Deskriptive Statistik und Wahrscheinlichkeitsrechnung

Vorlesung an der Universität des Saarlandes

Dr. Martin Becker

Sommersemester 2018

Organisatorisches I

- Vorlesung: Freitag, 12-14 Uhr, Gebäude B4 1, Audimax (HS 0.01)
- Übungen: nach gesonderter Ankündigung, Beginn: ab Montag, 16.04.
- Prüfung: 2-stündige Klausur nach Semesterende (1. Prüfungszeitraum)
  Anmeldung im ViPa nur vom 11.05. (8 Uhr) – 28.05. (15 Uhr)!
  (Abmeldung im ViPa bis 05.07., 12 Uhr)
- Hilfsmittel für Klausur
  - „Moderat" programmierbarer Taschenrechner, auch mit Grafikfähigkeit
  - 2 *beliebig gestaltete* DIN A 4–Blätter (bzw. 4, falls nur einseitig)
  - Benötigte Tabellen werden gestellt, aber **keine weitere Formelsammlung!**
- Durchgefallen — was dann?
  - „Wiederholungskurs" im kommenden (Winter-)Semester
  - „Nachprüfung" (voraussichtlich) erst März/April 2019 (2. Prüfungszeitraum)
  - „Reguläre" Vorlesung/Übungen wieder im Sommersemester 2019

Organisatorisches II

- Informationen und Materialien unter

  `http://www.lehrstab-statistik.de`

  bzw. spezieller

  `http://www.lehrstab-statistik.de/deskrwrss2018.html`

  (bei Problemen `http://www2.lehrstab-statistik.de` versuchen!)
- Kontakt: Dr. Martin Becker
  Geb. C3 1, 2. OG, Zi. 2.17
  e-Mail: `martin.becker@mx.uni-saarland.de`
- Sprechstunde nach Vereinbarung (Terminabstimmung per e-Mail)
- Vorlesungsunterlagen
  - Vorlesungsfolien (kein Kompaktskript)
  - Download (inklusive Drucker-freundlicher 2-auf-1 bzw. 4-auf-1 Versionen) in der Regel einige Tage vor der Vorlesung möglich

Organisatorisches III

- Übungsunterlagen
  - Wöchentliche Übungsblätter
  - Download i.d.R. kurz nach Ende der Vorlesung Freitag nachmittags möglich
  - Ebenfalls online: Ergebnisse *(keine Musterlösungen!)* zu einigen Aufgaben
  - Besprechung der Übungsblätter mit ausführlicheren Lösungsvorschlägen in den Übungsgruppen der folgenden Woche.
  - **Übungsaufgaben sollten unbedingt vorher selbst bearbeitet werden!**
  - Geplant: Freiwillige Bearbeitung und Abgabe von (höchstens zwei) Zusatzübungsblättern, die nach „Klausurmaßstäben" korrigiert zurückgegeben werden.
- Alte Klausuren
  - Aktuelle Klausuren inklusive der meisten Ergebnisse unter „Klausuren" auf Homepage des Lehrstabs verfügbar
  - Prüfungsrelevant sind (natürlich) alle in Vorlesung und Übungsgruppen besprochenen Inhalte, nicht nur die Inhalte der Altklausuren!

Was ist eigentlich „Statistik“?

- Der Begriff „Statistik“ hat verschiedene Bedeutungen, insbesondere:
  - Oberbegriff für die Gesamtheit der Methoden, die für die Erhebung und Verarbeitung empirischer Informationen relevant sind (→ statistische Methodenlehre)
  - (Konkrete) Tabellarische oder grafische Darstellung von Daten
  - (Konkrete) Abbildungsvorschrift, die in Daten enthaltene Informationen auf eine „Kennzahl“ (→ Teststatistik) verdichtet
- Grundlegende Teilgebiete der Statistik:
  - Deskriptive Statistik (auch: beschreibende Statistik, explorative Statistik)
  - Schließende Statistik (auch: inferenzielle Statistik, induktive Statistik)
- Typischer Einsatz von Statistik:

  Verarbeitung — insbesondere Aggregation — von (eventuell noch zu erhebenden) Daten mit dem Ziel, (informelle) Erkenntnisgewinne zu erhalten bzw. (formal) Schlüsse zu ziehen.

⇝ Bestimmte Informationen „ausblenden“, um neue Informationen zu erkennen

Vorurteile gegenüber Statistik

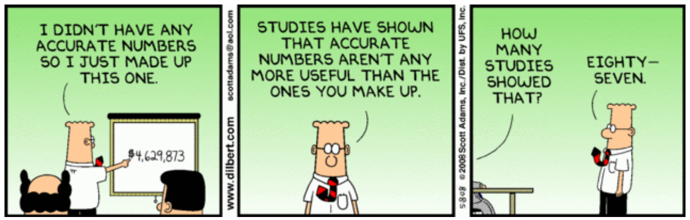
- Einige Zitate oder „Volksweisheiten“:
  - „Statistik ist pure Mathematik, und in Mathe war ich immer schlecht...“
  - „Mit Statistik kann man alles beweisen!“
  - „Ich glaube nur der Statistik, die ich selbst gefälscht habe.“ (häufig Winston Churchill zugeschrieben, aber eher Churchill von Goebbels' Propagandaministerium „in den Mund gelegt“)
  - „There are three kinds of lies: lies, damned lies, and statistics.“ (häufig Benjamin Disraeli zugeschrieben)

  ⇝ negative Vorurteile gegenüber der Disziplin „Statistik“
- Tatsächlich aber
  - benötigt man für viele statistische Methoden nur die vier Grundrechenarten.
  - ist „gesunder Menschenverstand“ viel wichtiger als mathematisches Know-How.
  - sind nicht die statistischen Methoden an sich schlecht oder gar falsch, sondern die korrekte Auswahl und Anwendung der Methoden zu hinterfragen.
  - werden viele (korrekte) Ergebnisse statistischer Untersuchungen lediglich falsch interpretiert.

Kann man mit Statistik lügen? I

Und falls ja, wie (schützt man sich dagegen)?

- Natürlich kann man mit Statistik „lügen“ bzw. täuschen!
- „Anleitung“ von Prof. Dr. Walter Krämer (TU Dortmund): *So lügt man mit Statistik, Piper, München, 2009*
- Offensichtliche Möglichkeit: Daten (vorsätzlich) manipulieren/fälschen:



Kann man mit Statistik lügen? II

Und falls ja, wie (schützt man sich dagegen)?

- Weitere Möglichkeiten zur Täuschung
  - Irreführende Grafiken
  - (Bewusstes) Weglassen relevanter Information
  - (Bewusste) Auswahl ungeeigneter statistischer Methoden
- Häufiges Problem (vor allem in den Medien): *Suggestion von Sicherheit durch hohe Genauigkeit angegebener Werte*

  ⇝ zusätzlich: Ablenkung vom „Adäquationsproblem“ (misst der angegebene Wert überhaupt das „Richtige“?)
- Schutz vor Täuschung:
  - Mitdenken!
  - „Gesunden Menschenverstand“ einschalten!
  - Gute Grundkenntnisse in Statistik!

Beispiel (Adäquationsproblem) I

vgl. Walter Krämer: So lügt man mit Statistik, Piper, München, 2009

- Frage: Was ist *im Durchschnitt* sicherer, Reisen mit Bahn oder Flugzeug?
- Statistik 1:

Bahn	9 Verkehrstote pro 10 Milliarden Passagierkilometer
Flugzeug	3 Verkehrstote pro 10 Milliarden Passagierkilometer

  $\rightsquigarrow$ Fliegen sicherer als Bahnfahren!
- Statistik 2:

Bahn	7 Verkehrstote pro 100 Millionen Passagierstunden
Flugzeug	24 Verkehrstote pro 100 Millionen Passagierstunden

  $\rightsquigarrow$ Bahnfahren sicherer als Fliegen!
- Widerspruch? Fehler?

Beispiel (Adäquationsproblem) II

vgl. Walter Krämer: So lügt man mit Statistik, Piper, München, 2009

- Nein, Unterschied erklärt sich durch höhere Durchschnittsgeschwindigkeit in Flugzeugen (Annahme: ca. 800 km/h vs. ca. 80 km/h)
- Wie wird „Sicherheit“ gemessen? Welcher „Durchschnitt“ ist geeigneter?
  $\rightsquigarrow$ Interpretation abhängig von der Fragestellung! Hier:
  - Steht man vor der Wahl, eine gegebene Strecke per Bahn oder Flugzeug zurückzulegen, so ist Fliegen sicherer.
  - Vor einem vierstündigen Flug ist dennoch eine größere „Todesangst“ angemessen als vor einer vierstündigen Bahnfahrt.

Beispiel („Schlechte“ Statistik) I

`http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html`

- Studie/Pressemitteilung des ACE Auto Club Europa *anläßlich des Frauentags am 8. März 2010*: „Autofahrerinnen im Osten am besten“
- Untersuchungsgegenstand:
  - Regionale Unterschiede bei Unfallhäufigkeit mit Frauen als Hauptverursacher
  - Vergleich Unfallhäufigkeit mit Frau bzw. Mann als Hauptverursacher
- Wesentliche Datengrundlage ist eine Publikation des Statistischen Bundesamts (Destatis): „Unfälle im Straßenverkehr nach Geschlecht 2008“

Beispiel („Schlechte“ Statistik) II

`http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html`

- Beginn der Pressemitteilung des ACE:

  „Von wegen schwaches Geschlecht: Hinterm Steuer sind Frauen besonders stark.“

  Weiter heißt es:

  “Auch die durch Autofahrerinnen verursachten Unfälle mit Personenschaden liegen wesentlich hinter den von Männern verursachten gleichartigen Karambolagen zurück.“

  und in einer Zwischenüberschrift

  „Schlechtere Autofahrerinnen sind immer noch besser als Männer“

Beispiel („Schlechte“ Statistik) III

http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html

- „Statistische“ Argumentation: Laut Destatis-Quelle sind (**angeblich!**)
  - mehr als 2/3 aller Unfälle mit Personenschaden 2008 (genauer: 217 843 von etwas über 320 000 Unfällen) durch PKW-fahrende Männer verursacht worden,
  - nur 37% aller Unfälle mit Personenschaden 2008 durch PKW-fahrende Frauen verursacht worden.
- Erste Auffälligkeit: 66.6% + 37% = 103.6% **(???)**
- Lösung: **Ablesefehler** (217 843 aller 320 614 Unfälle mit Personenschaden (67.9%) wurden mit **PKW-Fahrer** (geschlechtsunabhängig) als Hauptverursacher registriert)

Beispiel („Schlechte“ Statistik) IV

http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html

- Korrekte Werte:
  - Bei 210 905 der 217 843 Hauptunfallverursacher als PKW-Fahrzeugführer wurde Geschlecht registriert.
  - 132 757 waren männlich (62.95%), 78 148 weiblich (37.05%)
- **Also:** immer noch deutlich mehr Unfälle mit PKW-fahrenden Männern als Hauptverursacher im Vergleich zu PKW-Fahrerinnen.
- **Aber:** Absolute Anzahl von Unfällen geeignetes Kriterium für Fahrsicherheit?

Beispiel („Schlechte“ Statistik) V

http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html

- Modellrechnung des DIW aus dem Jahr 2004 schätzt
  - Anzahl Männer mit PKW-Führerschein auf 28.556 Millionen,
  - Anzahl Frauen mit PKW-Führerschein auf 24.573 Millionen.
- Weitere ältere Studie (von 2002) schätzt
  - durchschnittliche Fahrleistung von Männern mit PKW-Führerschein auf 30 km/Tag,
  - durchschnittliche Fahrleistung von Frauen mit PKW-Führerschein auf 12 km/Tag.
- Damit stehen also
  - bei Männern 132 757 verursachte Unfälle geschätzten $30 \cdot 365 \cdot 28.556 = 312688.2$ Millionen gefahrenen Kilometern,
  - bei Frauen 78 148 verursachte Unfälle geschätzten $12 \cdot 365 \cdot 24.573 = 107629.74$ Millionen gefahrenen Kilometern

  gegenüber.

