LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Datenanalyse – Sitzung 3

Streuung

Institut für Kommunikationswissenschaft und Medienforschung
Ludwig-Maximilians-Universität München

Ablauf der Sitzung

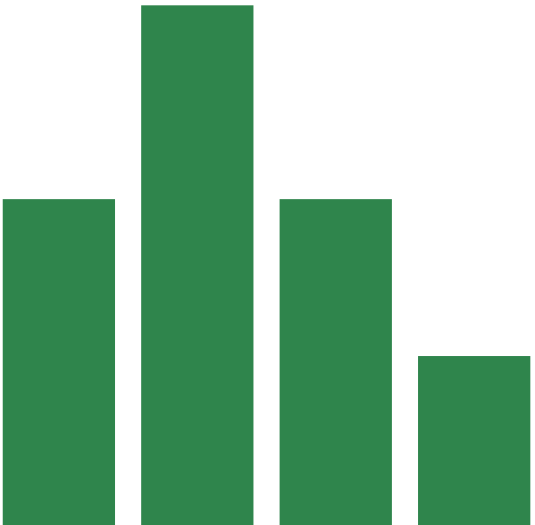
1. Kurze Wiederholung: Streuungsmaße
2. Übung 1: Streuungsmaße manuell berechnen
3. Streuungsmaße in R berechnen
4. Übung 2: Übung 1 in R umsetzen
5. Übung 3: Streuungsmaße in R berechnen

1. Kurze Wiederholung:
Streuungsmaße

Was sind Streuungsmaße?

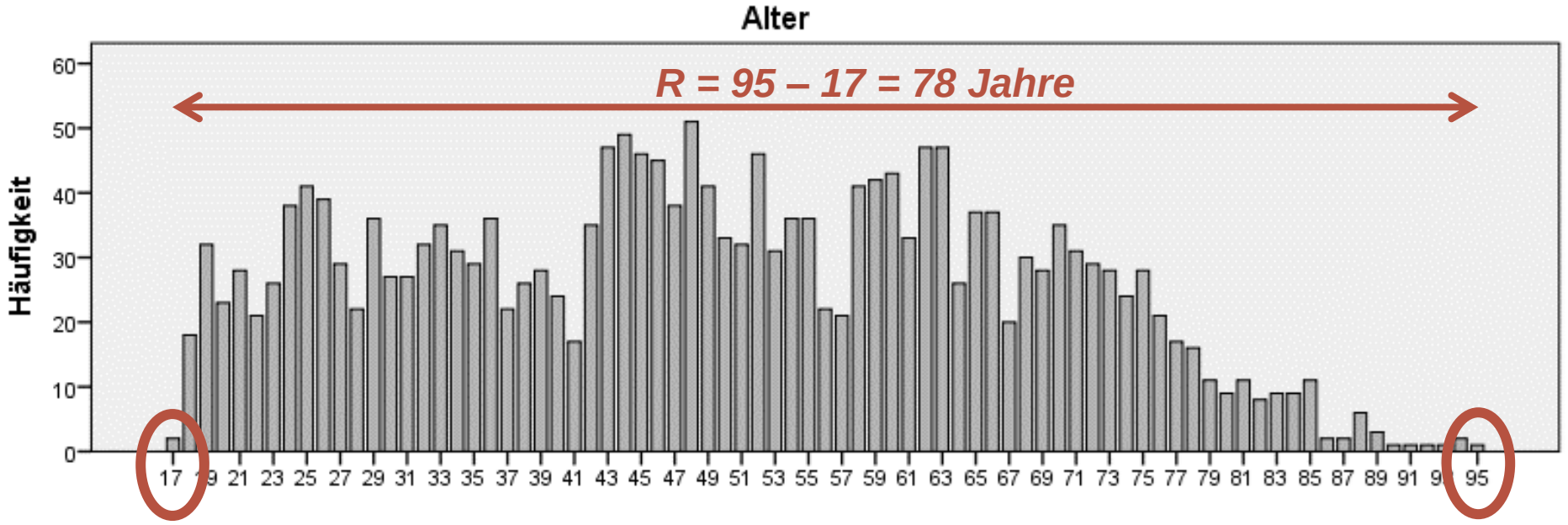
Streuungsmaße sind statistische Kennzahlen, die beschreiben, **wie die Datenpunkte innerhalb eines Datensatzes ausgebreitet oder „gestreut“ sind.** Sie geben an, wie sehr die Daten um ein Lagemaß (wie z.B. dem arithmetischen Mittel) herum variieren.

Dazu zählen: **Spannweite**, **Interquartils-Abstand**, **Varianz**, **Standardabweichung** und der **Variationskoeffizient**.



