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 $\leftrightarrow$ Sekundärerhebung:
    Neue Erhebung oder Nutzung von vorhandenem Datenmaterial
  - Vollerhebung $\leftrightarrow$ 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?

⇝ 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 Elementen 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:
  x_1 = *“Die Linke”*, x_2 = *“SPD”*, x_3 = *“CDU”*, x_4 = *“Die Linke”*, x_5 = *“SPD”*, x_6 = *“SPD”*, x_7 = *“Die Linke”*, x_8 = *“CDU”*, x_9 = *“FDP”*, x_{10} = *“Grüne”*, x_{11} = *“Die Linke”*, x_{12} = *“SPD”*, x_{13} = *“Die Linke”*, x_{14} = *“CDU”*, x_{15} = *“SPD”*, x_{16} = *“CDU”*, x_{17} = *“CDU”*, x_{18} = *“SPD”*, x_{19} = *“SPD”*, x_{20} = *“FDP”*, ...
- 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“, „divers“), 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

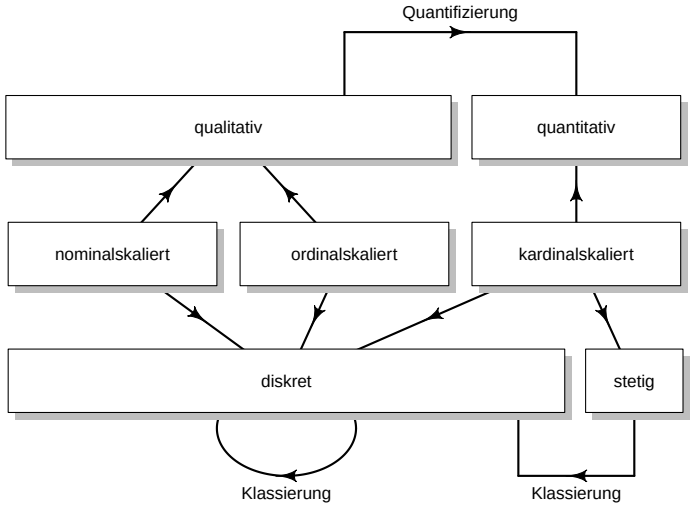
Umwandlung von Merkmalstypen I

- 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 entsprechen, 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 vernünftig, 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

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

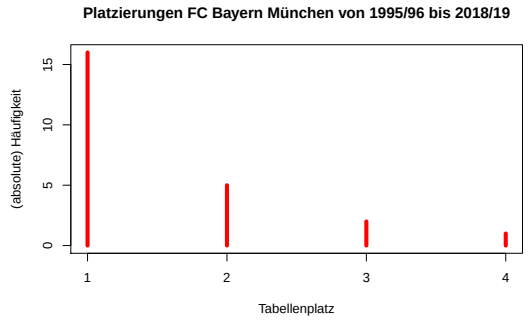
- Alternative grafische Darstellung bei (mindestens) ordinalskalierten Merkmalen mit numerischen Ausprägungen: **Stabdiagramm**

Häufigkeitsverteilungen IV

- Stabdiagramm zur Urliste

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

der finalen Tabellenplätze des FC Bayern München in der (ersten) Fußball-Bundesliga (Saison 1995/96 bis 2018/2019):



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

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

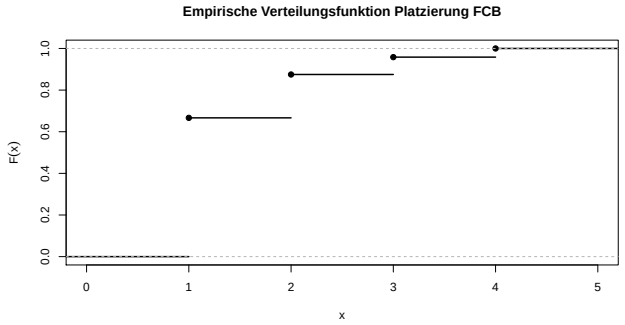
- Ein Vergleich mit den Definitionen von $h(a)$ und $r(a)$ offenbart (!), dass $F(x)$ auch mit Hilfe von $h(a)$ bzw. $r(a)$ berechnet werden kann; gibt es m Merkmalsausprägungen, so gilt:

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

- Beispiel: Empirische Verteilungsfunktion für FC Bayern-Platzierungen

$$F(x) = \begin{cases} 0 & \text{für} \quad x < 1 \\ \frac{16}{24} & \text{für} \quad 1 \leq x < 2 \\ \frac{21}{24} & \text{für} \quad 2 \leq x < 3 \\ \frac{23}{24} & \text{für} \quad 3 \leq x < 4 \\ 1 & \text{für} \quad x \geq 4 \end{cases} \approx \begin{cases} 0.000 & \text{für} \quad x < 1 \\ 0.667 & \text{für} \quad 1 \leq x < 2 \\ 0.875 & \text{für} \quad 2 \leq x < 3 \\ 0.958 & \text{für} \quad 3 \leq x < 4 \\ 1.000 & \text{für} \quad x \geq 4 \end{cases}$$

- 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

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

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: