WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Einführung in die Quantitative Datenanalyse

Sitzung 7: Maße der zentralen Tendenz und Variabilität

Proseminar an der Freien Universität Berlin
25.06.2017 - Marcus Spittler

Inhalt der 7. Sitzung

1. Maße der zentralen Tendenz
   - Modus
   - Median
   - (Arithmetisches) Mittel
2. Maße der Variabilität
   - Varianz
   - Standardabweichung
   - Variationsbreite
   - Variationskoeffizient
3. Form von Verteilungen
   - Symmetrie
   - Wölbung

Margaret Hamilton (1s969) mit dem Code der Apollo 11 Mission

Übersicht

Zulässige Berechnungen ab jeweiligem Skalenniveau:

Skalenniveau	Zentrale Tendenz	Variabilität
Nominalskala	Modus	Entropie
Ordinalskala	Median	Summenhäufigkeitsentropie
ab Intervallsk.	- Arithmetisches Mittel	- Standardabweichung
	- Harmonisches Mittel	- Varianz
	- Geometrisches Mittel	- Variationsbreite

Arithmetisches Mittel

- Arithmetisches Mittel wird mit $\bar{x}$ (*mean*) bezeichnet
- Entspricht dem **Schwerpunkt**/Zentrum der Verteilung
- Ab *intervallskaliertem* Merkmal

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

- Empfindlich gegenüber Ausreißern/Extremwerten

```
mean( c( 800,1100,1500,2500,2800, 3000) )
```

```
## [1] 1950
```

```
mean( c( 800,1100,1500,2500,2800, 3000, 80000) )
```

```
## [1] 13100
```

Arithmetisches Mittel

- Eigenschaften des **arithmetischen Mittels**
  1. Summe der Abweichungen vom Mittel ergibt immer *null*

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

  2. Summe der quadrierten Abweichungen vom Mittel ergibt immer ein *Minimum*. Diese Eigenschaft macht man sich bei der *Methode der kleinsten Quadrate* / *Least squares* zunutze

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

Modus

- **Modus/Modalwert** Mo (*Mode*)
  - Häufigster Wert einer Verteilung
  - Für jedes Skalenniveau geeignet
  - Sollten zwei verschiedene Werte die selbe Häufigkeit haben, spricht man von einer *bimodalen Verteilung*

```
icecream <- c("chocolate", "vanilla", "strawberry",
              "vanilla", "strawberry", "vanilla",
              "chocolate", "chocolate", "chocolate")
table(icecream)
```

```
## icecream
##  chocolate strawberry    vanilla
##          4          2          3
```

Median

- **Median** Md (*median*)
  - Der Median einer Stichprobe von Werten ist definiert als der Wert, der größer gleich 50% der Werte der Stichprobe ist.
  - $x_{MD} = x_{(0.5)} = min\{x_i \mid F(x_i) \geq 0.5\}$
  - Kennzeichnet die *Mitte* der Stichprobenwerte
  - Ab *ordinalskaliertem* Merkmal
  - Wichtige Eigenschaft: Robust gegen Ausreißer

Berechnung des Median

- Erster Schritt: Anordnung der Daten nach Größe geordnet (Rangreihe) bzw. theoretisch plausibler Rangordnung
- bei **ungeradem** n = Median $Md = Rangplatz(n+1)/2$
- bei **quantitativen** Merkmalen und **geradem** n hier definiert als das arithmetische Mittel zwischen oberem und unterem Rangplatz: $Md = x_{bar}$ von $x_{Rangplatz:n/2}$ und $x_{Rangplatz:(n/2)+1}$
- bei **ordinalskalierten** Merkmalen und **geradem** n ist der MD der untere Rangplatz: $Md = x_{Rangplatz:n/2}$
- Die Berechnung des Medians bei **geradem** n ist uneindeutig, hier ist Vorschlag präsentiert, es gibt jedoch verschiedene Methoden.

Median für ordinale Merkmale

- Beispiel: Wir haben 14 Menschen in einem Fast Food Restaurant bei ihrer Bestelllung beobachtet. Dabei haben wir erhoben, welche Größe das von ihnen bestellte Menü hatte.

- Die Größe des Menus haben wir in der Reihenfolge der Bestellungen notiert, z.B.:

```
menus <- c("sehr klein", "groß", "groß", "groß",
           "sehr groß", "mittel", "sehr klein",
           usw. ... )
```

Median für ordinale Merkmale

- Beispiel: Wir haben 14 Menschen in einem Fast Food Restaurant bei ihrer Bestelllung beobachtet. Dabei haben wir erhoben, welche Größe das von ihnen bestellte Menü hatte.

- Die Größe des Menus haben wir in der Reihenfolge der Bestellungen notiert, z.B.:

```
menus <- c("sehr klein", "groß", "groß", "groß",
           "sehr groß", "mittel", "sehr klein",
           usw. ... )
```

- Die Größe der Menus ist ein **ordinalskaliertes** Merkmal. Die Anzahl der beobachteten Personen n ist mit 14 **gerade**.

Median für ordinale Merkmale

Zuerst legen wir wie gewohnt einen neuen Vektor mit den Daten an

```
menus <- c(rep("sehr klein",3), rep("groß", 5), "mittel",
           rep("klein",4), "sehr groß")
table(menus)
```

```
## menus
##       groß      klein     mittel  sehr groß sehr klein
##          5          4          1          1          3
```

Danach "ordnen" wir den Vektor um für R die Reihenfolge festzulegen

```
menus <- ordered(menus, levels = c("sehr klein", "klein", "mittel",
                                   "groß", "sehr groß"))
table(menus)
```

```
## menus
## sehr klein      klein     mittel       groß  sehr groß
##          3          4          1          5          1
```

Median für ordinale Merkmale

```
# Package "DescTools" für die Berechnung des Median
# bei ordinalskalierten Variablen
library(DescTools)
Median(menus)
```

```
## [1] klein
## Levels: sehr klein < klein < mittel < groß < sehr groß
```

Der Median unserer Variable `menus` liegt bei "klein". Das heißt, die unteren 50% der Besteller haben ein "kleines" oder "sehr kleines" Menu bestellt.

Median bei Intervallskala

```
income <- c( 800,1100,1500,2500,2800, 3000, 80000)
summary(income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     800    1300    2500   13100    2900   80000
```

```
median(income)
```

```
## [1] 2500
```

```
mean(income)
```

```
## [1] 13100
```

Maße der Variabilität

- Während Maße der zentralen Tendenz uns Auskunft über die Mitte, bzw. das Zentrum der Werte liefern, informieren uns Maße der Variablität über die **Unterschiedlichkeit** der Werte.

Varianz und Std.Abweichung

- **(Stichproben-) Varianz** s^2 (*variance*)
  - Def.: Die Stichprobenvarianz ist die Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel, dividiert durch $n-1$.
  - $n-1$ bezeichnet man als **Freiheitsgrade**
  - Ab *intervallskaliertem* Merkmal

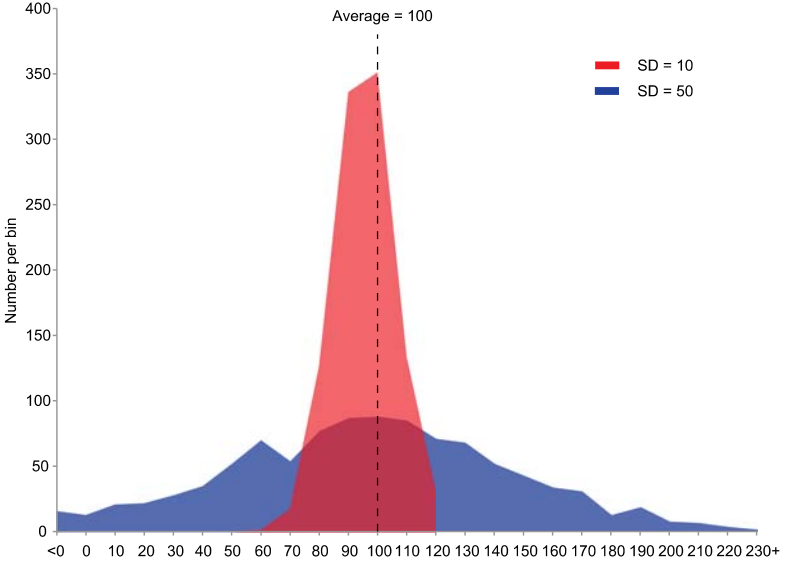
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Standardabweichung** s (*standard deviation*)
  - Da die Varianz quadiert ist, ist sie nur schwer inhaltlich interpretierbar

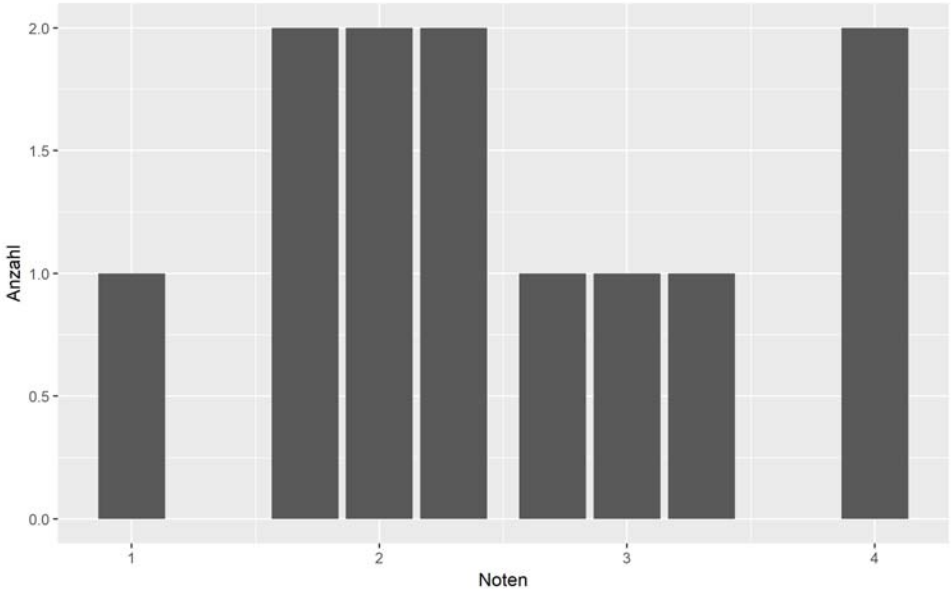
$$\sigma = \sqrt{\sigma^2}$$

Standardabweichung

Beispiel für zwei Verteilungen mit dem gleichen Mittelwert, aber unterschiedlichen Standardabweichungen:



Varianz Beispiel



Varianz Beispiel

ID i	Note x_i	1. Schritt $x_i - \bar{x}$	2. Schritt $(x_i - \bar{x})^2$
1	3.3	0.8	0.64
2	1.7	-0.8	0.64
3	2.0	-0.5	0.25
4	4.0	1.5	2.25
5	1.0	-1.5	2.25
6	2.0	-0.5	0.25
7	3.0	0.5	0.25
8	2.7	0.2	0.04
9	4.0	1.5	2.25

Varianz Beispiel

```
# Mittelwert
mean(grades)
```

```
## [1] 2.5
```

```
# Varianz
var(grades)
```

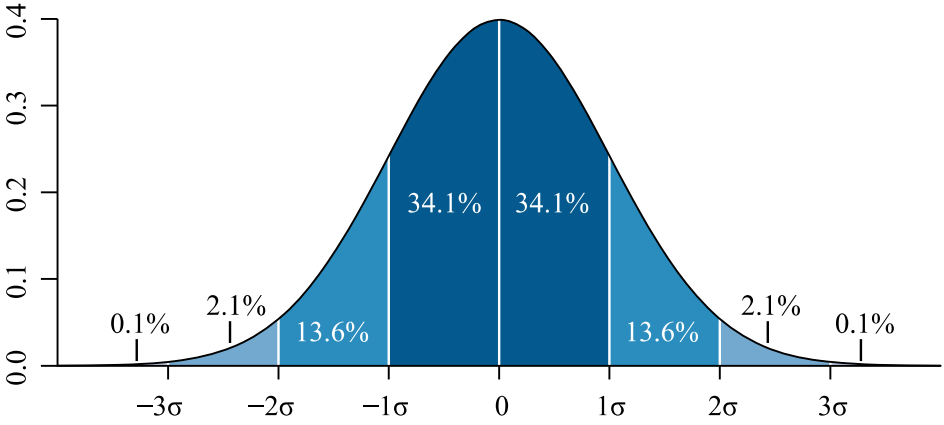
```
## [1] 0.8672727
```

```
# Standardabweichung
sd(grades)
```

```
## [1] 0.9312748
```