Beispiel („Schlechte“ Statistik) VI

http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html

- Dies führt im Durchschnitt
  - bei Männern zu 0.425 verursachten Unfällen mit Personenschaden pro eine Million gefahrenen Kilometern,
  - bei Frauen zu 0.726 verursachten Unfällen mit Personenschaden pro eine Million gefahrenen Kilometern.
- Pro gefahrenem Kilometer verursachen (schätzungsweise) weibliche PKW-Fahrer also durchschnittlich ca. **71% mehr** Unfälle als männliche!
- Anstatt dies zu konkretisieren, räumt die Studie lediglich weit am Ende ein entsprechendes Ungleichgewicht bei der jährlichen Fahrleistung ein.

Beispiel („Schlechte“ Statistik) VII

http://www.ace-online.de/nc/der-club/news/autofahrerinnen-im-osten-am-besten.html

- Welt Online (siehe http://www.welt.de/vermischtes/article6674754/Frauen-sind-bessere-Autofahrer-als-Maenner.html) beruft sich auf die ACE-Studie in einem Artikel mit der Überschrift

  „Frauen sind bessere Autofahrer als Männer“

  und der prägnanten Bildunterschrift

  „Männer glauben bloß, sie seien die besseren Autofahrer. Eine Unfall-Statistik beweist das Gegenteil.“

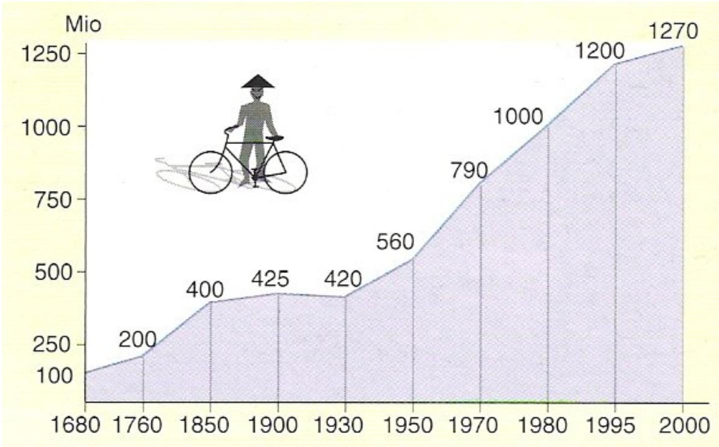
  Erst am Ende wird einschränkend erwähnt:

  „Fairerweise muss man erwähnen, dass Männer täglich deutlich mehr Kilometer zurücklegen. Und: Während 93 Prozent von ihnen einen Führerschein besitzen, sind es bei den Frauen lediglich 82 Prozent.“

Beispiel (Irreführende Grafik) I

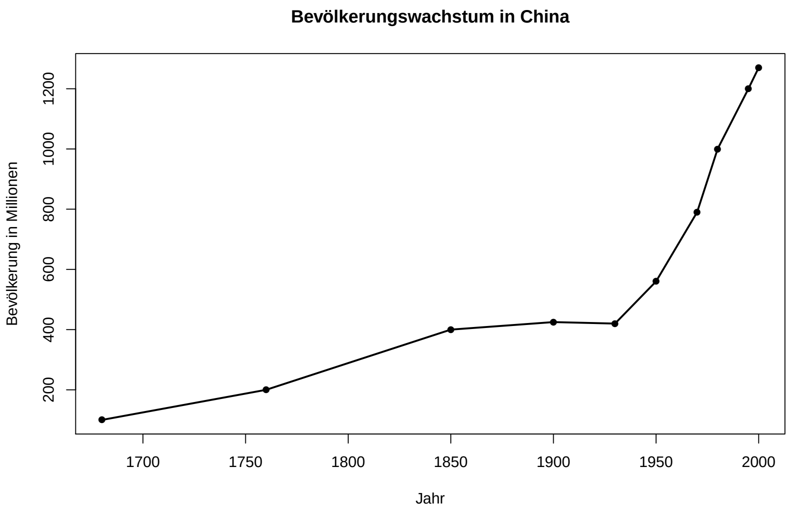
vgl. http://www.klein-singen.de/statistik/h/Wissenschaft/Bevoelkerungswachstum.html

Bevölkerungswachstum in China



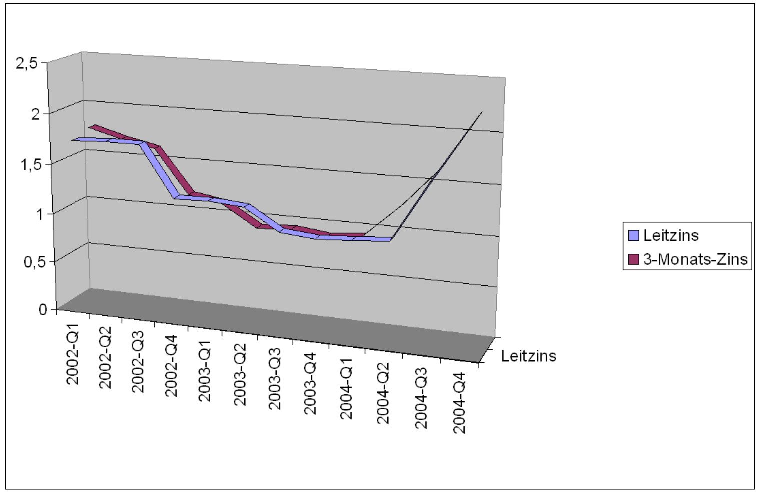
Beispiel (Irreführende Grafik) II

identischer Datensatz, angemessene Skala



Beispiel (Chartjunk)

Microsoft Excel mit Standardeinstellung für 3D-Liniendiagramme



Beispiel (Grafik ohne Chartjunk)

Statistik-Software **R**, identischer Datensatz

Kann Statistik auch nützlich sein?

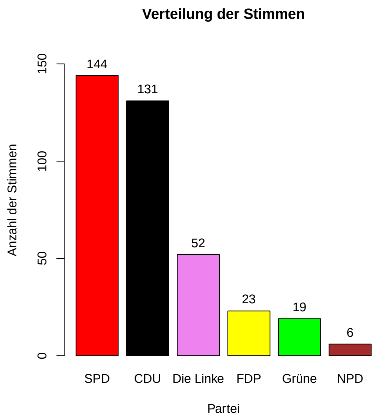
Welche Partei erhält wie viele Stimmen im Wahlbezirk 1.206 der Gemeinde Losheim am See bei den Erststimmen zur Bundestagswahl 2009? Stimmen:

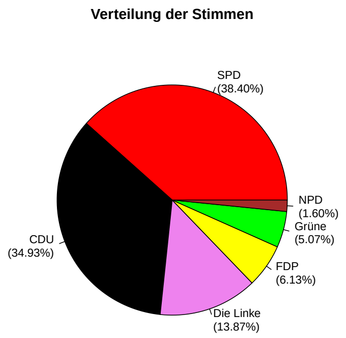
CDU, SPD, SPD, Die Linke, CDU, Die Linke, Die Linke, SPD, SPD, CDU, CDU, CDU, SPD, Grüne, FDP, SPD, SPD, NPD, SPD, FDP, Die Linke, Grüne, Grüne, CDU, CDU, Grüne, CDU, SPD, Die Linke, CDU, SPD, SPD, SPD, CDU, FDP, SPD, SPD, CDU, Grüne, SPD, FDP, SPD, SPD, SPD, SPD, Grüne, CDU, SPD, SPD, SPD, SPD, FDP, SPD, CDU, Grüne, CDU, CDU, SPD, SPD, CDU, FDP, CDU, SPD, CDU, SPD, CDU, SPD, SPD, CDU, SPD, Die Linke, SPD, CDU, Die Linke, Die Linke, Die Linke, SPD, CDU, NPD, FDP, Die Linke, Die Linke, Die Linke, CDU, SPD, CDU, SPD, CDU, CDU, CDU, CDU, CDU, SPD, SPD, SPD, SPD, CDU, CDU, SPD, Die Linke, SPD, CDU, CDU, SPD, SPD, CDU, CDU, FDP, SPD, SPD, Die Linke, SPD, NPD, CDU, CDU, CDU, SPD, CDU, Grüne, SPD, SPD, CDU, CDU, CDU, SPD, SPD, FDP, CDU, CDU, SPD, SPD, CDU, CDU, Die Linke, Die Linke, Die Linke, SPD, SPD, SPD, CDU, SPD, SPD, CDU, CDU, SPD, CDU, FDP, SPD, CDU, SPD, Die Linke, CDU, SPD, Die Linke, CDU, CDU, FDP, CDU, CDU, CDU, SPD, FDP, SPD, SPD, CDU, CDU, CDU, SPD, SPD, CDU, SPD, SPD, Die Linke, CDU, Grüne, Die Linke, SPD, SPD, SPD, Die Linke, CDU, SPD, SPD, SPD, Die Linke, Die Linke, SPD, SPD, CDU, SPD, CDU, Die Linke, FDP, FDP, CDU, CDU, Die Linke, SPD, SPD, CDU, Die Linke, CDU, SPD, CDU, CDU, SPD, CDU, CDU, SPD, SPD, SPD, SPD, SPD, SPD, CDU, Die Linke, SPD, Die Linke, CDU, SPD, Die Linke, SPD, CDU, Grüne, SPD, Die Linke, CDU, SPD, SPD, CDU, SPD, SPD, SPD, Grüne, Die Linke, Die Linke, FDP, SPD, CDU, SPD, CDU, SPD, CDU, CDU, Die Linke, Die Linke, SPD, CDU, Grüne, FDP, SPD, SPD, CDU, SPD, CDU, CDU, SPD, CDU, Die Linke, Grüne, Die Linke, Die Linke, Die Linke, SPD, Die Linke, CDU, CDU, CDU, Die Linke, CDU, SPD, Die Linke, Die Linke, SPD, SPD, SPD, SPD, SPD, CDU, SPD, CDU, SPD, Grüne, CDU, CDU, SPD, Die Linke, Grüne, CDU, FDP, Die Linke, Grüne, SPD, CDU, CDU, SPD, FDP, SPD, Die Linke, SPD, CDU, CDU, FDP, Die Linke, SPD, CDU, NPD, FDP, FDP, SPD, NPD, SPD, SPD, SPD, CDU, CDU, CDU, Grüne, SPD, SPD, SPD, FDP, CDU, CDU, SPD, Die Linke, CDU, Die Linke, SPD, CDU, SPD, Die Linke, CDU, Die Linke, CDU, CDU, CDU, SPD, SPD, SPD, Grüne, SPD, SPD, CDU, FDP, Grüne, CDU, CDU, CDU, CDU, CDU, SPD, NPD, CDU, SPD, CDU, SPD, CDU, SPD, CDU, SPD, SPD, SPD, CDU, CDU, CDU, CDU, Die Linke, CDU, CDU, SPD, CDU

- Mit etwas (deskriptiver) Statistik in tabellarischer Form:

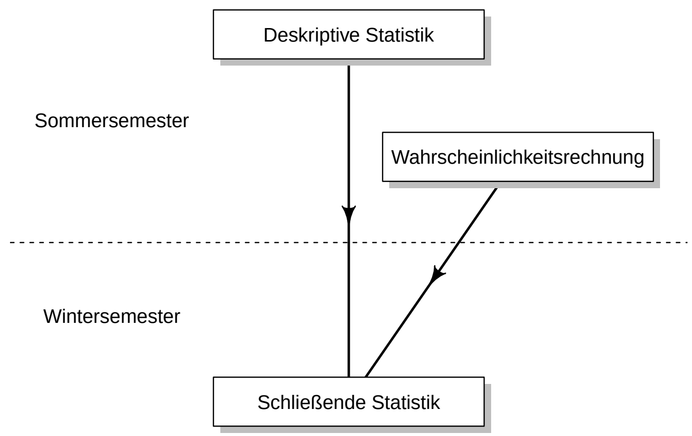
	SPD	CDU	Die Linke	FDP	Grüne	NPD	Summe
Anzahl der Stimmen	144	131	52	23	19	6	375
Stimmenanteil in %	38.40	34.93	13.87	6.13	5.07	1.60	100.00

- Grafisch aufbereitete Varianten:





Organisation der Statistik-Veranstaltungen



Teil I

Deskriptive Statistik

Datenerhebung I

- Beginn jeder (deskriptiven) statistischen Untersuchung: Datenerhebung
- Zu einer **Menge von Merkmalsträgern** (**statistische Masse**), eventuell Teil einer größeren **Grundgesamtheit**, werden ein oder mehrere **Merkmale** erhoben
- Unterscheidung nach
  - Primärerhebung ↔ Sekundärerhebung:
    Neue Erhebung oder Nutzung von vorhandenem Datenmaterial
  - Vollerhebung ↔ Teilerhebung:
    Erhebung der Merkmale für ganze Grundgesamtheit oder Teilgesamtheit

Datenerhebung II

- Bei Primärerhebung: Untersuchungsziel bestimmt
  - Auswahl bzw. Abgrenzung der statistischen Masse
  - Auswahl der zu erhebenden Merkmale
  - Art der Erhebung, z.B. Befragung (Post, Telefon, Internet, persönlich), Beobachtung, Experiment
- Sorgfalt bei Datenerhebung enorm wichtig:
  Fehler bei Datenerhebung sind später nicht mehr zu korrigieren!
- Ausführliche Diskussion hier aus Zeitgründen nicht möglich

Vorsicht vor „falschen Schlüssen“! I

- Deskriptive Statistik fasst lediglich Information über statistische Masse zusammen
- Schlüsse auf (größere) „Grundgesamtheit“ (bei Teilerhebung)
  $\rightsquigarrow$ Schließende Statistik
- Dennoch häufig zu beobachten:
  „Informelles“ Übertragen der Ergebnisse in der statistischen Masse auf größere Menge von Merkmalsträgern

$\rightsquigarrow$ Gefahr von falschen Schlüssen!

Vorsicht vor „falschen Schlüssen“! II

Beispiel: Bachelor-Absolventen (vgl. Krämer: So lügt man mit Statistik)

Hätte man am Ende des SS 2011 in der statistischen Masse der Absolventen des BWL-Bachelorstudiengangs in Saarbrücken die Merkmale „Studiendauer“ und „Abschlussnote“ erhoben, würde man wohl feststellen, dass alle Abschlüsse in Regelstudienzeit und im Durchschnitt mit einer guten Note erfolgt sind. Warum? Kann man dies ohne weiteres auf Absolventen anderer Semester übertragen?

$\rightsquigarrow$ Zur Interpretationsfähigkeit von Ergebnissen statistischer Untersuchungen:

- Abgrenzung der zugrundeliegenden statistischen Masse **sehr** wichtig
- (Möglichst) objektive Festlegung nach Kriterien zeitlicher, räumlicher und sachlicher Art

Definition 2.1 (Menge, Mächtigkeit, Tupel)

1. Eine (endliche) **Menge** M ist die Zusammenfassung (endlich vieler) unterschiedlicher Objekte (Elemente).
2. Zu einer endlichen Menge M bezeichnen $\#M$ oder auch $|M|$ die Anzahl der Elemente in M. $\#M$ bzw. $|M|$ heißen auch **Mächtigkeit** der Menge M.
3. Für eine Anzahl $n \geq 1$ von (nicht notwendigerweise verschiedenen!) Elementen $x_1, x_2, \ldots, x_n$ aus einer Menge M wird eine (nach ihrer Reihenfolge geordnete) Auflistung $(x_1, x_2, \ldots, x_n)$ bzw. $x_1, x_2, \ldots, x_n$ als n-**Tupel** aus der Menge M bezeichnet. 2-Tupel (x_1, x_2) heißen auch Paare.
4. Lassen sich die Elemente der Menge M (der Größe nach) ordnen, so sei (zu einer vorgegebenen Ordnung)
   1. mit $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ bzw. $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ das der Größe nach geordnete n-Tupel der n Elemente $x_1, x_2, \ldots, x_n$ aus M bezeichnet, es gelte also $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
   2. zu einer endlichen Teilmenge $A \subseteq M$ der Mächtigkeit m mit $(a_{(1)}, a_{(2)}, \ldots, a_{(m)})$ bzw. $a_{(1)}, a_{(2)}, \ldots, a_{(m)}$ das der Größe nach geordnete m-Tupel der Elemente $a_1, a_2, \ldots, a_m$ von A bezeichnet, es gelte also $a_{(1)} < a_{(2)} < \cdots < a_{(m)}$.

Merkmalswerte, Merkmalsraum, Urliste I

- Bei der Datenerhebung werden den Merkmalsträgern zu jedem erhobenen Merkmal **Merkmalswerte** oder **Beobachtungswerte** zugeordnet.
- Man nimmt an, dass man (im Prinzip auch vor der Erhebung) eine Menge M angeben kann, die alle vorstellbaren Merkmalswerte eines Merkmals enthält.
- Das n-Tupel $(x_1, \ldots, x_n)$ der Merkmalswerte $x_1, \ldots, x_n$ (aus der Menge M) zu einem bei den n Merkmalsträgern erhobenen Merkmal X bezeichnet man als **Urliste**.
- Die Menge A der (verschiedenen) in der Urliste (tatsächlich) auftretenden Merkmalswerte, in Zeichen

$$A := \{a \in M \mid \exists i \in \{1, \ldots, n\} \text{ mit } x_i = a\} \ ,$$

heißt **Merkmalsraum**, ihre Elemente **Merkmalsausprägungen**.

Merkmalswerte, Merkmalsraum, Urliste II

Beispiel Wahlergebnis

- Urliste (siehe Folie 22) aus gewählten Parteien der 375 abgegebenen gültigen Stimmen:

$$\begin{aligned} &x_1 = \text{“CDU”}, x_2 = \text{“SPD”}, x_3 = \text{“SPD”}, x_4 = \text{“Die Linke”}, x_5 = \text{“CDU”}, \\ &x_6 = \text{“Die Linke”}, x_7 = \text{“Die Linke”}, x_8 = \text{“SPD”}, x_9 = \text{“SPD”}, x_{10} = \text{“CDU”}, \\ &x_{11} = \text{“CDU”}, x_{12} = \text{“CDU”}, x_{13} = \text{“SPD”}, x_{14} = \text{“Grüne”}, x_{15} = \text{“FDP”}, \\ &x_{16} = \text{“SPD”}, x_{17} = \text{“SPD”}, x_{18} = \text{“NPD”}, x_{19} = \text{“SPD”}, x_{20} = \text{“FDP”}, \ldots \end{aligned}$$

- Merkmalsraum: $A = \{\text{SPD, CDU, Die Linke, FDP, Grüne, NPD}\}$

Merkmalstypen I

Definition 2.2 (Merkmalstypen)

1. Ein Merkmal heißt
   - **nominalskaliert**, wenn seine Ausprägungen lediglich unterschieden werden sollen,
   - **ordinalskaliert** oder **rangskaliert**, wenn (darüberhinaus) eine (Rang-)Ordnung auf den Ausprägungen vorgegeben ist,
   - **kardinalskaliert** oder **metrisch skaliert**, wenn (darüberhinaus) ein „Abstand" auf der Menge der Ausprägungen vorgegeben ist, also wenn das Ausmaß der Unterschiede zwischen verschiedenen Ausprägungen gemessen werden kann.
2. Ein Merkmal heißt **quantitativ**, wenn es kardinalskaliert ist, **qualitativ** sonst.
3. Ein Merkmal heißt
   - **diskret**, wenn es qualitativ ist oder wenn es quantitativ ist und die Menge der möglichen Ausprägungen endlich oder abzählbar unendlich ist.
   - **stetig**, wenn es quantitativ ist und für je zwei mögliche Merkmalsausprägungen auch alle Zwischenwerte angenommen werden können.

Merkmalstypen II

- Welche der in Definition 2.2 erwähnten Eigenschaften für ein Merkmal zutreffend sind, hängt von der jeweiligen Anwendungssituation ab.
- Insbesondere ist die Abgrenzung zwischen stetigen und diskreten Merkmalen oft schwierig (allerdings meist auch nicht besonders wichtig).
- Damit ein Merkmal (mindestens) ordinalskaliert ist, muss die verwendete Ordnung — insbesondere bei Mehrdeutigkeit — eindeutig festgelegt sein.
- Häufig findet man zusätzlich zu den in 2.2 erläuterten Skalierungen auch die Begriffe **Intervallskala**, **Verhältnisskala** und **Absolutskala**. Diese stellen eine feinere Unterteilung der Kardinalskala dar.
- *Unabhängig vom Skalierungsniveau* heißt ein Merkmal **numerisch**, wenn seine Merkmalsausprägungen Zahlenwerte sind.

Merkmalstypen III

Beispiel (Merkmalstypen)

- nominalskalierte Merkmale: Geschlecht (Ausprägungen: „männlich", „weiblich"), Parteien (siehe Wahlergebnis-Beispiel)
- ordinalskalierte Merkmale: Platzierungen, Zufriedenheit („sehr zufrieden", „eher zufrieden", „weniger zufrieden", „unzufrieden")
- kardinalskalierte Merkmale: Anzahl Kinder, Anzahl Zimmer in Wohnung, Preise, Gewichte, Streckenlängen, Zeiten
  - davon diskret: Anzahl Kinder, Anzahl Zimmer in Wohnung,
  - davon (eher) stetig: Preise, Gewichte, Streckenlängen, Zeiten

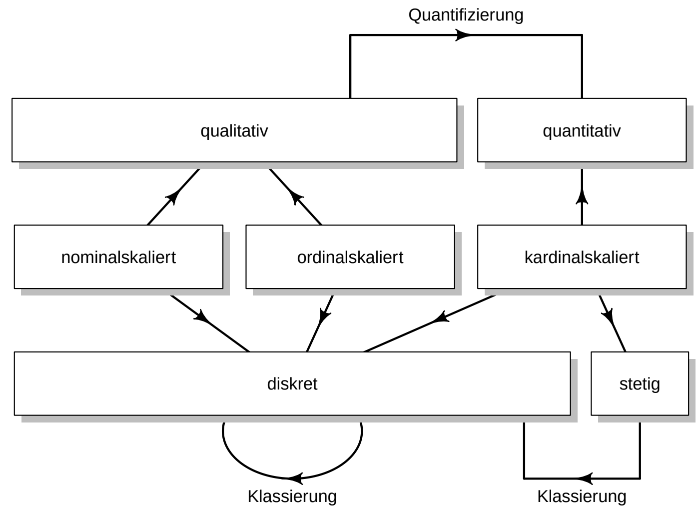
Umwandlung von Merkmalstypen I

- Umwandlung qualitativer in quantititative Merkmale durch **Quantifizierung**:
  - Ersetzen des qualitativen Merkmals „Berufserfahrung" mit den Ausprägungen „Praktikant", „Lehrling", „Geselle", „Meister" durch quantitatives Merkmal, dessen Ausprägungen den (mindestens) erforderlichen Jahren an Berufspraxis entspricht, die zum Erreichen des Erfahrungsgrades erforderlich sind.
  - Ersetzen des qualitativen Merkmals Schulnote mit den Ausprägungen „sehr gut", „gut", „befriedigend", „ausreichend", „mangelhaft", „ungenügend" (eventuell feiner abgestuft durch Zusätze „+" und „-") durch quantitatives Merkmal, z.B. mit den Ausprägungen $15, 14, \ldots, 00$ oder den Ausprägungen $1.0, 1.3, 1.7, 2.0, 2.3, \ldots, 4.7, 5.0, 6.0$.
  - **Vorsicht:** Umwandlung nur sinnvoll, wenn Abstände tatsächlich (sinnvoll) interpretiert werden können!

Umwandlung von Merkmalstypen II

- Umwandlung stetiger in diskrete Merkmale durch **Klassierung** oder **Gruppierung**, d.h. Zusammenfassen ganzer Intervalle zu einzelnen Ausprägungen, z.B. Gewichtsklassen beim Boxsport.
  - Klassierung ermöglicht auch Umwandlung diskreter Merkmale in (erneut) diskrete Merkmale mit unterschiedlichem Merkmalsraum, z.B. Unternehmensgrößen kleiner und mittlerer Unternehmen nach Anzahl der Beschäftigten mit Ausprägungen „1-9", „10-19", „20-49", „50-249".
  - Klassierung erfolgt regelmäßig (aber nicht immer) bereits vor der Datenerhebung.