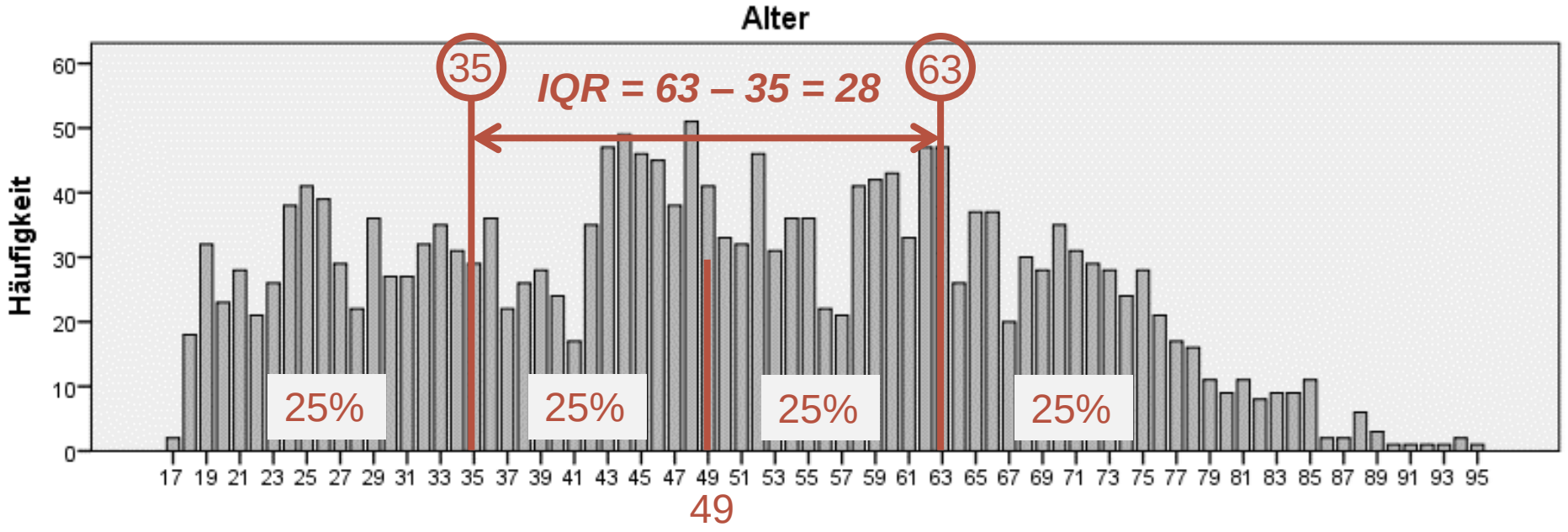
Spannweite (Range, R)

- $R = \max\{x_i\} - \min\{x_i\}$
- Nachteil: Anfällig gegenüber Ausreißern



Interquartils-Abstand (IQR)

- $IQR = Q_{0,75} - Q_{0,25}$ (die „mittleren“ 50% der Stichprobe)
- Vorteil: Robust gegenüber Ausreißern



Varianz (s^2)

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

- Bewertet jeden einzelnen Wert dahingehend, wie weit er vom Mittelwert entfernt ist
- Das Quadrieren sorgt für positive Werte und für eine „Übergewichtung“ von größeren Abweichungen
- Für metrische Daten geeignet (d.h. mindestens intervallskaliert) (nicht für nominal- oder ordinalskalierte Variablen!)

Varianz (s^2)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**3**	**3**	**3**	**4**	**4**	**4**	**4**	**5**

Anzahl	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1x	0	-2,625	6,891
2x	1	-1,625	2,641
5x	2	-0,625	0,391
3x	3	+0,375	0,141
4x	4	+1,375	1,891
1x	5	+2,375	5,641
$N = 16$	$\Sigma = 42$		$\Sigma = 27{,}75$
	$\bar{x} = 2{,}625$		$s_x^2 = 1{,}85$

Wichtig:
Anzahl
für Σ nicht
vergessen

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

$$s^2 = \frac{1}{16-1} \cdot 27{,}75 \approx 1{,}85$$

Wichtig:
Die -1 für die empirische
Varianz nicht vergessen!

Standardabweichung (s)

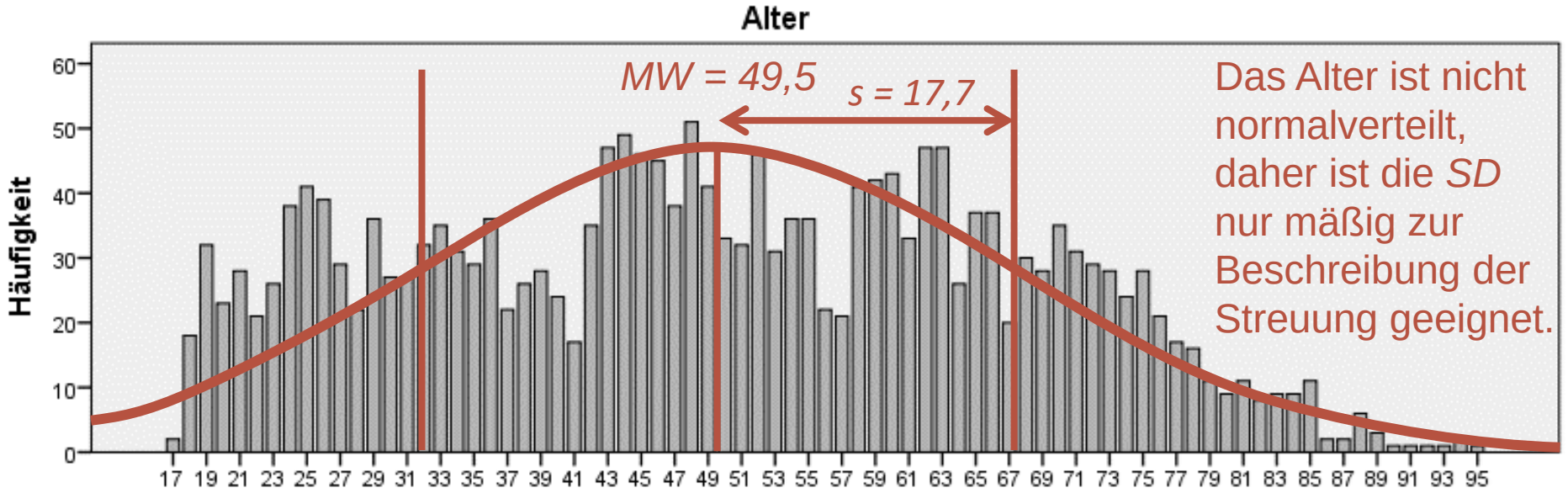
$$s = \sqrt{s^2}$$

(in der Literatur auch *SD* bezeichnet)

- Definiert als Quadratwurzel der Varianz
- Führt den Wert von der quadrierten Größenordnung zurück in die Größenordnung der Messwerte
- Ist ein Maß dafür, wie weit die einzelnen Messwerte um den Mittelwert streuen
- **Interpretation:** Bei normalverteilten Werten liegen 2/3 der Fälle (68,3 %) innerhalb einer Standardabweichung s um den Mittelwert
- **Interpretation:** Bei normalverteilten Daten gilt $IQR = 1{,}35 \cdot s$
- Für metrische Daten geeignet (d.h. mindestens intervallskaliert) (nicht für nominal- oder ordinalskalierte Variablen!)

Standardabweichung (s)

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



Variationskoeffizient (v)

$$v = \frac{s}{\bar{x}}$$

- Streuung relativ zum Mittelwert
- Vergleich von Merkmalen unterschiedlicher Größenordnung hinsichtlich ihrer Streuung
- Nur für verhältnisskalierte bzw. ratioskalierte Daten (!)
  (ohne echten Nullpunkt ist der Mittelwert beliebig verschiebbar, z.B. °C vs. Kelvin)

  (nicht für nominal-, ordinal- oder intervallskalierte Variablen!)
- Nur wenn der Mittelwert nicht 0 ist

Wichtige Take-Aways

- Streuungsmaße: statistische Kennzahlen, die beschreiben, wie die Datenpunkte innerhalb eines Datensatzes ausgebreitet / „gestreut“ sind
- Spannweite (Range, R): Größe des Bereichs, in dem die Daten liegen
- Interquartils-Abstand bzw. IQR: Größe des Bereichs, in dem die mittleren 50% der Fälle liegen
- Varianz (s^2): Mittlere quadratische Abweichung vom arithmetischen Mittel, erst ab Intervallskalen anwendbar
- Standardabweichung (s): Wurzel aus der Varianz (Größenordnung der Werte), erst ab Intervallskalen anwendbar
- Variationskoeffizient (v): Normierte Standardabweichung (nur für Verhältnis- bzw. Ratio-Skalen)

2. Übung 1

Streuungsmaße manuell berechnen

Übung 1

Eine kleine Befragung hat ergeben, dass Personen nur eine begrenzte Anzahl an Fernsehsendern nutzen. Die gemessenen Werte für die Anzahl der Sender sind wie folgt (geordnet):

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25

Bestimmen Sie (ohne R)...

