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 $k_0, k_1, \ldots, k_l$ von l (rechtsseitig abgeschlossenen) Intervallen

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

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

Häufigkeitsverteilungen klassierter Daten III

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

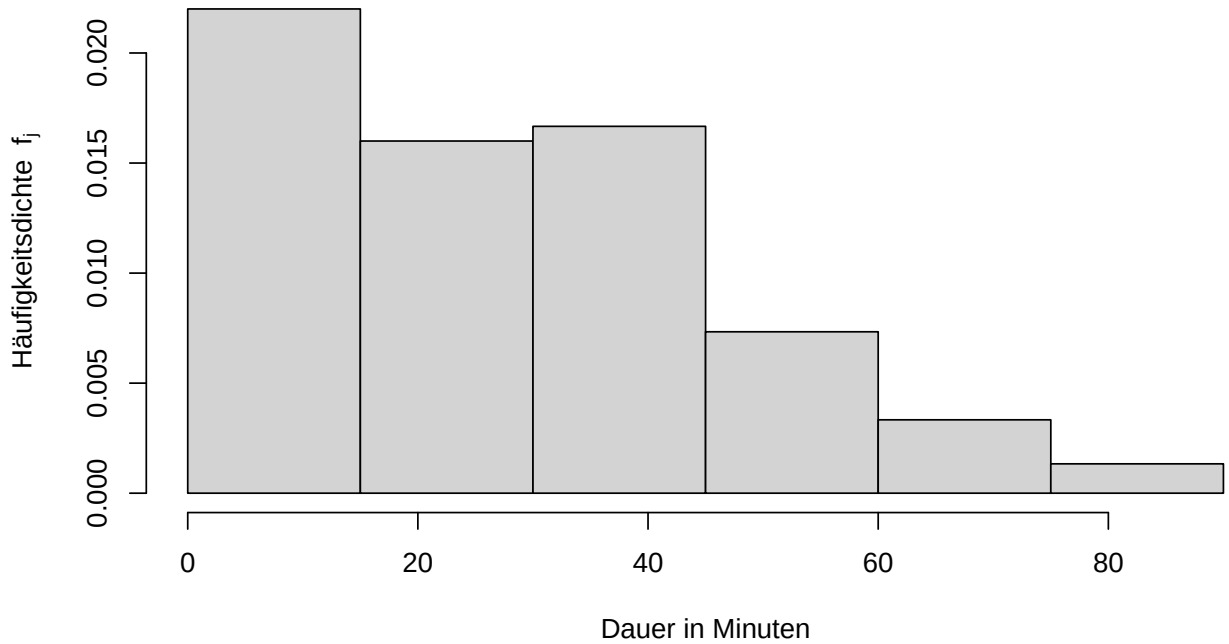
  Klassenbreiten $b_j := k_j - k_{j-1}$
  Klassenmitten $m_j := \frac{k_{j-1}+k_j}{2}$
  absolute Häufigkeiten $h_j := \#\{i \in \{1,\ldots,n\} \mid k_{j-1} < x_i \leq k_j\}$
  relative Häufigkeiten $r_j := \frac{h_j}{n}$
  Häufigkeitsdichten $f_j := \frac{r_j}{b_j}$

  (jeweils für $j \in \{1,\ldots,l\}$).
- 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:

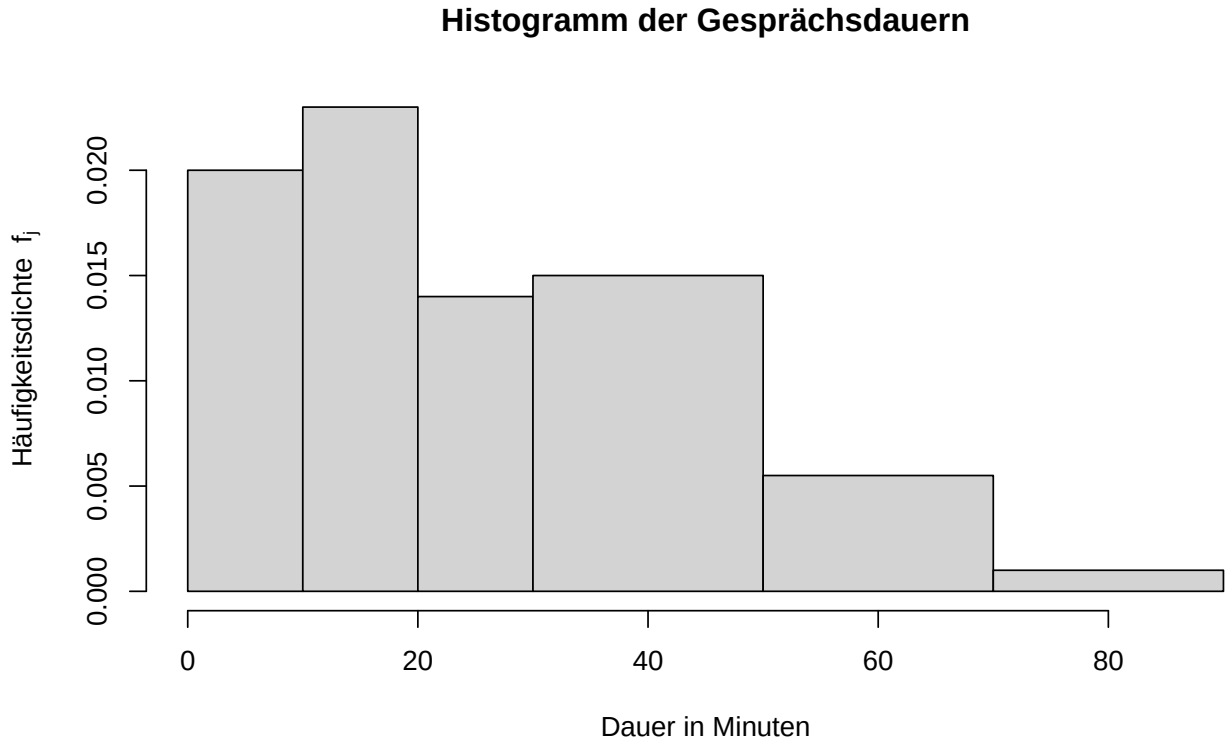
Nr. j	Klasse $K_j = (k_{j-1}, k_j]$	Klassenbreite b_j	Klassenmitte m_j	absolute Häufigkeit h_j	relative Häufigkeit $r_j = \frac{h_j}{n}$	Häufigkeitsdichte $f_j = \frac{r_j}{b_j}$	Verteilungsfunktion $F(k_j)$
1	$(0, 15]$	15	7.5	33	0.33	0.022	0.33
2	$(15, 30]$	15	22.5	24	0.24	0.016	0.57
3	$(30, 45]$	15	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

Histogramm der Gesprächsdauern



- Alternativ mit 6 Klassen bei 2 verschiedenen Breiten:

Nr. j	Klasse $K_j = (k_{j-1}, k_j]$	Klassenbreite b_j	Klassenmitte m_j	absolute Häufigkeit h_j	relative Häufigkeit $r_j = \frac{h_j}{n}$	Häufigkeitsdichte $f_j = \frac{r_j}{b_j}$	Verteilungsfunktion $F(k_j)$
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

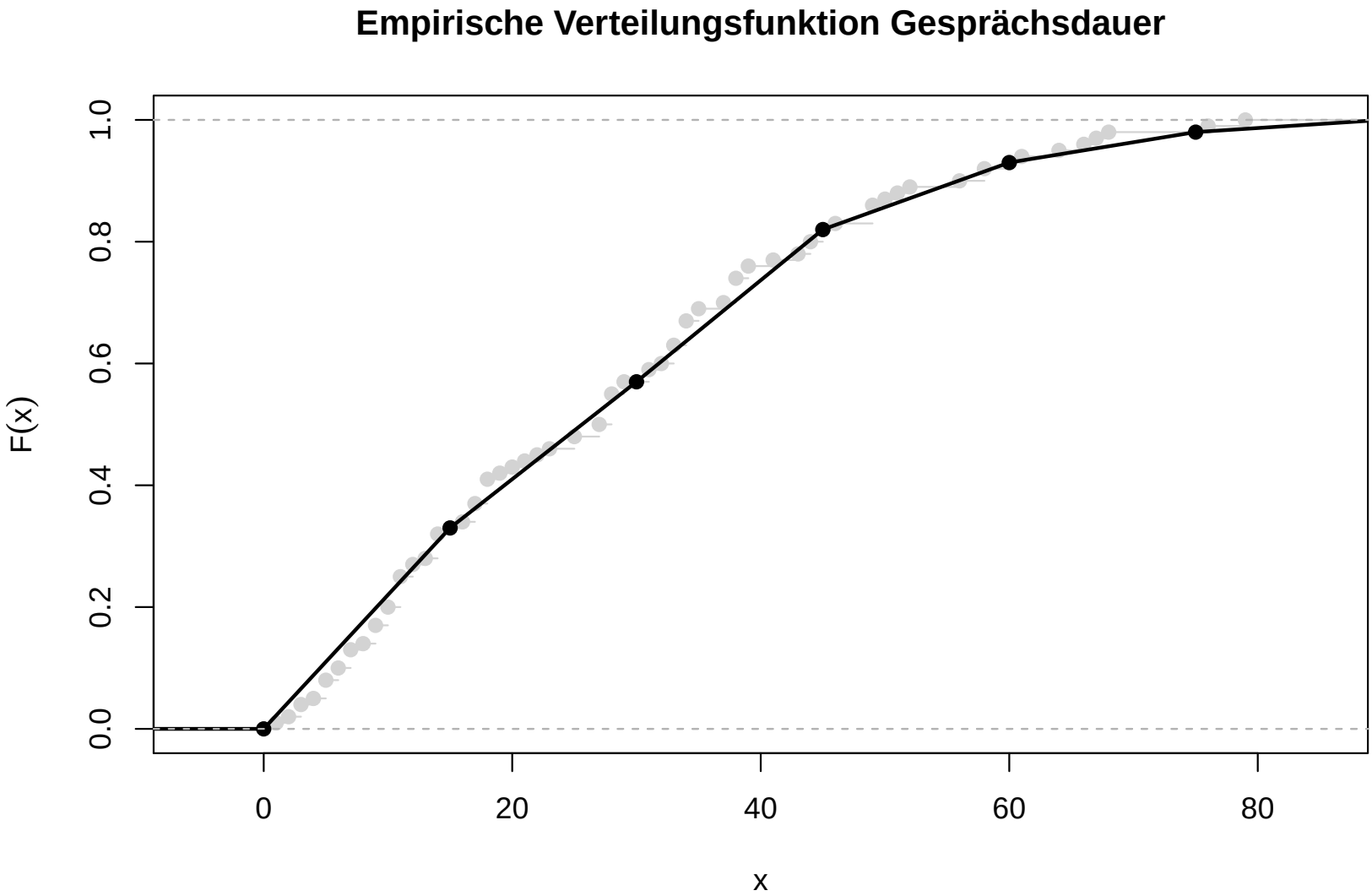
$$F(x) = \begin{cases} 0 & \text{für} \quad x \leq k_0 \\ F(k_{j-1}) + f_j \cdot (x - k_{j-1}) & \text{für} \quad k_{j-1} < x \leq k_j,\ j \in \{1, \ldots, l\} \\ 1 & \text{für} \quad x > k_l \end{cases}$$

- Am Beispiel der Gesprächdauern (Klassierung aus Folie 52)

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

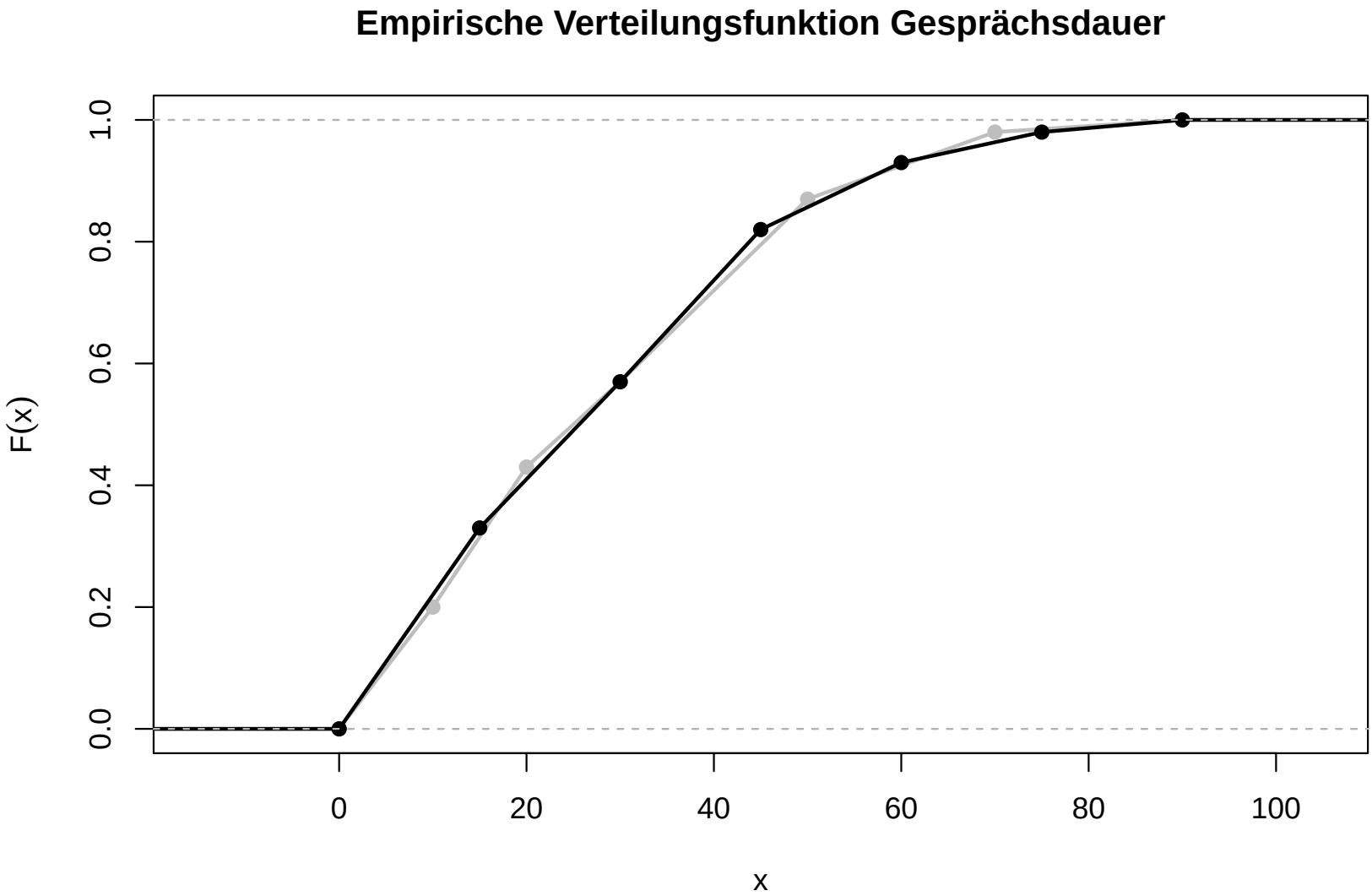
Grafik: Verteilungsfunktion bei klassierten Daten

(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 $A = \{a_1, \ldots, a_m\}$ und relativer Häufigkeitsverteilung r.
Dann heißt jedes Element $a_{mod} \in A$ mit

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

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

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

  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 $x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)}$ 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 $x_{med} = x_{\left(\frac{n+1}{2}\right)}$.*
- *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 $M = \{\text{sehr gut}, \text{gut}, \text{befriedigend}, \text{ausreichend}, \text{mangelhaft}, \text{ungenügend}\}$ 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 $x_1, \ldots, x_n$ 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:
    $R = \sqrt[5]{(1 + 0.015) \cdot (1 + 0.0175) \cdot (1 + 0.02) \cdot (1 + 0.025) \cdot (1 + 0.035)} - 1$
    $\rightsquigarrow \quad R = 2.2476\%.$
- Der in diesem Beispiel für die Zinsfaktoren (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:

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

  ⇝ 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 3.1

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 3.1 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 von Begrenzungen bei der Einlagensicherung möchte ein Anleger Kapital von 500 000 € 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

  $\rightsquigarrow$ 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 $t \in \mathbb{R}$