Übersichtsdarstellung Merkmalstypen



Inhaltsverzeichnis

(Ausschnitt)

Häufigkeitsverteilungen I

- Geeignetes Mittel zur Verdichtung der Information aus Urlisten vor allem bei diskreten Merkmalen mit „wenigen" Ausprägungen: **Häufigkeitsverteilungen**
- Zur Erstellung einer Häufigkeitsverteilung: Zählen, wie oft jede Merkmalsausprägung a aus dem Merkmalsraum $A = \{a_1, \ldots, a_m\}$ in der Urliste $(x_1, \ldots, x_n)$ vorkommt.
  - Die **absoluten Häufigkeiten** $h(a)$ geben für die Merkmalsausprägung $a \in A$ die (absolute) Anzahl der Einträge der Urliste mit der Ausprägung a an, in Zeichen
  $$h(a) := \#\{i \in \{1, \ldots, n\} \mid x_i = a\} .$$
  - Die **relativen Häufigkeiten** $r(a)$ geben für die Merkmalsausprägung $a \in A$ den (relativen) Anteil der Einträge der Urliste mit der Ausprägung a an der gesamten Urliste an, in Zeichen
  $$r(a) := \frac{h(a)}{n} = \frac{\#\{i \in \{1, \ldots, n\} \mid x_i = a\}}{n} .$$

Häufigkeitsverteilungen II

- Die absoluten Häufigkeiten sind natürliche Zahlen und summieren sich zu n auf (i.Z. $\sum_{j=1}^{m} h(a_j) = n$).
- Die relativen Häufigkeiten sind Zahlen zwischen 0 und 1 (bzw. zwischen 0% und 100%) und summieren sich zu 1 (bzw. 100%) auf (i.Z. $\sum_{j=1}^{m} r(a_j) = 1$).
- Ist die Anordnung (Reihenfolge) der Urliste unwichtig, geht durch Übergang zur Häufigkeitsverteilung keine relevante Information verloren.
- Häufigkeitsverteilungen werden in der Regel in tabellarischer Form angegeben, am Beispiel des Wahlergebnisses:

	SPD	CDU	Die Linke	FDP	Grüne	NPD	Summe
a_j	a_1	a_2	a_3	a_4	a_5	a_6	Σ
$h(a_j)$	144	131	52	23	19	6	375
$r(a_j)$	0.3840	0.3493	0.1387	0.0613	0.0507	0.0160	1.0000

Häufigkeitsverteilungen III

- Grafische Darstellung (insbesondere bei nominalskalierten Merkmalen) durch **Balkendiagramme** (auch: Säulendiagramme) oder **Kuchendiagramme** (siehe Folie 23).
- Balkendiagramme meist geeigneter als Kuchendiagramme (außer, wenn die anteilige Verteilung der Merkmalsausprägungen im Vordergrund steht)
- Oft mehrere Anordnungen der Spalten/Balken/Kreissegmente bei nominalskalierten Merkmalen plausibel, absteigende Sortierung nach Häufigkeiten $h(a_j)$ meist sinnvoll.
- Bei ordinalskalierten Merkmalen zweckmäßig: Sortierung der Merkmalsausprägungen nach vorgegebener Ordnung, also

$$a_1 = a_{(1)}, a_2 = a_{(2)}, \ldots, a_m = a_{(m)}$$

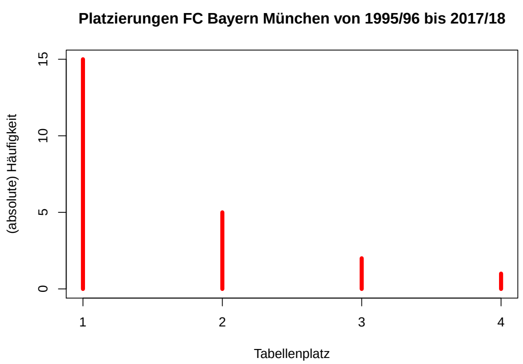
- Alternative grafische Darstellung bei (mindestens) ordinalskalierten Merkmalen mit numerischen Ausprägungen: **Stabdiagramm**

Häufigkeitsverteilungen IV

- Stabdiagramm zur Urliste

$$2, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 4, 1, 2, 1, 3, 2, 1, 1, 1, 1, 1, 1$$

der finalen Tabellenplätze des FC Bayern München in der (ersten) Fußball-Bundesliga (Saison 1995/96 bis 2017/2018):



Empirische Verteilungsfunktion

- Bei (mindestens ordinalskalierten) numerischen Merkmalen interessante Fragestellungen:
  - Wie viele Merkmalswerte sind kleiner/größer als ein vorgegebener Wert?
  - Wie viele Merkmalswerte liegen in einem vorgegebenem Bereich (Intervall)?
- Hierzu nützlich: **(relative) kumulierte Häufigkeitsverteilung**, auch bezeichnet als **empirische Verteilungsfunktion**
- Die empirische Verteilungsfunktion $F(x)$ ordnet einer Zahl x den Anteil der Merkmalswerte $x_1, \ldots, x_n$ zu, die kleiner oder gleich x sind, also

$$F(x) := \frac{\#\{i \in \{1, \ldots, n\} \mid x_i \leq x\}}{n} .$$

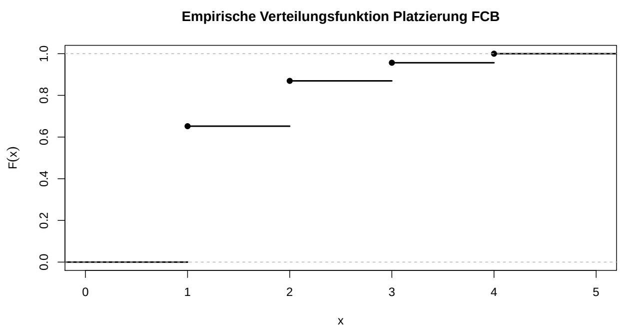
- Ein Vergleich mit den Definitionen von $h(a)$ und $r(a)$ offenbart (!), dass $F(x)$ auch mit Hilfe von $h(a)$ bzw. $r(a)$ berechnet werden kann; gibt es m Merkmalsausprägungen, so gilt:

$$F(x) = \frac{1}{n} \sum_{\substack{a_j \leq x \\ 1 \leq j \leq m}} h(a_j) = \sum_{\substack{a_j \leq x \\ 1 \leq j \leq m}} r(a_j)$$

- Beispiel: Empirische Verteilungsfunktion für FC Bayern-Platzierungen

$$F(x)=\begin{cases}0 & \text{für} \quad x<1\\ \frac{15}{23} & \text{für} \quad 1\le x<2\\ \frac{20}{23} & \text{für} \quad 2\le x<3\\ \frac{22}{23} & \text{für} \quad 3\le x<4\\ 1 & \text{für} \quad x\ge 4\end{cases} \approx \begin{cases}0.000 & \text{für} \quad x<1\\ 0.652 & \text{für} \quad 1\le x<2\\ 0.870 & \text{für} \quad 2\le x<3\\ 0.957 & \text{für} \quad 3\le x<4\\ 1.000 & \text{für} \quad x\ge 4\end{cases}$$

- Grafische Darstellung der empirischen Verteilungsfunktion:



Relative Häufigkeiten von Intervallen I

(bei numerischen Merkmalen)

- Relative Häufigkeit $r(a)$ ordnet Ausprägungen $a \in A$ zugehörigen Anteil von a an den Merkmalswerten zu.
- $r(\cdot)$ kann auch für $x \in \mathbb{R}$ mit $x \notin A$ ausgewertet werden ($\rightsquigarrow r(x) = 0$).
- „Erweiterung" von $r(\cdot)$ auch auf Intervalle möglich:
- $F(b)$ gibt für $b \in \mathbb{R}$ bereits Intervallhäufigkeit

$$F(b) = r\left((-\infty, b]\right) = r\left(\{x \in \mathbb{R} \mid x \le b\}\right)$$

an.

Relative Häufigkeiten von Intervallen II

(bei numerischen Merkmalen)

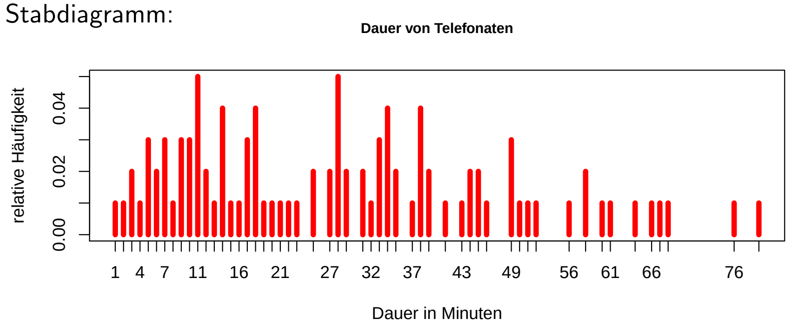
- Relative Häufigkeit des offenen Intervalls $(-\infty, b)$ als Differenz

$$r\left((-\infty, b)\right) = r\left((-\infty, b]\right) - r(b) = F(b) - r(b)$$

- Analog: relative Häufigkeiten weiterer Intervalle:
  - $r\left((a, \infty)\right) = 1 - F(a)$
  - $r\left([a, \infty)\right) = 1 - (F(a) - r(a)) = 1 - F(a) + r(a)$
  - $r\left([a, b]\right) = F(b) - (F(a) - r(a)) = F(b) - F(a) + r(a)$
  - $r\left((a, b]\right) = F(b) - F(a)$
  - $r\left([a, b)\right) = (F(b) - r(b)) - (F(a) - r(a)) = F(b) - r(b) - F(a) + r(a)$
  - $r\left((a, b)\right) = (F(b) - r(b)) - F(a) = F(b) - r(b) - F(a)$

Häufigkeitsverteilungen klassierter Daten I

- Bisherige Analysemethoden schlecht geeignet für stetige Merkmale bzw. diskrete Merkmale mit „vielen" Ausprägungen
- (Fiktives) Beispiel: Dauer von 100 Telefonaten (in Minuten)
  - Urliste: 44, 35, 22, 5, 50, 5, 3, 17, 19, 67, 49, 52, 16, 34, 11, 27, 14, 1, 35, 11, 3, 49, 18, 58, 43, 34, 79, 34, 7, 38, 28, 21, 27, 51, 9, 17, 10, 60, 14, 32, 9, 18, 11, 23, 25, 10, 76, 28, 13, 15, 28, 7, 31, 45, 66, 61, 39, 25, 17, 33, 4, 41, 29, 38, 18, 44, 28, 12, 64, 6, 38, 8, 37, 38, 28, 5, 7, 34, 11, 2, 31, 14, 33, 39, 12, 49, 14, 58, 45, 56, 46, 68, 18, 6, 11, 10, 29, 33, 9, 20
  - Stabdiagramm:



Häufigkeitsverteilungen klassierter Daten II

- **Problem:** viele Merkmalswerte treten nur einmalig (oder „selten“) auf
  $\rightsquigarrow$ Aussagekraft von Häufigkeitstabellen und Stabdiagrammen gering
- **Lösung:** Zusammenfassen mehrerer Merkmalsausprägungen in Klassen
- Zu dieser **Klassierung** erforderlich: **Vorgabe** der Grenzen $k_0, k_1, \dots, k_l$ von l (rechtsseitig abgeschlossenen) Intervallen

$$K_1 := (k_0, k_1],\; K_2 := (k_1, k_2],\; \dots,\; K_l := (k_{l-1}, k_l]\ ,$$

die alle n Merkmalswerte überdecken
(also mit $k_0 < x_i \le k_l$ für alle $i \in \{1, \dots, n\}$)

Häufigkeitsverteilungen klassierter Daten III

- Wichtige Kennzahlen der Klassierung (bzw. der klassierten Daten):

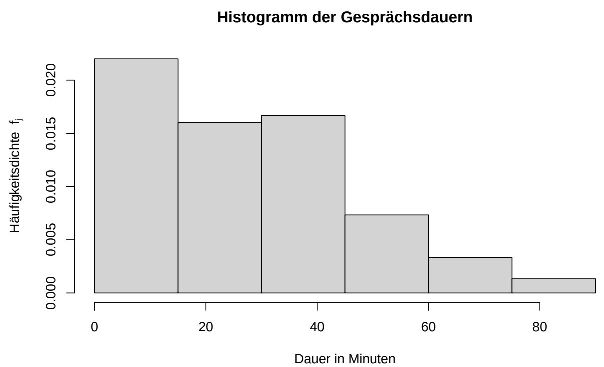
Klassenbreiten $b_j := k_j - k_{j-1}$
Klassenmitten $m_j := \frac{k_{j-1}+k_j}{2}$
absolute Häufigkeiten $h_j := \#\{i \in \{1, \dots, n\} \mid k_{j-1} < x_i \le k_j\}$
relative Häufigkeiten $r_j := \frac{h_j}{n}$
Häufigkeitsdichten $f_j := \frac{r_j}{b_j}$

(jeweils für $j \in \{1, \dots, l\}$).

- Übliche grafische Darstellung von klassierten Daten: **Histogramm**
- Hierzu: Zeichnen der Rechtecke mit Höhen f_j über den Intervallen K_j (also der Rechtecke mit den Eckpunkten $(k_{j-1}, 0)$ und (k_j, f_j))

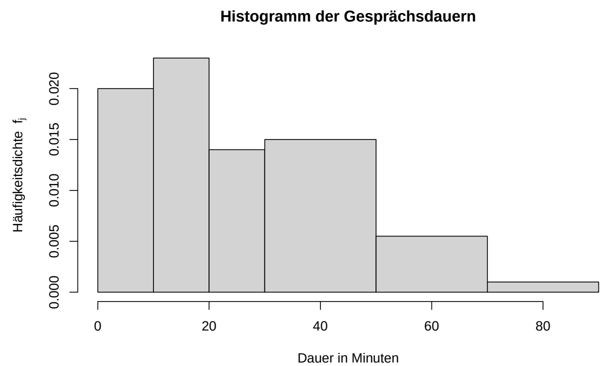
- Am Beispiel der Gesprächsdauern bei 6 Klassen zu je 15 Minuten Breite:

Nr. j	Klasse $K_j = (k_{j-1}, k_j]$	Klassen-breite b_j	Klassen-mitte m_j	absolute Häufigkeit h_j	relative Häufigkeit $r_j = \frac{h_j}{n}$	Häufigkeits-dichte $f_j = \frac{r_j}{b_j}$	Verteilungs-funktion $F(k_j)$
1	$(0, 15]$	15	7.5	33	0.33	0.022	0.33
2	$(15, 30]$	15	22.5	24	0.24	0.016	0.57
3	$(30, 45]$	15	37.5	25	0.25	$0.01\overline{6}$	0.82
4	$(45, 60]$	15	52.5	11	0.11	$0.007\overline{3}$	0.93
5	$(60, 75]$	15	67.5	5	0.05	$0.00\overline{3}$	0.98
6	$(75, 90]$	15	82.5	2	0.02	$0.001\overline{3}$	1.00



- Alternativ mit 6 Klassen bei 2 verschiedenen Breiten:

Nr. j	Klasse $K_j = (k_{j-1}, k_j]$	Klassen-breite b_j	Klassen-mitte m_j	absolute Häufigkeit h_j	relative Häufigkeit $r_j = \frac{h_j}{n}$	Häufigkeits-dichte $f_j = \frac{r_j}{b_j}$	Verteilungs-funktion $F(k_j)$
1	$(0, 10]$	10	5	20	0.20	0.0200	0.20
2	$(10, 20]$	10	15	23	0.23	0.0230	0.43
3	$(20, 30]$	10	25	14	0.14	0.0140	0.57
4	$(30, 50]$	20	40	30	0.30	0.0150	0.87
5	$(50, 70]$	20	60	11	0.11	0.0055	0.98
6	$(70, 90]$	20	80	2	0.02	0.0010	1.00



Bemerkungen I

- Der **Flächeninhalt** der einzelnen Rechtecke eines Histogramms entspricht der relativen Häufigkeit der zugehörigen Klasse
  - $\rightsquigarrow$ Die Summe aller Flächeninhalte beträgt 1
  - $\rightsquigarrow$ Die Höhe der Rechtecke ist nur dann proportional zu der relativen Häufigkeit der Klassen, falls alle Klassen die gleiche Breite besitzen!
- Die Klassierung ist abhängig von der Wahl der Klassengrenzen, unterschiedliche Klassengrenzen können einen Datensatz auch sehr unterschiedlich erscheinen lassen $\rightsquigarrow$ Potenzial zur Manipulation
- Es existieren verschiedene Algorithmen zur automatischen Wahl von Klassenanzahl und -grenzen (z.B. nach Scott, Sturges, Freedman-Diaconis)

Bemerkungen II

- Durch Klassierung geht Information verloren!
  - Spezielle Verfahren für klassierte Daten vorhanden
  - Verfahren approximieren ursprüngliche Daten in der Regel durch die Annahme gleichmäßiger Verteilung innerhalb der einzelnen Klassen
  - (Approximative) Verteilungsfunktion (ebenfalls mit $F(x)$ bezeichnet) zu klassierten Daten entsteht so durch lineare Interpolation der an den Klassengrenzen k_j bekannten (und auch nach erfolgter Klassierung noch exakten!) Werte der empirischen Verteilungsfunktion $F(k_j)$
  - Näherungsweise Berechnung von Intervallhäufigkeiten dann gemäß Folie 46 f. mit der approximativen empirischen Verteilungsfunktion $F(x)$.

(Approx.) Verteilungsfunktion bei klassierten Daten

Approximative Verteilungsfunktion bei klassierten Daten

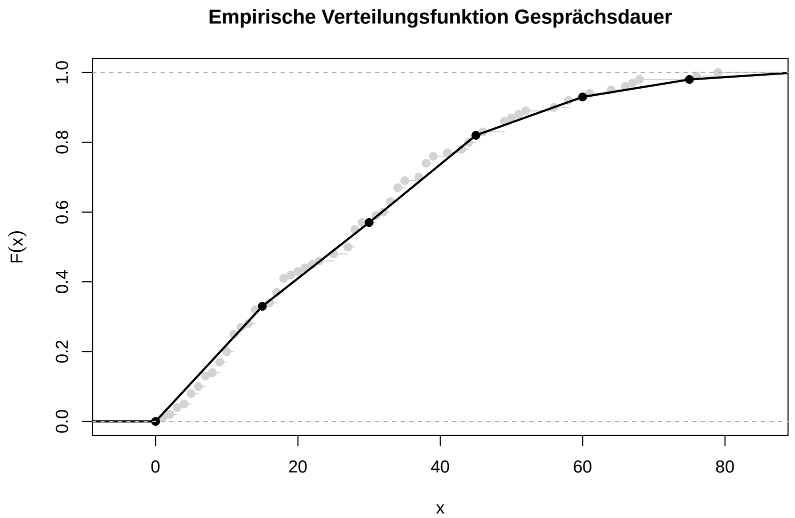
$$F(x) = \begin{cases} 0 & \text{für} \quad x \leq k_0 \\ F(k_{j-1}) + f_j \cdot (x - k_{j-1}) & \text{für} \quad k_{j-1} < x \leq k_j,\ j \in \{1, \ldots, l\} \\ 1 & \text{für} \quad x > k_l \end{cases}$$

- Am Beispiel der Gesprächdauern (Klassierung aus Folie 52)

$$F(x) = \begin{cases} 0 & \text{für} \quad x \leq 0 \\ 0.0200 \cdot (x - 0) & \text{für} \quad 0 < x \leq 10 \\ 0.20 + 0.0230 \cdot (x - 10) & \text{für} \quad 10 < x \leq 20 \\ 0.43 + 0.0140 \cdot (x - 20) & \text{für} \quad 20 < x \leq 30 \\ 0.57 + 0.0150 \cdot (x - 30) & \text{für} \quad 30 < x \leq 50 \\ 0.87 + 0.0055 \cdot (x - 50) & \text{für} \quad 50 < x \leq 70 \\ 0.98 + 0.0010 \cdot (x - 70) & \text{für} \quad 70 < x \leq 90 \\ 1 & \text{für} \quad x > 90 \end{cases}$$

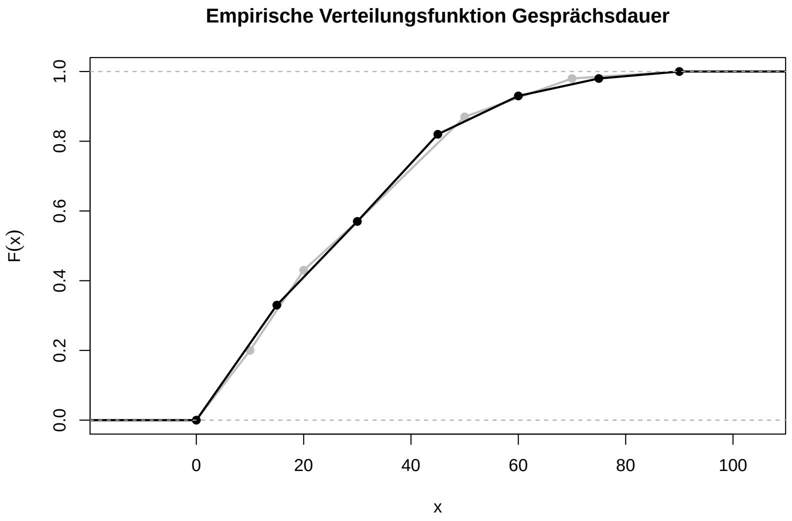
Grafik: Verteilungsfunktion bei klassierten Daten

(Empirische Verteilungsfunktion der unklassierten Daten in hellgrau)



Grafik: Verteilungsfunktion bei verschiedenen Klassierungen

(Klassierung aus Folie 51 in schwarz, Klassierung aus Folie 52 in grau)



Lagemaße

- Aggregation von Merkmalswerten zu Häufigkeitsverteilungen (auch nach erfolgter Klassierung) nicht immer ausreichend.
- Häufig gewünscht: einzelner Wert, der die Verteilung der Merkmalswerte geeignet charakterisiert $\rightsquigarrow$ „Mittelwert"
- **Aber:**
  - Gibt es immer einen „Mittelwert"?
    Was ist der Mittelwert der Merkmalswerte *rot, gelb, gelb, blau*?
    $\rightsquigarrow$ allgemeinerer Begriff: „Lagemaß"
  - Gibt es verschiedene „Mittelwerte"?
    Falls ja, welcher der Mittelwerte ist (am Besten) geeignet?

Lagemaße für nominalskalierte Merkmale

- Verschiedene Merkmalsausprägungen können lediglich unterschieden werden
- „Typische" Merkmalswerte sind also solche, die häufig vorkommen
- Geeignetes Lagemaß: häufigster Wert (es kann mehrere geben!)

Definition 3.1 (Modus, Modalwert)

Sei X ein (mindestens) nominalskaliertes Merkmal mit Merkmalsraum $A = \{a_1, \dots, a_m\}$ und relativer Häufigkeitsverteilung r.
Dann heißt jedes Element $a_{\text{mod}} \in A$ mit

$$r(a_{\text{mod}}) \geq r(a_j) \text{ für alle } j \in \{1, \dots, m\}$$

Modus oder **Modalwert** von X.

- Beispiele:
  - Modus der Urliste *rot, gelb, gelb, blau*:
    $a_{\text{mod}} = \text{gelb}$
  - Modalwerte der Urliste $1, 5, 3, 3, 4, 2, 6, 7, 6, 8$:
    $a_{\text{mod},1} = 3$ und $a_{\text{mod},2} = 6$

Lagemaße für ordinalskalierte Merkmale I

- Durch die vorgegebene Anordnung auf der Menge der möglichen Ausprägungen M lässt sich der Begriff „mittlerer Wert" mit Inhalt füllen.
- In der geordneten Folge von Merkmalswerten

$$x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)}$$

  bietet sich als Lagemaß also ein Wert „in der Mitte" der Folge an.
- Ist n gerade, gibt es keine eindeutige Mitte der Folge, und eine zusätzliche Regelung ist erforderlich.

Lagemaße für ordinalskalierte Merkmale II

Definition 3.2 (Median)

Sei X ein (mindestens) ordinalskaliertes Merkmal auf der Menge der vorstellbaren Merkmalsausprägungen M und $x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)}$ die gemäß der vorgegebenen Ordnung sortierte Urliste zum Merkmal X.

- Ist n ungerade, so heißt $x_{\left(\frac{n+1}{2}\right)}$ der **Median** von X, in Zeichen $x_{\text{med}} = x_{\left(\frac{n+1}{2}\right)}$.
- Ist n gerade, so heißen alle (möglicherweise viele verschiedene) Elemente von M *zwischen* (bezogen auf die auf M gegebene Ordnung) $x_{\left(\frac{n}{2}\right)}$ und $x_{\left(\frac{n}{2}+1\right)}$ (einschließlich dieser beiden Merkmalswerte) **Mediane** von X.

- Bei stetigen Merkmalen kann für die Definition des Medians auch für gerades n Eindeutigkeit erreicht werden, indem spezieller der Mittelwert

$$\frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

der beiden „mittleren“ Merkmalswerte als Median festgelegt wird.

Lagemaße für ordinalskalierte Merkmale III

- Beispiele:
  - Ist $M = \{\text{sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend}\}$ als Menge der möglichen Ausprägungen eines ordinalskalierten Merkmals X mit der üblichen Ordnung von Schulnoten von „sehr gut“ bis „ungenügend“ versehen, so ist die sortierte Folge von Merkmalswerten zur Urliste

    gut, ausreichend, sehr gut, mangelhaft, mangelhaft, gut

    durch

    sehr gut, gut, gut, ausreichend, mangelhaft, mangelhaft

    gegeben und sowohl „gut“ als auch „befriedigend“ und „ausreichend“ sind Mediane von X.
  - Der oben beschriebenen Konvention für stetige Merkmale folgend ist der Median des stetigen Merkmals zur Urliste

    1.85, 6.05, 7.97, 11.16, 17.19, 18.87, 19.82, 26.95, 27.25, 28.34

    von 10 Merkmalsträgern durch $x_{\text{med}} = \frac{1}{2} \cdot (17.19 + 18.87) = 18.03$ gegeben.

Lagemaße für kardinalskalierte Merkmale

- Bei kardinalskalierten Merkmalen ist oft eine „klassische“ Mittelung der Merkmalswerte als Lagemaß sinnvoll, man erhält so aus der Urliste $x_1, \dots, x_n$ das **„arithmetische Mittel“** $\overline{x} := \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- *Beispiel:*
  Die Haushalts-Nettoeinkommen (in €) von 6 Haushalten eines Mehrparteien-Wohnhauses sind:

Haushalt	1	2	3	4	5	6
Nettoeinkommen	1000	400	1500	2900	1800	2600

  Frage: Wie groß ist das durchschnittliche Nettoeinkommen?
  Antwort: $\frac{1}{6} \cdot (1000 + 400 + 1500 + 2900 + 1800 + 2600) = 1700$
- Bei klassierten Daten wird der Mittelwert als gewichtetes arithmetisches Mittel der l Klassenmitten näherungsweise berechnet:

$$\overline{x} := \frac{1}{n}\sum_{j=1}^{l} h_j \cdot m_j = \sum_{j=1}^{l} r_j \cdot m_j \ .$$

- Arithmetisches Mittel für viele (nicht alle!) Anwendungen adäquates „Mittel“
- *Beispiel:*
  Ein Wachstumssparvertrag legt folgende Zinssätze fest:

Jahr	1	2	3	4	5
Zinssatz	1.5%	1.75%	2.0%	2.5%	3.5%

  Wie groß ist der Zinssatz *im Durchschnitt*?
  - Aus Zinsrechnung bekannt: Kapital K inkl. Zinsen nach 5 Jahren bei Startkapital S beträgt

    $$K = S \cdot (1 + 0.015) \cdot (1 + 0.0175) \cdot (1 + 0.02) \cdot (1 + 0.025) \cdot (1 + 0.035)$$
  - Gesucht ist (für 5 Jahre gleichbleibender) Zinssatz R, der gleiches Endkapital K produziert, also R mit der Eigenschaft

    $$K \stackrel{!}{=} S \cdot (1 + R) \cdot (1 + R) \cdot (1 + R) \cdot (1 + R) \cdot (1 + R)$$
  - Ergebnis:
    $R = \sqrt[5]{(1 + 0.015) \cdot (1 + 0.0175) \cdot (1 + 0.02) \cdot (1 + 0.025) \cdot (1 + 0.035)} - 1$
    $\rightsquigarrow R = 2.2476\%$.
- Der in diesem Beispiel für die Bruttorenditen (1+Zinssatz) sinnvolle Mittelwert heißt **„geometrisches Mittel“**.

- *Beispiel:*
  Auf einer Autofahrt von insgesamt 30 [km] werden $s_1 = 10$ [km] mit einer Geschwindigkeit von $v_1 = 30$ [km/h], $s_2 = 10$ [km] mit einer Geschwindigkeit von $v_2 = 60$ [km/h] und $s_3 = 10$ [km] mit einer Geschwindigkeit von $v_3 = 120$ [km/h] zurückgelegt.
  Wie hoch ist die durchschnittliche Geschwindigkeit?
  - Durchschnittliche Geschwindigkeit: Quotient aus Gesamtstrecke und Gesamtzeit
  - Gesamtstrecke: $s_1 + s_2 + s_3 = 10\text{ [km]} + 10\text{ [km]} + 10\text{ [km]} = 30\text{ [km]}$
  - Zeit für Streckenabschnitt: Quotient aus Streckenlänge und Geschwindigkeit
  - Einzelzeiten also:

$$\frac{s_1}{v_1} = \frac{10\text{ [km]}}{30\text{ [km/h]}}, \quad \frac{s_2}{v_2} = \frac{10\text{ [km]}}{60\text{ [km/h]}} \text{ und } \frac{s_3}{v_3} = \frac{10\text{ [km]}}{120\text{ [km/h]}}$$

  $\leadsto$ Durchschnittsgeschwindigkeit

$$\frac{s_1+s_2+s_3}{\frac{s_1}{v_1}+\frac{s_2}{v_2}+\frac{s_3}{v_3}} = \frac{30\text{ [km]}}{\frac{10}{30}\text{ [h]} + \frac{10}{60}\text{ [h]} + \frac{10}{120}\text{ [h]}} = \frac{30}{\frac{7}{12}}\text{ [km/h]} = 51.429\text{ [km/h]}$$

- Der in diesem Beispiel für die Geschwindigkeiten sinnvolle Mittelwert heißt **„harmonisches Mittel“**.

Zusammenfassung: Mittelwerte I

Definition 3.3 (Mittelwerte)

Seien $x_1, x_2, \ldots, x_n$ die Merkmalswerte zu einem kardinalskalierten Merkmal X. Dann heißt

1. $\overline{x} := \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ das arithmetische Mittel,
2. $\overline{x}^{(g)} := \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^n x_i} = \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}$ das geometrische Mittel,
3. $\overline{x}^{(h)} := \dfrac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)} = \dfrac{1}{\frac{1}{n}\sum_{i=1}^n \frac{1}{x_i}}$ das harmonische Mittel

von $x_1, \ldots, x_n$.

Zusammenfassung: Mittelwerte II

Bemerkung 3.4

Liegt die absolute (bzw. relative) Häufigkeitsverteilung h (bzw. r) eines kardinalskalierten Merkmals X mit Merkmalsraum $A = \{a_1, \ldots, a_m\}$ vor, so gilt

1. $\overline{x} = \frac{1}{n}\sum_{j=1}^m h(a_j) \cdot a_j = \sum_{j=1}^m r(a_j) \cdot a_j$
2. $\overline{x}^{(g)} = \sqrt[n]{\prod_{j=1}^m a_j^{h(a_j)}} = \prod_{j=1}^m a_j^{r(a_j)}$
3. $\overline{x}^{(h)} = \dfrac{1}{\frac{1}{n}\sum_{j=1}^m \frac{h(a_j)}{a_j}} = \dfrac{n}{\sum_{j=1}^m \frac{h(a_j)}{a_j}} = \dfrac{1}{\sum_{j=1}^m \frac{r(a_j)}{a_j}}$

- Die in Bemerkung 3.4 berechneten Mittelwerte können als sogenannte *gewichtete Mittelwerte* der aufgetreten Merkmalswerte $a_1, \ldots, a_m$ aufgefasst werden, wobei die Gewichte durch die absoluten Häufigkeiten $h(a_1), \ldots, h(a_m)$ (bzw. durch die relativen Häufigkeiten $r(a_1), \ldots, r(a_m)$) der aufgetretenen Merkmalswerte gegeben sind.

Weitere Beispiele I

- Pauschale Aussagen, wann welcher Mittelwert geeignet ist, nicht möglich!
- *Beispiel Zinssätze:*
  Aufgrund Begrenzungen der europäischen Einlagensicherung möchte ein Anleger Kapital von 100 000 € gleichmäßig auf 5 Banken verteilen, die für die vorgegebene Anlagedauer folgende Zinsen anbieten:

Bank	1	2	3	4	5
Zinssatz	2.5%	2.25%	2.4%	2.6%	2.55%

  Frage: Wie groß ist der durchschnittliche Zinssatz?
  Antwort: $\frac{1}{5} \cdot (2.5\% + 2.25\% + 2.4\% + 2.6\% + 2.55\%) = 2.46\%$

Weitere Beispiele II

- *Beispiel Geschwindigkeiten:*
  Auf einer Autofahrt von insgesamt 30 [Min.] Fahrzeit werden $t_1 = 10$ [Min.] mit einer Geschwindigkeit von $v_1 = 30$ [km/h], $t_2 = 10$ [Min.] mit $v_2 = 60$ [km/h] und $t_3 = 10$ [Min.] mit $v_3 = 120$ [km/h] zurückgelegt.
  Wie hoch ist die durchschnittliche Geschwindigkeit?
  - Durchschnittliche Geschwindigkeit: Quotient aus Gesamtstrecke und -zeit
  - Gesamtzeit: $t = t_1 + t_2 + t_3 = 10\text{ [Min.]} + 10\text{ [Min.]} + 10\text{ [Min.]} = 30\text{ [Min.]}$
  - Länge der Streckenabschnitte: Produkt aus Geschwindigkeit und Fahrzeit
  - ⇝ Durchschnittsgeschwindigkeit

$$\frac{v_1 \cdot t_1 + v_2 \cdot t_2 + v_3 \cdot t_3}{t} = \frac{1}{3}\cdot 30\text{ [km/h]} + \frac{1}{3}\cdot 60\text{ [km/h]} + \frac{1}{3}\cdot 120\text{ [km/h]} = 70\text{ [km/h]}$$

Bemerkungen I

- Insbesondere bei diskreten Merkmalen wie z.B. einer Anzahl muss der erhaltene (arithmetische, geometrische, harmonische) Mittelwert weder zum Merkmalsraum A noch zur Menge der vorstellbaren Merkmalsausprägungen M gehören (z.B. „im Durchschnitt 2.2 Kinder pro Haushalt“).
- Auch der/die Median(e) gehören (insbesondere bei numerischen Merkmalen) häufiger nicht zur Menge A der Merkmalsausprägungen; lediglich der/die Modalwert(e) kommen stets auch in der Liste der Merkmalswerte vor!
- **Vorsicht** vor falschen Rückschlüssen vom Mittelwert auf die Häufigkeitsverteilung!

Bemerkungen II

Mobilfunknutzung Europa in 2006

In einem Online-Artikel der Zeitschrift „Computerwoche“ vom 03.04.2007 (siehe `http://www.computerwoche.de/a/statistik-jeder-europaeer-telefoniert-mobil,590888`) wird aus der Tatsache, dass die Anzahl der Mobiltelefone in Europa größer ist als die Anzahl der Europäer, also das arithmetische Mittel des Merkmals *Anzahl Mobiltelefone pro Person* in Europa größer als 1 ist, die folgende Aussage in der Überschrift abgeleitet:

Statistik: Jeder Europäer telefoniert mobil

Zusammenfassend heißt es außerdem:

Laut einer aktuellen Studie telefoniert jeder Europäer mittlerweile mit mindestens einem Mobiltelefon.

Wie sind diese Aussagen zu beurteilen? Welcher Fehlschluss ist gezogen worden?

Optimalitätseigenschaften

einiger Lagemaße bei kardinalskalierten Daten

- Für kardinalskalierte Merkmale besitzen Mediane und arithmetische Mittelwerte spezielle (Optimalitäts-)Eigenschaften.
- Für jeden Median x_{med} eines Merkmals X mit den n Merkmalswerten $x_1, \ldots, x_n$ gilt:

$$\sum_{i=1}^{n} |x_i - x_{\text{med}}| \leq \sum_{i=1}^{n} |x_i - t| \quad \text{für alle } t \in \mathbb{R}$$

- Für das arithmetische Mittel $\overline{x}$ eines Merkmals X mit den n Merkmalswerten $x_1, \ldots, x_n$ gilt:
  1. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$
  2. $\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - t)^2$ für alle $t \in \mathbb{R}$

Weitere Lagemaße: Quantile/Perzentile I

- Für jeden Median x_{med} gilt: Mindestens 50% der Merkmalswerte sind kleiner gleich x_{med} und ebenso mindestens 50% größer gleich x_{med}.
- Verallgemeinerung dieser Eigenschaft auf beliebige Anteile geläufig, also auf Werte, zu denen mindestens ein Anteil p kleiner gleich und ein Anteil $1-p$ größer gleich ist, sog. p-**Quantilen** (auch p-**Perzentile**) x_p.
- Mediane sind dann gleichbedeutend mit 50%-Quantilen bzw. 0.5-Quantilen, es gilt also insbesondere bei eindeutigen Medianen

$$x_{med} = x_{0.5} \ .$$

Weitere Lagemaße: Quantile/Perzentile II

Definition 3.5 (Quantile/Perzentile, Quartile)

Sei X ein (mindestens) ordinalskaliertes Merkmal auf der Menge der vorstellbaren Merkmalsausprägungen M mit den Merkmalswerten $x_1, \dots, x_n$.
Für $0 < p < 1$ heißt jeder Wert $x_p \in M$ mit der Eigenschaft

$$\frac{\#\{i \in \{1,\dots,n\} \mid x_i \leq x_p\}}{n} \geq p \quad \text{und} \quad \frac{\#\{i \in \{1,\dots,n\} \mid x_i \geq x_p\}}{n} \geq 1-p$$

p-**Quantil** (auch p-**Perzentil**) von X. Man bezeichnet spezieller das 0.25-Quantil $x_{0.25}$ als **unteres Quartil** sowie das 0.75-Quantil $x_{0.75}$ als **oberes Quartil**.

Weitere Lagemaße: Quantile/Perzentile III

- p-Quantile kann man auch mit der emp. Verteilungsfunktion F bestimmen:
- Mit der Abkürzung

$$F(x-0) := \lim_{\substack{h \to 0 \\ h > 0}} F(x-h), \qquad x \in \mathbb{R},$$

für linksseitige Grenzwerte empirischer Verteilungsfunktionen F ist x_p ist genau dann ein p-Quantil, wenn gilt:

$$F(x_p - 0) \leq p \leq F(x_p)$$

- Spezieller ist x_p genau dann ein p-Quantil, wenn
  - bei Vorliegen der exakten Häufigkeitsverteilung r und Verteilungsfunktion F

$$F(x_p) - r(x_p) \leq p \leq F(x_p) \ ,$$

  - bei Verwendung der approximativen Verteilungsfunktion F bei klassierten Daten (wegen der Stetigkeit der Approximation!)

$$F(x_p) = p$$

gilt.

Weitere Lagemaße: Quantile/Perzentile IV

- Genauso wie der Median muss ein p-Quantil nicht eindeutig bestimmt sein.
- Bei stetigen Merkmalen kann Eindeutigkeit *zum Beispiel* durch die gängige Festlegung

$$x_p = \begin{cases} x_{(\lfloor n \cdot p \rfloor + 1)} & \text{für} \quad n \cdot p \notin \mathbb{N} \\ \frac{1}{2} \cdot \left(x_{(n \cdot p)} + x_{(n \cdot p + 1)} \right) & \text{für} \quad n \cdot p \in \mathbb{N} \end{cases}$$

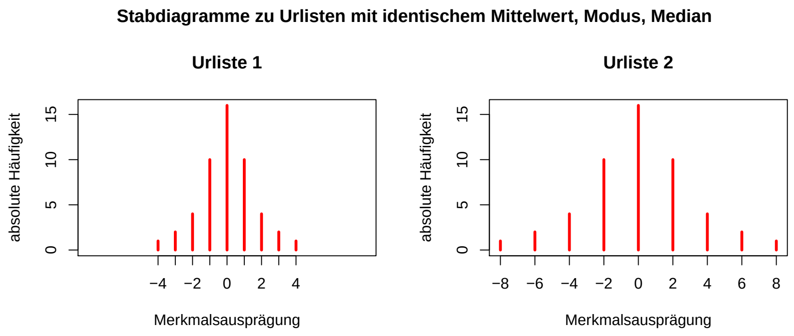
erreicht werden, wobei $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ die gemäß der vorgegebenen Ordnung sortierte Urliste ist und mit $\lfloor y \rfloor$ für $y \in \mathbb{R}$ die größte ganze Zahl kleiner gleich y bezeichnet wird.
- Zum Beispiel ist für die (bereits sortierte) Urliste

$$6.77, 7.06, 8.84, 9.98, 11.87, 12.18, 12.7, 14.92$$

der Länge $n = 8$ das 0.25-Quantil $x_{0.25}$ wegen $n \cdot p = 8 \cdot 0.25 = 2 \in \mathbb{N}$ nicht eindeutig bestimmt, sondern alle Werte $x_{0.25} \in [7.06, 8.84]$ sind 0.25-Quantile. Die eindeutige Festlegung nach obiger Konvention würde dann die „Auswahl“ $x_{0.25} = \frac{1}{2}(7.06 + 8.84) = 7.95$ treffen.

Streuungsmaße I

- Verdichtung der Merkmalswerte auf einen Lageparameter als einzige Kennzahl recht unspezifisch.
- Starke Unterschiede trotz übereinstimmender Lagemaße möglich:



Streuungsmaße II

- Bei kardinalskalierten Merkmalen: zusätzliche Kennzahl für Variation bzw. Streuung der Merkmalswerte von Interesse
- Ähnlich wie bei Lagemaßen: verschiedene Streuungsmaße gängig
- Allen Streuungsmaßen gemeinsam:
  Bezug zu „Abstand" zwischen Merkmalswerten
- *Ein* möglicher Abstand: (Betrag der) Differenz zwischen Merkmalswerten

Streuungsmaße III

Definition 3.6 (Spannweite, IQA, mittlere abs. Abweichung)

Seien $x_1, \ldots, x_n$ die Urliste zu einem kardinalskalierten Merkmal X, x_{med} der Median und $x_{0.25}$ bzw. $x_{0.75}$ das untere bzw. obere Quartil von X.
Dann heißt

1. $SP := \left(\max_{i\in\{1,\ldots,n\}} x_i\right) - \left(\min_{i\in\{1,\ldots,n\}} x_i\right) = x_{(n)} - x_{(1)}$ die **Spannweite** von X,
2. $IQA := x_{0.75} - x_{0.25}$ der **Interquartilsabstand (IQA)** von X,
3. $MAA := \frac{1}{n}\sum_{i=1}^{n} |x_i - x_{\text{med}}|$ die **mittlere absolute Abweichung** von X.

Streuungsmaße IV

- Die Betragsstriche in Teil 1 und 2 von Definition 3.6 fehlen, da sie überflüssig sind.
- Um Eindeutigkeit in Teil 2 und 3 von Definition 3.6 zu erhalten, sind die für kardinalskalierte Merkmale üblichen Konventionen zur Berechnung von Median und Quantilen aus Folie 61 bzw. 76 anzuwenden.
- Verwendung von $\overline{x}$ statt x_{med} in Teil 3 von Definition 3.6 prinzipiell möglich, aber: Beachte Folie 72!
- Weiterer möglicher Abstand: Quadrate der Differenzen zwischen Merkmalswerten

Streuungsmaße V

Definition 3.7 (empirische Varianz, empirische Standardabweichung)

Seien $x_1, \ldots, x_n$ die Urliste zu einem kardinalskalierten Merkmal X, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ das arithmetische Mittel von X. Dann heißt

1. $s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ die **(empirische) Varianz** von X,
2. die (positive) Wurzel $s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ die **(empirische) Standardabweichung** von X.

Streuungsmaße VI

- Empirische Varianz bzw. Standardabweichung sind die gebräuchlichsten Streuungsmaße.
- Standardabweichung s hat dieselbe Dimension wie die Merkmalswerte, daher i.d.R. besser zu interpretieren als Varianz.
- Für Merkmale mit positivem Mittelwert $\overline{x}$ als relatives Streuungsmaß gebräuchlich: **Variationskoeffizient** $VK := \frac{s}{\overline{x}}$
- „Rechenregeln" zur alternativen Berechnung von s bzw. s^2 vorhanden.

Satz 3.8 (Verschiebungssatz)

Seien $x_1, \ldots, x_n$ die Urliste zu einem kardinalskalierten Merkmal X, $\overline{x}$ das arithmetische Mittel und s^2 die empirische Varianz von X. Dann gilt

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

Streuungsmaße VII

- Mit der Schreibweise $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ erhält man aus Satz 3.8 die kürzere Darstellung $s^2 = \overline{x^2} - \overline{x}^2$.
- Liegt zum Merkmal X die absolute Häufigkeitsverteilung $h(a)$ bzw. die relative Häufigkeitsverteilung $r(a)$ auf der Menge der Ausprägungen $A = \{a_1, \ldots, a_m\}$ vor, so kann s^2 auch durch

$$s^2 = \frac{1}{n}\sum_{j=1}^{m} h(a_j)\cdot(a_j - \overline{x})^2 = \sum_{j=1}^{m} r(a_j)\cdot(a_j - \overline{x})^2$$

berechnet werden. (Berechnung von $\overline{x}$ dann mit Häufigkeiten als $\overline{x} = \frac{1}{n}\sum_{j=1}^{m} h(a_j)\cdot a_j = \sum_{j=1}^{m} r(a_j)\cdot a_j$, siehe Bemerkung 3.4 auf Folie 67)
- Natürlich kann alternativ auch Satz 3.8 verwendet und $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ mit Hilfe der Häufigkeitsverteilung durch

$$\overline{x^2} = \frac{1}{n}\sum_{j=1}^{m} h(a_j)\cdot a_j^2 = \sum_{j=1}^{m} r(a_j)\cdot a_j^2$$

berechnet werden.