- Spannweite
- Quartile und Interquartils-Abstand
- Varianz und Standardabweichung
- 80%-Perzentil

Lösung für Übung 1

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25

Spannweite

$\mathrm{R} = x_{max} - x_{min} = \mathbf{25 - 1 = 24}$

Lösung für Übung 1

1, 1, 1, 1, 1, 2, 2, 2, 2, **2**, **3**, 3, 3, 3, 3, 3, 3, 3, 3, **3**,
3, 3, 4, 4, 4, 4, 4, 4, 4, **5**, **5**, 5, 5, 6, 6, 7, 9, 10, 15, 25

Quartile und IQR

$$\tilde{x}_p = \begin{cases} \frac{1}{2}\left(x_{N\cdot p} + x_{Np+1}\right) & falls\ N \cdot p\ ganzzahlig \\ x_{[N\cdot p]} & falls\ N \cdot p\ nicht\ ganzzahlig \end{cases}$$

N = **40**;

für $Q_{0,25}$ gilt $N \cdot p = 40 \cdot 0{,}25 = 10 \rightarrow ganzzahlig$

$$Q_{0,25} = \frac{1}{2}\left(x_{N\cdot p} + x_{N\cdot p+1}\right) = \mathbf{\frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(2 + 3) = 2,5}$$

$Q_{0,5} = \mathbf{3}$ und $Q_{0,75} = \mathbf{5}$

$IQR = Q_{0,75} - Q_{0,25} = \mathbf{5 - 2,5 = 2,5}$

Lösung für Übung 1

Anzahl	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5x	1	-3,425	11,73
5x	2	-2,425	5,88
12x	3	-1,425	2,03
7x	4	-0,425	0,18
4x	5	0,575	0,33
2x	6	1,575	2,48
1x	7	2,575	6,63
1x	9	4,575	20,93
1x	10	5,575	31,08
1x	15	10,575	111,83
1x	25	20,575	423,33
$N = 40$	$\Sigma = 177$ $\bar{x} = 4{,}425$		$\Sigma = 713{,}8$

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25

Varianz und Standardabweichung

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{\mathbf{713,8}}{\mathbf{40-1}} \approx \mathbf{18,3}$$

$$s = \sqrt{s^2} = \sqrt{\mathbf{18,3}} \approx \mathbf{4,28}$$

Lösung für Übung 1

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, **5**, **5**, 6, 6, 7, 9, 10, 15, 25

8 Fälle = 20% von N = 40 Fällen

80% Perzentil

$$\tilde{x}_p = \begin{cases} \frac{1}{2}\left(x_{N\cdot p} + x_{Np+1}\right) & falls\ N \cdot p\ ganzzahlig \\ x_{[N\cdot p]} & falls\ N \cdot p\ nicht\ ganzzahlig \end{cases}$$

$$Q_{0,8} = \frac{1}{2}\left(x_{N\cdot p} + x_{N\cdot p+1}\right) = \mathbf{\frac{1}{2}(x_{32} + x_{33}) = \frac{1}{2}(5 + 5) = 5}$$

3. Streuungsmaße in R berechnen

Streuungsmaße berechnen mittels `tidyverse (dplyr)`

Sie können Streuungsmaße ganz genau wie Lagemaße mittels **`dplyr`** berechnen. Wie auch bei den Lagemaßen benötigen Sie dazu die **`dplyr`**-Funktion:

1. **`summarise()`** / **`summarize()`**
2. [ggf. in Kombination mit **`group_by()`**].

Struktur:

```
data %>%
   summarize(
      Range = max(variable) - min(variable),
      Q25 = quantile(variable, 0.25),
      Q75 = quantile(variable, 0.75),
      Perc80 = quantile(variable, 0.80),
      IQR = IQR(variable),
      Variance = var(variable),
      SD = sd(variable),
      V = sd(variable) / mean(variable)
    )
```

Tipp:
`, na.rm = TRUE` lässt `summarize()` NAs ignorieren.

Streuungsmaße berechnen mittels `tidycomm`

Sie können die Streuungsmaße alternativ mit dem `tidycomm` Package berechnen. Dafür benötigen Sie je nach Skalenniveau der Sie interessierenden Variablen folgende Funktionen:

1. `tab_percentiles()*` [Perzentiltabelle]
2. `describe()` [Streuungsmaße für metrische Variablen]

* Eselsbrücke: tab_percentiles steht für „tabulate percentiles“

Struktur:

```
data %>%
   tab_percentiles(variable)


data %>%
   describe(variable)
```

4. Übung 2

Übung 1 in R berechnen

Übung 1 in R umsetzen

Erinnerung: Eine kleine Befragung hat ergeben, dass Personen nur eine begrenzte Anzahl an Fernsehsendern nutzen. Die gemessenen Werte für die Anzahl der Sender sind wie folgt (geordnet):

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25

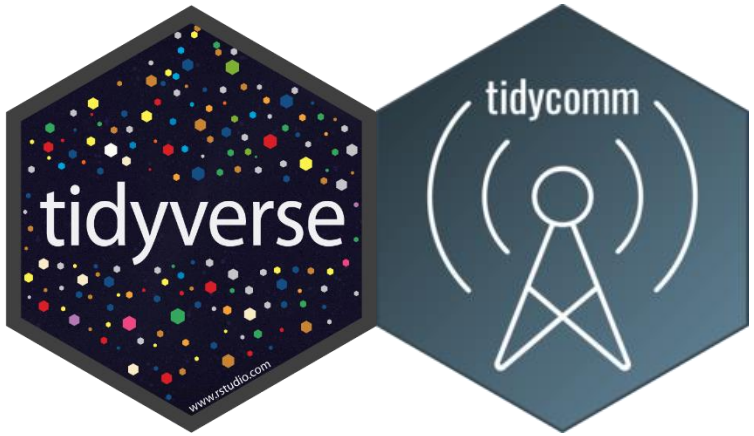
Bestimmen Sie mit R (`tidycomm`)...

- Spannweite
- Quartile und Interquartils-Abstand
- Varianz und Standardabweichung
- 80%-Perzentil

Übung 1 in R umsetzen

Zunächst aktivieren wir die Packages, die die Zusatzfunktionen enthalten, die wir gleich zur Berechnung der Streuungsmaße nutzen wollen. Dazu benutzen wir die Funktion `library()` und aktivieren unsere zwei Packages:

```
library(tidyverse)
library(tidycomm)
```



Übung 1 in R umsetzen

Als nächstes legen wir die 40 Werte samt der ID der Befragten als Vektoren und Dataframe an:

```r
# Anlegen des Datensatzes
tvsender <- c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3,
              3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 9, 10, 15, 25)
befragten_id <- c(1:40)
data <- cbind(befragten_id, tvsender)
data <- tidyr::as_tibble(data)
```

Lösung für Übung 1a in R: `tidycomm`

Übung 1a): Ausgeben der Spannweite.

Sie können die Aufgabe mittels `describe()` aus `tidycomm` lösen, da die Variable metrisch skaliert ist.

Der Code:

```
13  data %>% describe(tvsender)
```

Das Ergebnis:

```
# A tibble: 1 × 15
  Variable     N Missing     M    SD   Min   Q25   Mdn   Q75   Max Range CI_95_LL CI_95_UL Skewness Kurtosis
* <chr>    <int>   <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 tvsender    40       0  4.42  4.28     1  2.75     3     5    25    24     3.06     5.79     3.28     15.2
```

Lösung für Übung 1b in R: `tidycomm`

Übung 1b): Ausgeben der Quartile und des Interquartils-Abstand.

Sie können die Aufgabe mittels **`describe()`** aus **`tidycomm`** lösen, da die Variable metrisch skaliert ist.

Der Code:

```
16  data %>%
17    describe(tvsender) %>%
18    mutate(IQR = Q75 - Q25)
```

Das Ergebnis:

```
# A tibble: 1 × 16
  Variable     N Missing     M    SD   Min   Q25   Mdn   Q75   Max Range CI_95_LL CI_95_UL Skewness Kurtosis   IQR
  <chr>    <int>   <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <dbl>
1 tvsender    40       0  4.42  4.28     1  2.75     3     5    25    24     3.06     5.79     3.28     15.2  2.25
```

Achtung:
Den IQR berechnet Ihnen tidycomm nicht automatisch. Sie können den IQR aber mittels **`dplyr::mutate()`** dank Ihrer Statistik-Kenntnisse berechnen als Q75 - Q25 = 5 – 2.75 = 2.25.

Lösung für Übung 1c in R: `tidycomm`

Übung 1c): Ausgeben der Varianz und Standardabweichung.

Sie können die Aufgabe mittels **`describe()`** aus **`tidycomm`** lösen, da die Variable metrisch skaliert ist.

Der Code:

```
22  data %>%
23    describe(tvsender) %>%
24    mutate(Variance = SD^2)
```

Das Ergebnis:

```
# A tibble: 1 × 16
  Variable     N Missing     M    SD   Min   Q25   Mdn   Q75   Max Range CI_95_LL CI_95_UL Skewness Kurtosis Variance
  <chr>    <int>   <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 tvsender    40       0  4.42  4.28     1  2.75     3     5    25    24     3.06     5.79     3.28     15.2     18.3
```

Achtung:

Den Varianz berechnet Ihnen tidycomm nicht automatisch. Sie können die Varianz aber mittels **`dplyr::mutate()`** dank Ihrer Statistik-Kenntnisse berechnen als $SD^2 = 4.28^2 = 18.3$

Lösung für Übung 1d in R: `tidycomm`

Übung 1d): Ausgeben des 80%-Perzentil.

Sie können die Aufgabe mittels `tab_percentiles()` aus `tidycomm` lösen.

