Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Kennwerte von Häufigkeitsverteilungen

- Häufig ist man nicht primär an der Untersuchung der Häufigkeitsverteilung einer Urliste interessiert
- Vielmehr will man zur Beantwortung spezifischer Fragestellungen die Daten durch eine oder mehrere Größen adäquat beschreiben
- Derartige Größen werden als *Kennwerte* oder auch *Statistiken* einer Stichprobe bezeichnet
- Die Kennwerte können dabei für eine Fragestellung sinnvoll, für andere aber völlig ungeeignet sein

Kennwerte von Häufigkeitsverteilungen

Man unterscheidet zwei Klassen von Kennwerten:

Maße der zentralen Tendenz ("Lage" der Stichprobenwerte)

- Modus
- Median
- Quantil
- Arithmetisches Mittel
- Geometrischer Mittelwert
- Harmonisches Mittel

Streuungsmaße (Dispersionsmaße)

- Spannweite
- Quartilabstand
- Mittlere absolute Abweichungen
- Varianz bzw. Standardabweichung

Modus

Der Modus kennzeichnet den häufigsten Wert

$$x_{mod} = a_k \text{ gdw. } f(a_k) \geq f(a_j) \text{ für alle } j = 1, \ldots, m,$$

wenn die Verteilung unimodal ist und dieser Wert eindeutig bestimmt ist

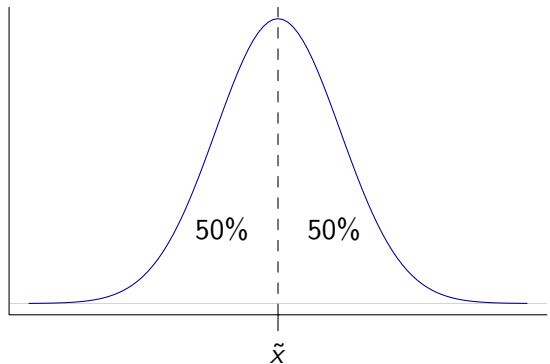
Beispiel: Noten in Klassenarbeit

$$x_{mod} = 3$$

Einzig sinnvolle Beschreibung der zentralen Tendenz bei qualitativen (nominalen) Messwerten (z. B. Partei mit den meisten Stimmen als Wahlsieger)

Median (Zentralwert)

Mindestens die Hälfte der Werte ist kleiner oder gleich $\tilde{x}$, und mindestens die Hälfte der Werte ist größer oder gleich $\tilde{x}$



Median (Zentralwert)

Für die geordnete Liste von Stichprobenwerten (kenntlich gemacht durch Indizes in Klammern)

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

definiert man

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] & \text{falls } n \text{ gerade} \end{cases}$$

Median (Zentralwert)

Beispiele

- Für die Stichprobenwerte

  3 6 8 9 11

  ist $\tilde{x} = x_{(3)} = 8$

- Für die Stichprobenwerte

  3 6 8 9 11 15

  ist $\tilde{x} = \frac{1}{2}[x_{(3)} + x_{(4)}] = 8.5$

Median (Zentralwert)

Beispiele

- Berechnung mit R

```
x <- c(3, 6, 8, 9, 11)
median(x)
[1] 8
```

und

```
x <- c(3, 6, 8, 9, 11, 15)
median(x)
[1] 8.5
```

Median (Zentralwert)

Minimumeigenschaft

- Haben die Beobachtungswerte metrisches Skalenniveau, dann ist die Summe der absoluten Abweichungen

$$\sum_{i=1}^{n} |x_i - a|$$

  minimal für $a = \tilde{x}$

- Zum Nachweis dieser Eigenschaft ist das Minimum der angegebenen Summe in Abhängigkeit von a zu bestimmen, was grundsätzlich folgende Vorgehensweise erfordert

Median (Zentralwert)

Nachweis der Minimumeigenschaft

- Ableiten der Summe nach a und Bestimmung desjenigen Wertes, für den die Ableitung Null wird
- Wir betrachten die Summe

$$\sum_{i=1}^{n} |x_i - a|$$

als Funktion von a mit

$$|x_i - a| = \begin{cases} x_i - a & \text{für } x_i > a \\ 0 & \text{für } x_i = a \\ a - x_i & \text{für } x_i < a \end{cases}$$

Median (Zentralwert)

Nachweis der Minimumeigenschaft

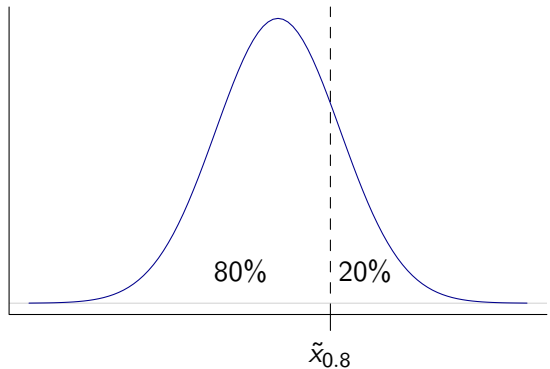
- Für die Ableitung von $|x_i - a|$ nach a gilt dann

$$\frac{d\,|x_i - a|}{d\,a} = \begin{cases} -1 & \text{für } x_i > a \\ +1 & \text{für } x_i < a \end{cases}$$

- Die Summe ist genau dann minimal, wenn deren Ableitung gleich Null ist, wenn also die beiden Fälle $x_i > a$ und $x_i < a$ jeweils gleich häufig vorkommen
- Daher minimiert der Median $a = \tilde{x}$ die Summe der absoluten Abweichungen

p-Quantil (Anteil)

Ein Anteil von mindestens $p \cdot 100\%$ Werten ist kleiner oder gleich $\tilde{x}_p$, und ein Anteil von mindestens $(1-p) \cdot 100\%$ Werten ist größer oder gleich $\tilde{x}_p$



p-Quantil (Anteil)

Für die geordnete Liste von Stichprobenwerten

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

definiert man

$$\tilde{x}_p = \begin{cases} x_{(k)} & \text{falls } n \cdot p \text{ nicht ganzzahlig} \\ & \text{und } k \text{ die nächstgrößere ganze Zahl} \\ \frac{1}{2}[x_{(k)} + x_{(k+1)}] & \text{falls } k = n \cdot p \text{ ganzzahlig} \end{cases}$$

p-Quantil (Anteil)

Beispiele

- Stichprobenwerte

$$3 \quad 6 \quad 8 \quad 9 \quad 11 \quad 15$$

- Für $p = 1/4$ ist $n \cdot p = 1.5$ und daher

$$\tilde{x}_{1/4} = x_{(2)} = 6$$

- Für $p = 2/3$ ist $n \cdot p = 4$ und daher

$$\tilde{x}_{2/3} = \tfrac{1}{2}[x_{(4)} + x_{(5)}] = 10$$

p-Quantil (Anteil)

Beispiele

- Berechnung mit R

```
x <- c(3, 6, 8, 9, 11, 15)
quantile(x, probs = c(1/4, 2/3), type = 2)
     25% 66.66667%
       6        10
```

- Die Bestimmung von Quantilen erfolgt in der Literatur in unterschiedlicher Weise, die in der Funktion `quantile()` durch die Spezifierung des Arguments `type` (Werte 1 bis 9) implementiert sind
- Mit `type = 2` wird der zu obiger Definition korrespondierende Fall ausgewählt

p-Quantil (Anteil)

Spezielle Quantile

- Quartile

$$Q_1 = \tilde{x}_{0.25}, \quad Q_2 = \tilde{x}_{0.5} = \tilde{x}, \quad Q_3 = \tilde{x}_{0.75}$$

  Die Differenz $Q_{dif} = Q_3 - Q_1$ wird als Quartilabstand bezeichnet

- Dezile

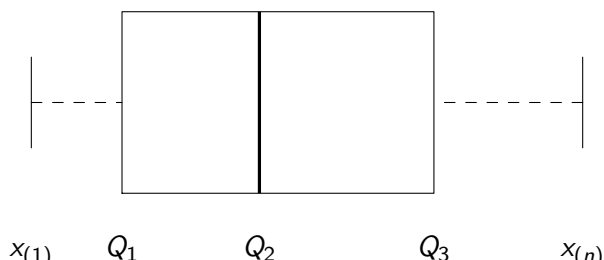
$$D_1 = \tilde{x}_{0.1}, \quad D_2 = \tilde{x}_{0.2}, \quad \ldots$$

- Perzentile

$$P_1 = \tilde{x}_{0.01}, \quad P_2 = \tilde{x}_{0.02}, \quad \ldots$$

Box-Plot

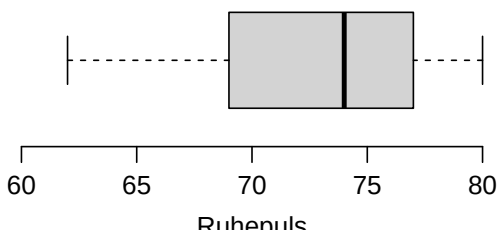
In einem Box-Plot (oder Box-and-Whisker-Plot) werden neben dem minimalen und dem maximalen Beobachtungswert auch die Quartile dargestellt



Box-Plot

Beispiel: Ruhepuls von 12 Studentinnen

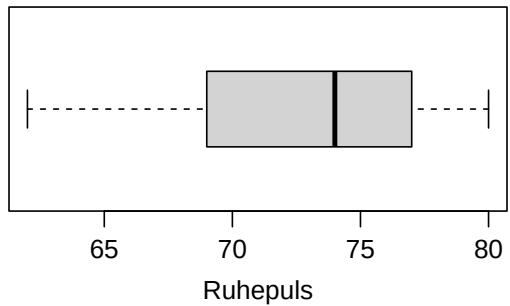
$$\mathbf{x} = (62, 64, 68, 70, 70, 74, 74, 76, 76, 78, 78, 80)$$



Box-Plot

Erstellen der Grafik in R

```
x <- c(62,64,68,70,70,74,74,76,76,78,78,80)
boxplot(x, horizontal = TRUE, range = 0, xlab = "Ruhepuls")
```



Box-Plot

Erstellen der Grafik in R

- Mit dem Argument `range` kann man auch spezifizieren, dass die "Whiskers" nicht den minimalen und maximalen Beobachtungswert einschließen, wie sie dies für `range = 0` tun
- Für einen positiven Wert von `range` schließen die "Whiskers" die Beobachtungswerte ein, die nicht mehr als das `range`-fache des Quartilabstands $Q_3 - Q_1$ von der Box entfernt liegen (bei `range = 1` also um nicht mehr als der Quartilabstand)
- Der voreingestellte Wert ("default") ist `range = 1.5`

Box-Plot

Erstellen der Grafik in R

```
y <- c(58, x, 89)  # zwei zusaetzliche Beobachtungen
boxplot(y, horizontal = TRUE, range = 1, xlab = "Ruhepuls",
        main = "Erweiterter Datensatz")
rug(y)             # markieren der Beobachtungen
```