Standardabweichung

Standardabweichung in der Gaussschen Normalverteilung



Variationsbreite

- **Variationsbreite** (*range*)
  - Differenz aus dem größten und kleinsten Messwert

$$x_n - x_1$$

```
range(grades)
```

```
## [1] 1 4
```

```
range(income)
```

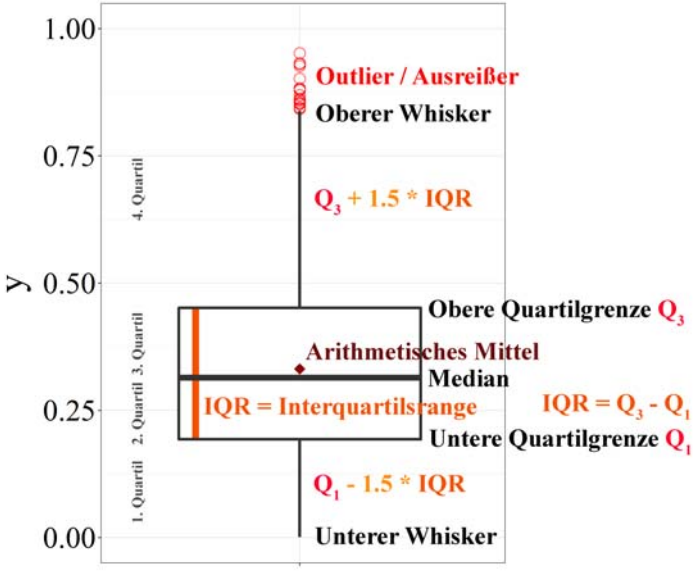
```
## [1]   800 80000
```

Interquartilabstand

- **Interquartilabstand** (*Hinge / IQR*)
  - Auch Tukey-Angelpunkte
  - Drückt die Länge jenes Bereichs aus, über den die mittleren 50% der Verteilung streuen.
  - Berechnet sich analog zum Median.
  - Q_1 ist der *untere Angelpunkt* unterhalb dem 25% der Verteilung liegen.

$$IQR = Q_3 - Q_1$$

Boxplot

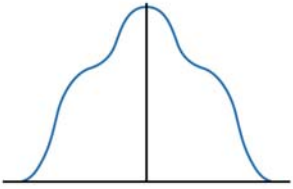


Weitere Erläuterung

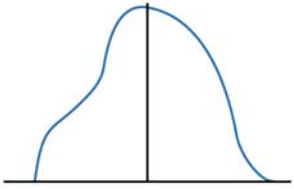
Form von Verteilungen

- **Verteilungen** können:
  - uni- oder bimodal
  - symmetrisch oder schief (*skewness*)
  - spezielle Funktionen sein (z.B. Normalverteilung)

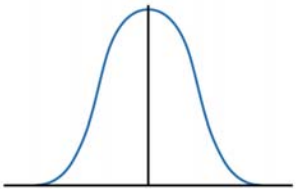
Form von Verteilung



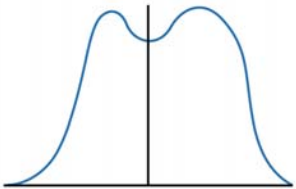
a symmetrisch



b asymmetrisch

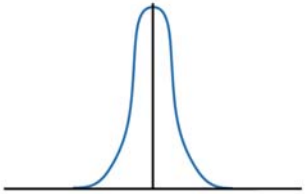


c unimodal

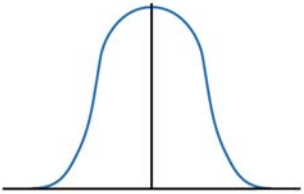


d bimodal

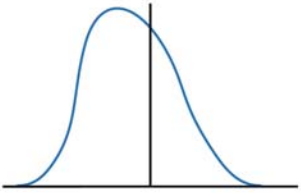
Form von Verteilung



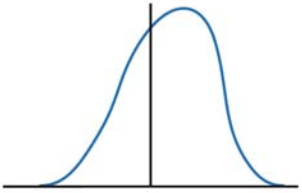
e schmalgipflig



f breitgipflig

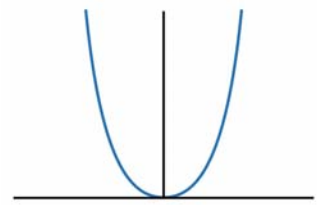


g linkssteil

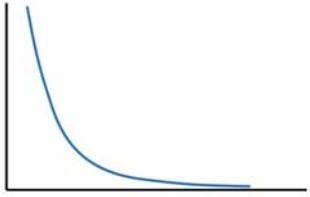


h rechtssteil

Form von Verteilung



i u-förmig



j abfallend

Vielen Dank für die Aufmerksamkeit