Empirische Varianz bei klassierten Daten

- Bei klassierten Daten: auch für empirische Varianz nur Approximation möglich.
- Analog zur Berechnung von s^2 aus Häufigkeitsverteilungen:
  - Näherungsweise Berechnung von s^2 aus Klassenmitten m_j und absoluten bzw. relativen Klassenhäufigkeiten h_j bzw. r_j der l Klassen als

$$s^2 = \frac{1}{n}\sum_{j=1}^{l} h_j\cdot(m_j - \overline{x})^2 \quad \text{mit} \quad \overline{x} = \frac{1}{n}\sum_{j=1}^{l} h_j\cdot m_j$$

bzw.

$$s^2 = \sum_{j=1}^{l} r_j\cdot(m_j - \overline{x})^2 \quad \text{mit} \quad \overline{x} = \sum_{j=1}^{l} r_j\cdot m_j\ .$$

  - Alternativ: Verwendung von Satz 3.8 mit

$$\overline{x} := \frac{1}{n}\sum_{j=1}^{l} h_j\cdot m_j = \sum_{j=1}^{l} r_j\cdot m_j$$

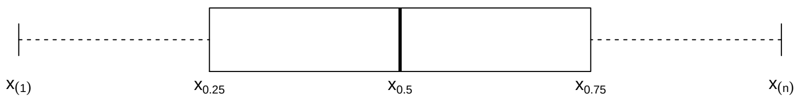
und

$$\overline{x^2} := \frac{1}{n}\sum_{j=1}^{l} h_j\cdot m_j^2 = \sum_{j=1}^{l} r_j\cdot m_j^2\ .$$

Box-and-whisker-Plot I

- Häufig von Interesse:
  Visueller Vergleich **eines** Merkmals für **verschiedene** statistische Massen
- Dazu nötig: Grafische Darstellung mit Ausdehnung (im Wesentlichen) nur in einer Dimension (2. Dimension für Nebeneinanderstellung der Datensätze)

⇝ **Box-and-whisker-Plot** oder kürzer **Box-Plot**:
Zur Urliste $x_1, \ldots, x_n$ eines kardinalskalierten Merkmals werden *im Prinzip* die 5 Kennzahlen $x_{(1)}, x_{0.25}, x_{0.5}, x_{0.75}, x_{(n)}$ in Form eines durch $x_{0.5}$ geteilten „Kästchens“ (Box) von $x_{0.25}$ bis $x_{0.75}$ und daran anschließende „Schnurrhaare“ (Whisker) bis zum kleinsten Merkmalswert $x_{(1)}$ und zum größten Merkmalswert $x_{(n)}$ dargestellt:

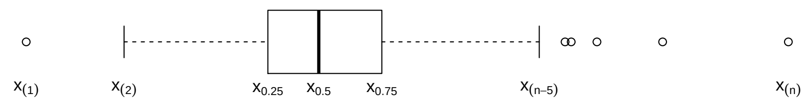


Box-and-whisker-Plot II

- (Häufig auftretende!) Ausnahme:
  $x_{(1)}$ und/oder $x_{(n)}$ liegen weiter als der 1.5-fache Interquartilsabstand (IQA) $x_{0.75} - x_{0.25}$ von der Box entfernt (also weiter als die 1.5-fache Breite der Box)

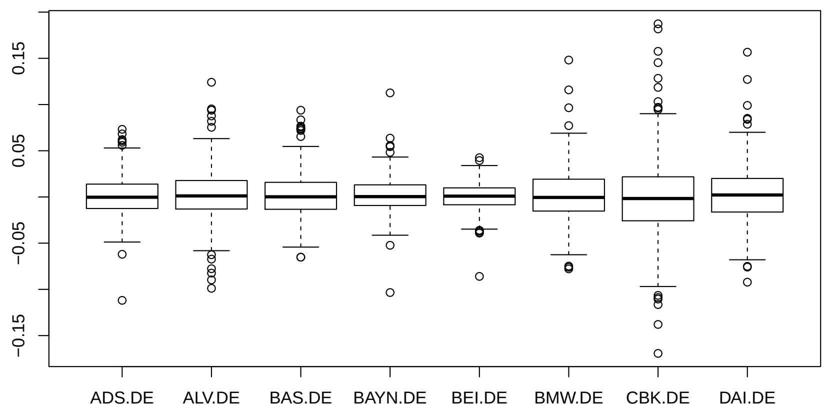
⇝ Dann: Whiskers nur bis zu äußersten Merkmalswerten innerhalb dieser Distanz und separates Eintragen der „Ausreißer“, d.h. aller Urlisteneinträge, die nicht von der Box und den Whiskers abgedeckt werden.

- Beispiel mit „Ausreißern“:



Box-and-whisker-Plot III

- Beispiel für Gegenüberstellung mehrerer Datensätze
  (Diskrete Tagesrenditen verschiedener DAX-Papiere)



Symmetrie(-maß), Schiefe I

- Neben Lage und Streuung bei kardinalskalierten Merkmalen auch interessant: **Symmetrie** (bzw. Asymmetrie oder Schiefe) und **Wölbung**
- Ein Merkmal X ist symmetrisch (um $\overline{x}$), wenn die Häufigkeitsverteilung von $X - \overline{x}$ mit der von $\overline{x} - X$ übereinstimmt.
  (Dabei ist mit $X - \overline{x}$ das Merkmal mit den Urlistenelementen $x_i - \overline{x}$ für $i \in \{1, \ldots, n\}$ bezeichnet, dies gilt analog für $\overline{x} - X$.)
- Symmetrie eines Merkmals entspricht also der Achsensymmetrie des zugehörigen Stabdiagramms um $\overline{x}$.
- Ist ein Merkmal nicht symmetrisch, ist die **empirische Schiefe** bzw. **empirische Skewness** ein geeignetes Maß für die Stärke der Asymmetrie.

Symmetrie(-maß), Schiefe II

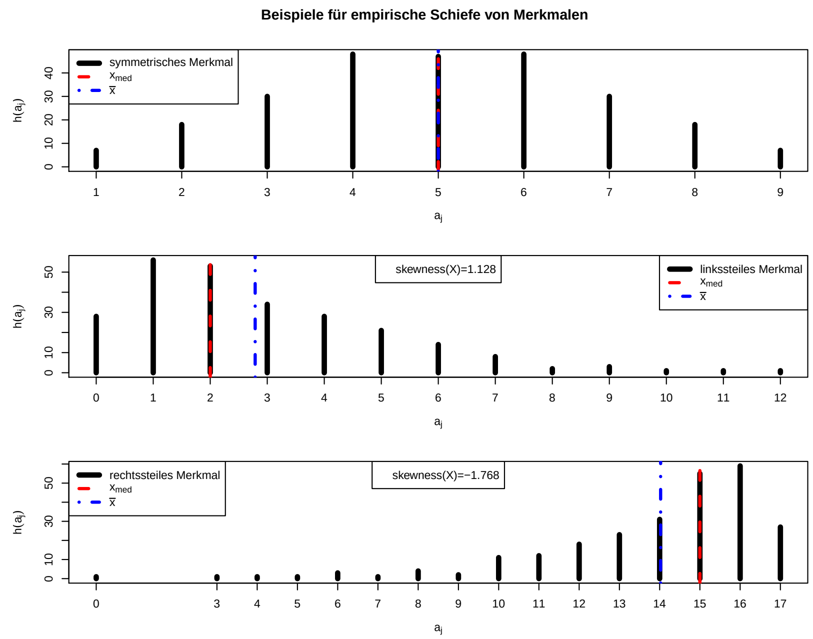
Definition 3.9 (empirische Schiefe, Skewness)

Sei X ein Merkmal mit der Urliste $x_1, \dots, x_n$. Dann heißt

$$\text{skewness}(X) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

mit $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ die **empirische Schiefe (Skewness)** von X.

- Man kann zeigen: X symmetrisch $\Rightarrow$ skewness$(X) = 0$
- X heißt **linkssteil** oder **rechtsschief**, falls skewness$(X) > 0$.
- X heißt **rechtssteil** oder **linksschief**, falls skewness$(X) < 0$.
- Für symmetrische Merkmale ist $\overline{x}$ gleichzeitig Median von X, bei linkssteilen Merkmalen gilt *tendenziell* $\overline{x} > x_{\text{med}}$, bei rechtssteilen *tendenziell* $\overline{x} < x_{\text{med}}$.



Wölbungsmaß (Kurtosis) I

Definition 3.10 (empirische Wölbung, Kurtosis)

Sei X ein Merkmal mit der Urliste $x_1, \dots, x_n$. Dann heißt

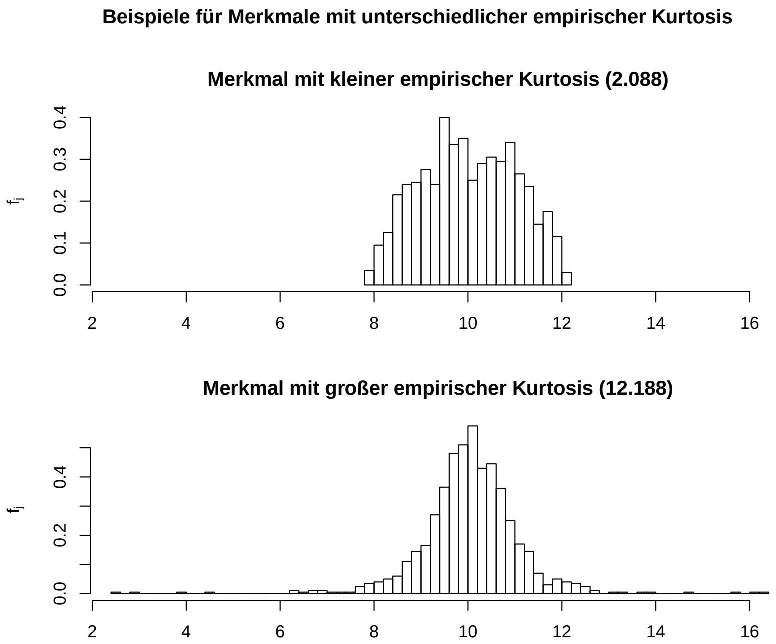
$$\text{kurtosis}(X) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4$$

mit $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ die **empirische Wölbung (Kurtosis)** von X.

- Kurtosis misst bei Merkmalen mit *einem* Modalwert, wie „flach" (kleiner Wert) bzw. „spitz" (großer Wert) der „Gipfel" um diesen Modalwert ist.

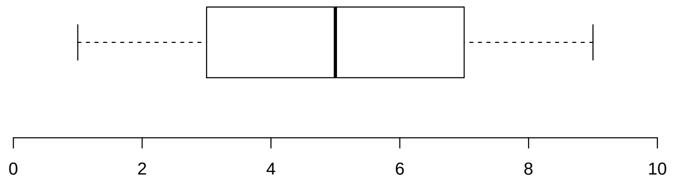
Wölbungsmaß (Kurtosis) II

- Bei gleicher mittlerer quadratischer Abweichung vom Mittelwert ($\rightsquigarrow$ Varianz) müssen Merkmale mit größerer emp. Kurtosis (mehr Werten in der Nähe des Gipfels) auch mehr weit vom Gipfel entfernte Merkmalswerte besitzen.
- Der Wert 3 wird als „normaler" Wert für die empirische Kurtosis angenommen, Merkmale mit $1 \leq$ kurtosis$(X) < 3$ heißen platykurtisch, Merkmale mit kurtosis$(X) > 3$ leptokurtisch.
- *Vorsicht:* Statt der Kurtosis von X wird oft die **Exzess-Kurtosis** von X angegeben, die der um den Wert 3 verminderten Kurtosis entspricht.



Schiefe und Wölbung in grafischen Darstellungen I

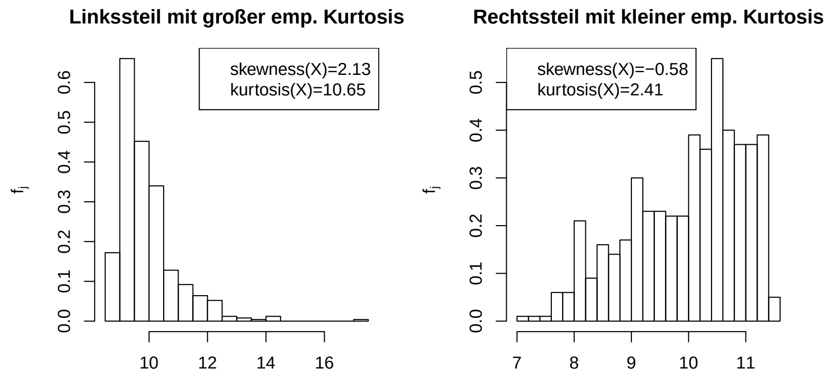
- Box-Plots lassen auch auf empirische Schiefe und Kurtosis schließen.
- Bei symmetrischen Merkmalen sind auch die Box-Plots symmetrisch.
  Beispiel: Box-Plot zur Urliste $1, 2, 3, 4, 5, 6, 7, 8, 9$:



Schiefe und Wölbung in grafischen Darstellungen II

- Bei linkssteilen Merkmalen hat *tendenziell* der rechte/obere Teil (rechter/oberer Teil der Box und rechter/oberer Whisker) eine **größere** Ausdehnung als der linke/untere Teil.
- Bei rechtssteilen Merkmalen hat *tendenziell* der rechte/obere Teil (rechter/oberer Teil der Box und rechter/oberer Whisker) eine **kleinere** Ausdehnung als der linke/untere Teil.
- Bei Merkmalen mit **großer** empirischer Kurtosis gibt es *tendenziell* **viele** „Ausreißer", also separat eingetragene Merkmalswerte außerhalb der Whiskers (wenigstens auf einer Seite).
- Bei Merkmalen mit **kleiner** empirischer Kurtosis gibt es häufig **wenige** oder **gar keine** „Ausreißer".

- Beispiele für Merkmale mit unterschiedlicher empirischer Schiefe/Kurtosis



- Zugehörige Box-Plots: