praktische Berechnung von G aus den Wertepaaren $(u_i; v_i)$ stehen folgende alternative Formeln zur Verfügung:

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

oder alternativ

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}(v_{i-1} + v_i)$$

Wertebereich des Gini-Koeffizienten:

$$0 \leq G \leq \frac{n-1}{n}$$

Gini-Koeffizient

Normierter Gini-Koeffizient G^+:

Der Gini-Koeffizient wird auf folgende Weise normiert:

$$G^+ = \frac{n}{n-1}G$$

Er hat somit den Wertebereich

$$0 \leq G \leq 1,$$

wobei 0 für *keine Konzentration* (Gleichverteilung) und 1 für *vollständige Konzentration* (Monopol) steht.

Herfindahl-Index

$x_1, \ldots x_n$ seien die Daten mit $x_i \geq 0$.

Die Anteile der Einheiten i sind wie folgt definiert:

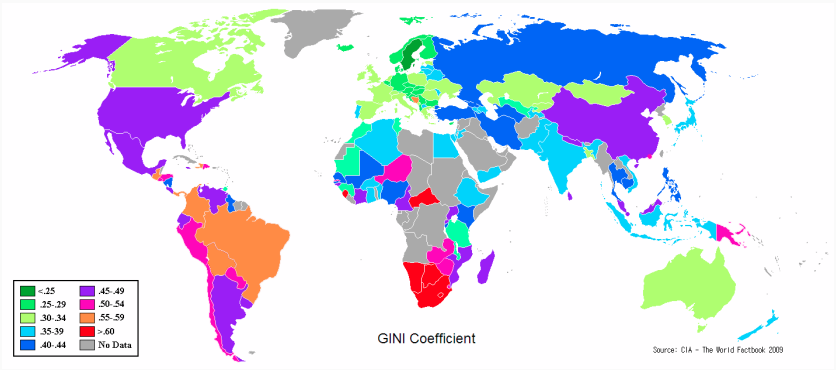
$$p_i := \frac{x_i}{\sum_{j=1}^{n} x_j}$$

Der Herfindahhl-Index ist

$$H_i := \sum_{j=1}^{n} p_i^2$$

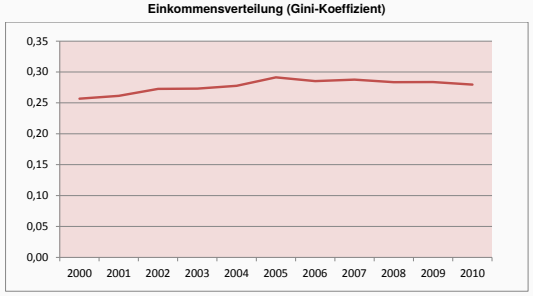
\end{block} Der Wertebereich ist von \$1/n \$ (Identische x) bis 1 (Monopol)

Gini-Index Einkommen weltweit



(World CIA Report 2009, Wikipedia)

Gini-Index Einkommen Deutschland



Quelle: Berechnungen des DIW Berlin auf Basis SOEP 2011.

Ein weiteres Verteilungsmaß ist der Gini-Koeffizient. Er beschreibt auf einer Skala von null bis eins die Ungleichheit der Verteilung. Je höher der Wert, umso ungleicher ist die Verteilung. Dieses Maß zeigt eine nach 2007 rückläufige Ungleichheit der Nettoäquivalenzeinkommen auf Haushaltsebene an. Dies umfasst alle Einkommensarten (insbesondere Einkommen aus Erwerb, Renten und Pensionen, aus Vermögen und Sozialtransfers). Der Trend einer Zunahme zwischen 2000 und 2005 hat sich also in der Zeit danach umgekehrt. Die Ungleichheit der Einkommen nimmt derzeit ab.

(Armuts- & Reichtumsbericht 2013, Bundesregierung)