Der Code:

```
26  data %>% tab_percentiles(tvsender)
```

Das Ergebnis:

```
> data %>% tab_percentiles(tvsender)
# A tibble: 1 × 11
  Variable   p10   p20   p30   p40   p50   p60   p70   p80   p90  p100
* <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 tvsender     1     2     3     3     3     4     4     5   7.2    25
```

5. Übung 3

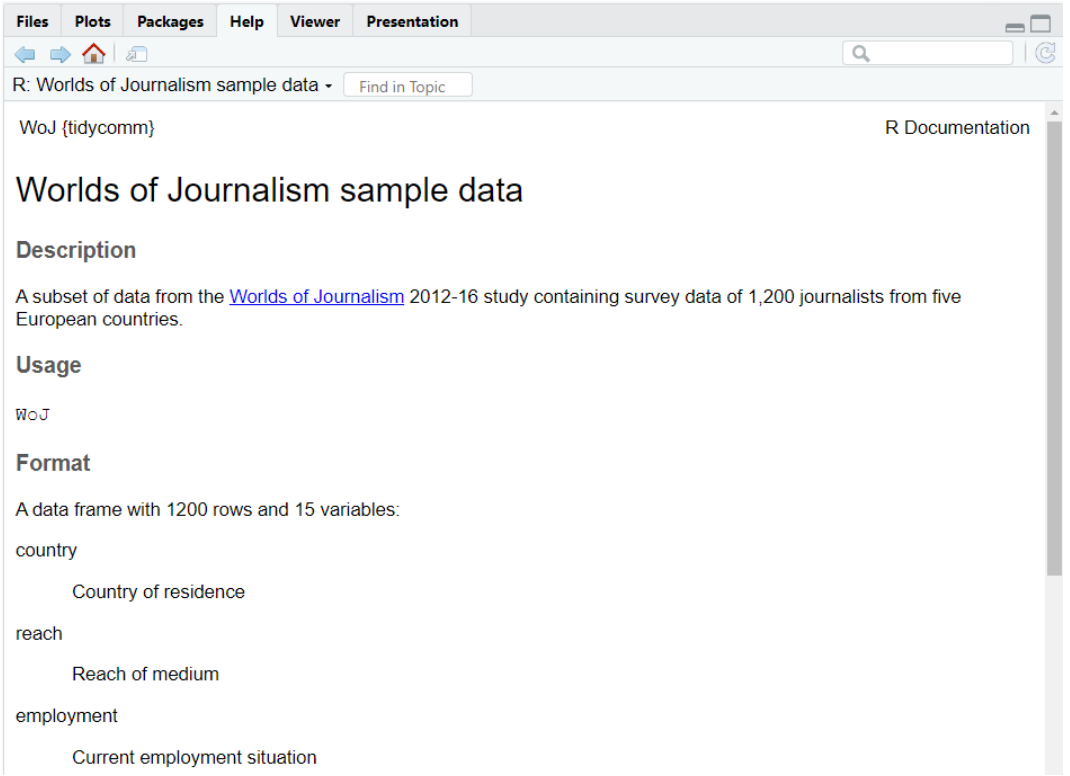
Streuungsmaße in R berechnen und visualisieren

Übung 3

Laden Sie nun erneut den Datensatz „WoJ“ vom tidycomm-Package in Ihre RStudio-Environment.

```
29  WoJ <- WoJ
```

Erinnerung: Hilfeseite zum WoJ-Datensatz (`?WoJ`)



Übung 3

Die Journalist*innen im WoJ-Datensatz wurden gefragt, wie lange Sie bereits in der Branche arbeiten («work_experience») und wie sehr Sie Politiker*innen trauen («trust_politicians») .

Beantworten Sie mittels `tidycomm`:

a) In welchem Wertebereich liegen die mittleren 80% der Antworten?
b) Stellen Sie die Verteilung der Antworten im Histogramm / Balkendiagramm dar.
c) Stellen Sie die Verteilung der Antworten im Boxplot dar.
d) In welcher der beiden Eigenschaften unterscheiden sich die Journalist*innen untereinander stärker?

Lösung für Übung 3a mittels `tidycomm`

a) In welchem Wertebereich liegen die mittleren 80% der Antworten?

Der Code:

```
33  WoJ %>%
34    tab_percentiles(work_experience)
```

Ergebnis:

```
# A tibble: 1 × 11
  Variable          p10   p20   p30   p40   p50   p60   p70   p80   p90  p100
* <chr>           <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 work_experience     4     7    10    14    17    20    25    28    33    53
```

Bereich zwischen $Q_{10\%}$ und $Q_{90\%}$ Beufserfahrung: [4; 33] Jahre

Der Code:

```
36  WoJ %>%
37    tab_percentiles(trust_politicians)
```

Ergebnis:

```
# A tibble: 1 × 11
  Variable          p10   p20   p30   p40   p50   p60   p70   p80   p90  p100
* <chr>           <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 trust_politicians     2     2     2     2     3     3     3     3     3     4
```

Bereich zwischen $Q_{10\%}$ und $Q_{90\%}$ Vertrauen in Politiker*innen: [2; 3]

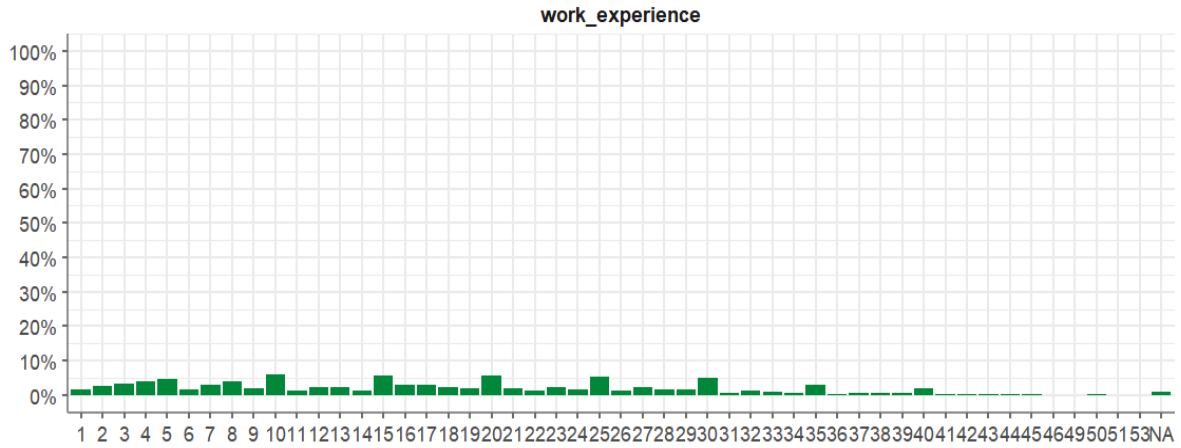
Lösung für Übung 3b mittels `tidycomm`

b) Stellen Sie die Verteilung der Antworten im Histogramm / Balkendiagramm dar.

Der Code:

```
39  WoJ %>%
40    tab_frequencies(work_experience)
41    visualize()
```

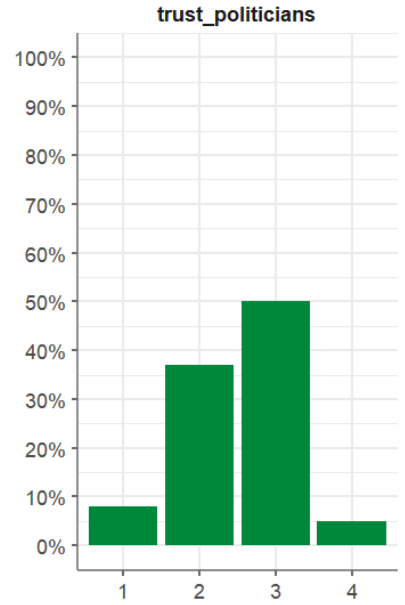
Das Ergebnis:



Der Code:

```
43  WoJ %>%
44    tab_frequencies(trust_politicians) %>%
45    visualize()
```

Das Ergebnis:



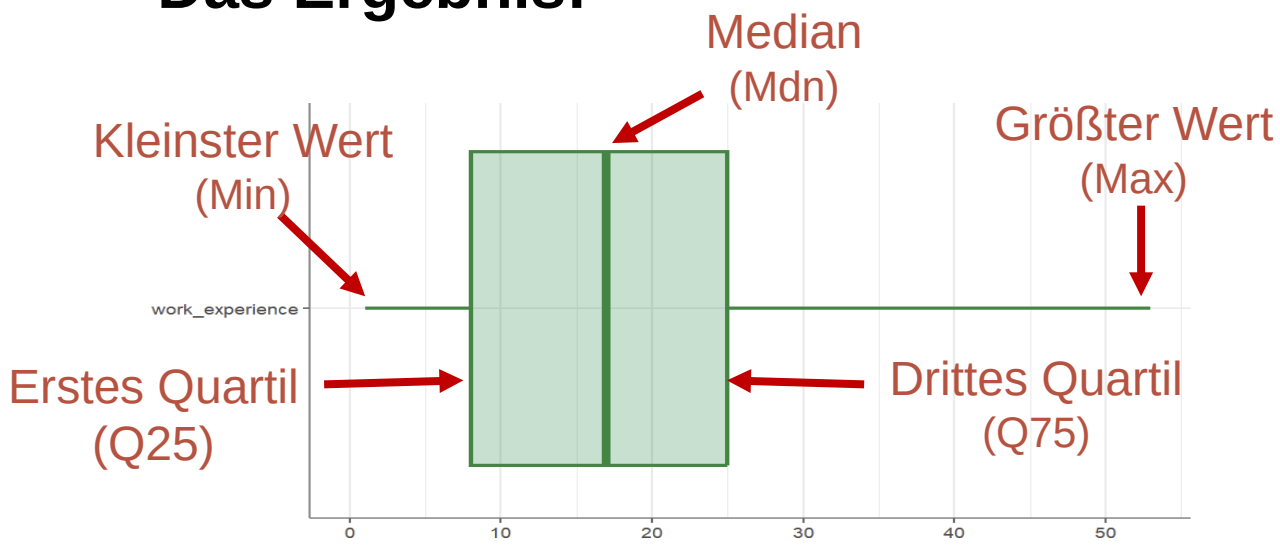
Lösung für Übung 3c mittels `tidycomm`

c) Stellen Sie die Verteilung der Antworten im Boxplot dar.

Der Code:

```
48  WoJ %>%
49    describe(work_experience) %>%
50    visualize()
```

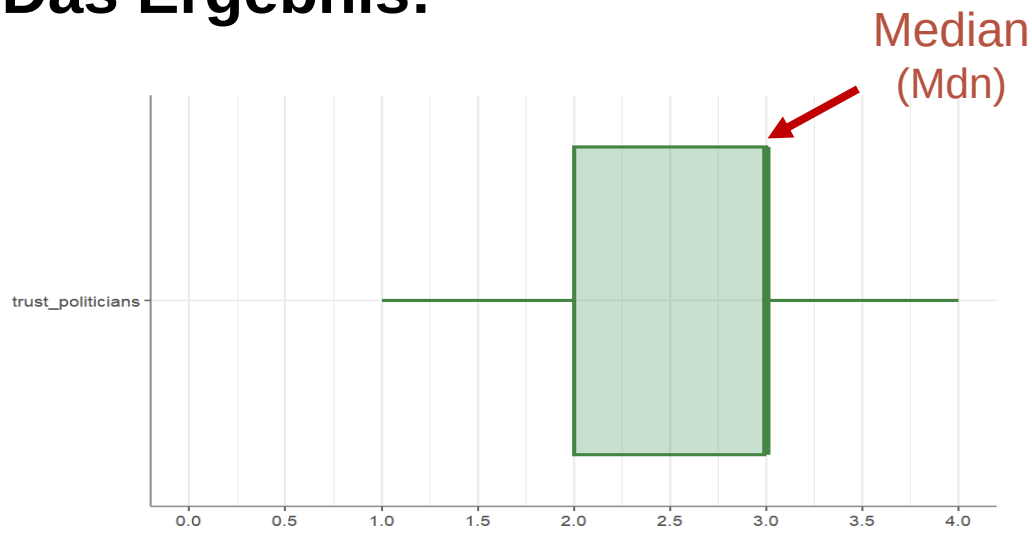
Das Ergebnis:



Der Code:

```
52  WoJ %>%
53    describe(trust_politicians) %>%
54    visualize()
```

Das Ergebnis:



Lösung für Übung 3d mittels `tidyverse (dplyr)`

d) In welcher der beiden Eigenschaften unterscheiden sich die Journalist*innen untereinander stärker?

Variationskoeffizient: $v = \frac{s}{\bar{x}}$

Der Code:

```
57  WoJ %>%
58    describe(work_experience) %>%
59    mutate(V_WE = SD / M) %>%
60    select(V_WE, SD, M)
```

Das Ergebnis:

```
# A tibble: 1 × 3
   V_WE    SD     M
  <dbl> <dbl> <dbl>
1 0.614  10.9  17.8
```

$v_{WE} = \frac{s}{\bar{x}} = \frac{10{,}9}{17{,}8} = 0{,}61$

Der Code:

```
62  WoJ %>%
63    describe(trust_politicians) %>%
64    mutate(V = SD / M) %>%
65    select(V, SD, M)
```

Das Ergebnis:

```
# A tibble: 1 × 3
   V_TP    SD     M
  <dbl> <dbl> <dbl>
1 0.282 0.712  2.52
```

$v_{TP} = \frac{s}{\bar{x}} = \frac{0{,}71}{2{,}52} = 0{,}28$

→ Größere Unterschiede in Berufserfahrung

Wichtige Take-Aways

- Streuungsmaße in R: berechnen mit `dplyr::summarize()` oder `tidycomm::describe()` / `tidycomm::tab_percentiles()`
- `dplyr::summarize()`: flexibler als `tidycomm`, benötigt aber mehr Subfunktionen wie `mean()`, `var()`, uvm.
- `tidycomm::describe()`: berechnet per Default *M*, *SD*, *Q25*, *Q75* & *Range*, in Kombination mit `dplyr::mutate()` auch *IQR*, *Varianz* und den *Variationskoeffizient*
- `tidycomm::tab_percentiles()` berechnet Perzentile