

# 03

## Statistische Messgrößen

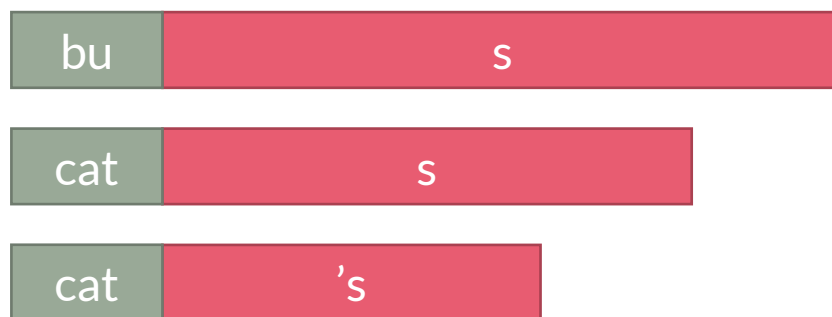
Dominic Schmitz & Janina Esser

# Statistische Messgrößen

- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

## **The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords<sup>1</sup>**

- Wort-finales /s/ zeigt je nach Bedeutung unterschiedliche Dauern



<sup>1</sup>Schmitz, D., Baer-Henney, D., & Plag, I. (2021). The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. *Phonetica*, 78(5-6), 571-616. doi: 10.1515/phon-2021-2013

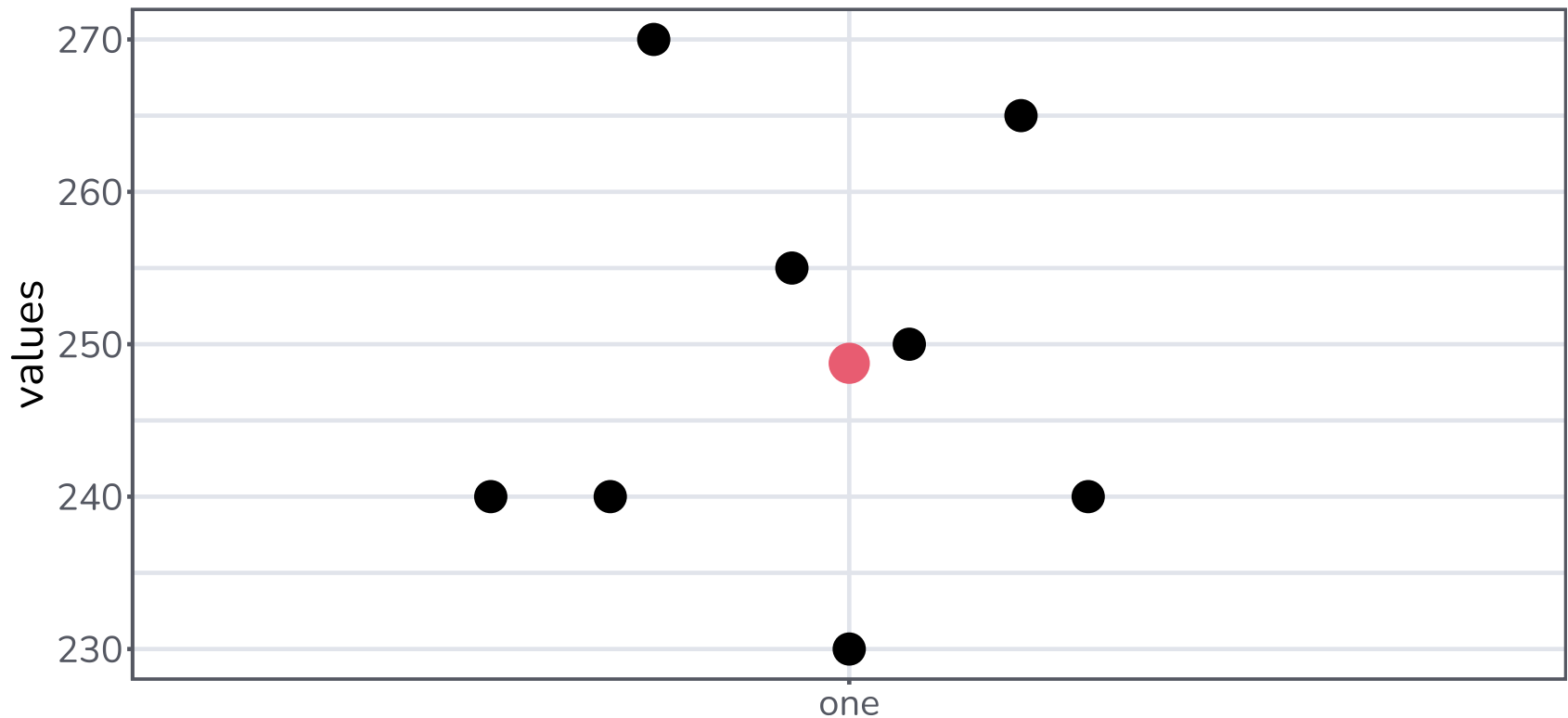
# Statistische Messgrößen

- **Measures of Central Tendency / Lagemaße**
  - Mean / Durchschnitt, Arithmetisches Mittel
  - Median
  - Mode / Modus
- **Measures of Dispersion / Streuungsmaße**
  - Range / Spannweite
  - Interquartile Range / Interquartilsabstand
  - Sample Variance / Stichprobenvarianz
  - Standard Deviation / Standardabweichung
  - Standard Error / Standardfehler
- **Shape of Distribution / Form der Verteilung**
  - Skewness / Schiefe

# Mean / Durchschnitt, Arithmetisches Mittel

Lagemaße

Die Summe aller Werte geteilt durch die Anzahl der Werte



# Mean / Durchschnitt, Arithmetisches Mittel

Die Summe aller Werte geteilt durch die Anzahl der Werte

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + \dots + a_n}{n}$$

Beispiel:

$$\bar{x} = \frac{270 + 240 + 240 + 255 + 250 + 265 + 230 + 240}{8} = 248.75$$

# Mean / Durchschnitt, Arithmetisches Mittel

Die Summe aller Werte geteilt durch die Anzahl der Werte

```
mean(data_s$sDur)
```

```
## [1] 0.1315305
```

```
mean(data_s$baseDur)
```

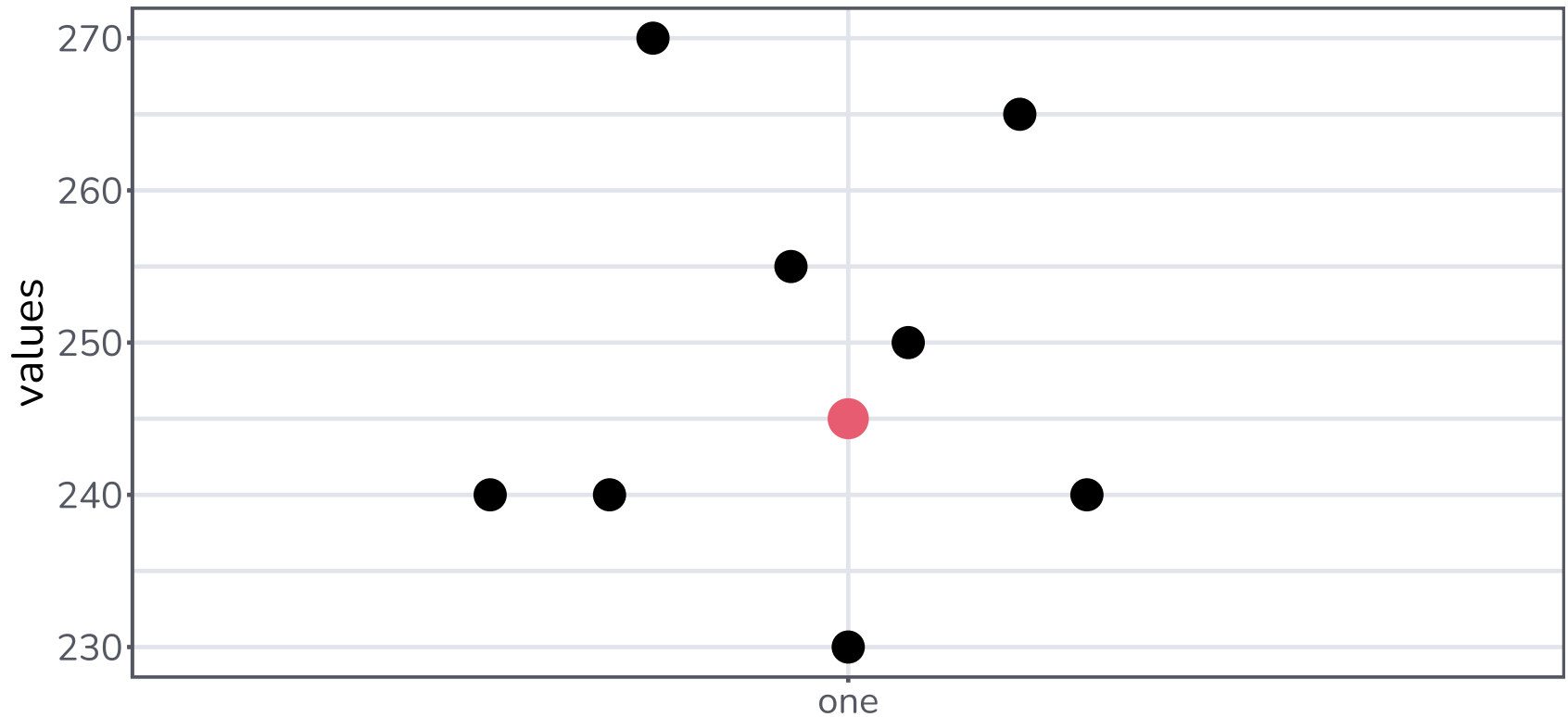
```
## [1] 0.3190967
```

```
mean(data_s$speakingRate)
```

```
## [1] 3.449667
```

# Median

Der mittlere Wert in einer aufsteigend geordneten Reihe



Der mittlere Wert in einer aufsteigend geordneten Reihe

$$\tilde{m}(a) = \frac{a_{\lfloor \#x \div 2 \rfloor} + a_{\lfloor \#x \div 2 + 0.5 \rfloor}}{2}$$

Beispiel:

↓ 245  
230, 240, 240, 240, 250, 255, 265, 270  
↓  
230, 240, 240, 240, 250, 255, 265



Der mittlere Wert in einer aufsteigend geordneten Reihe

```
median(data_s$sDur)
```

```
## [1] 0.118175
```

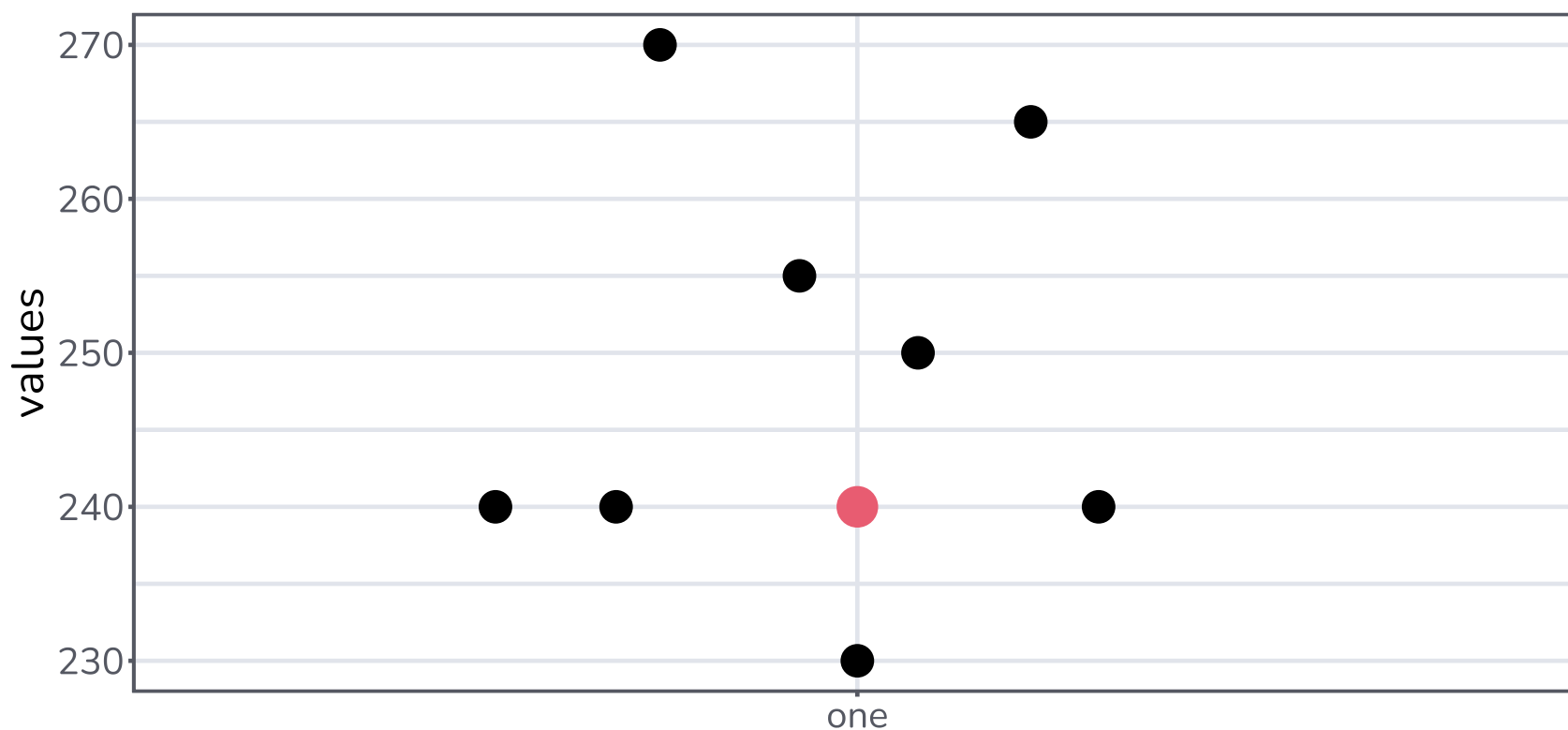
```
median(data_s$baseDur)
```

```
## [1] 0.306315
```

```
median(data_s$speakingRate)
```

```
## [1] 3.355
```

Der Wert, der am häufigsten in einer Gruppe von Werten vorkommt

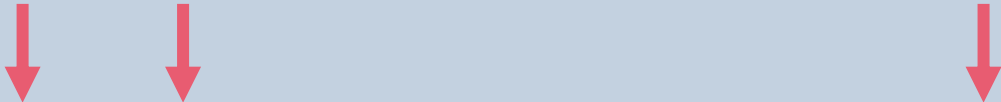


Der Wert, der am häufigsten in einer Gruppe von Werten vorkommt

$$D = L + \frac{(f_m - f_1)h}{2f_m - f_1 - f_2}$$

Beispiel:

270, 240, 240, 255, 250, 265, 230, 240



Der Wert, der am häufigsten in einer Gruppe von Werten vorkommt

```
mode_stat(data_s$sDur)
```

```
## [1] 0.1311
```

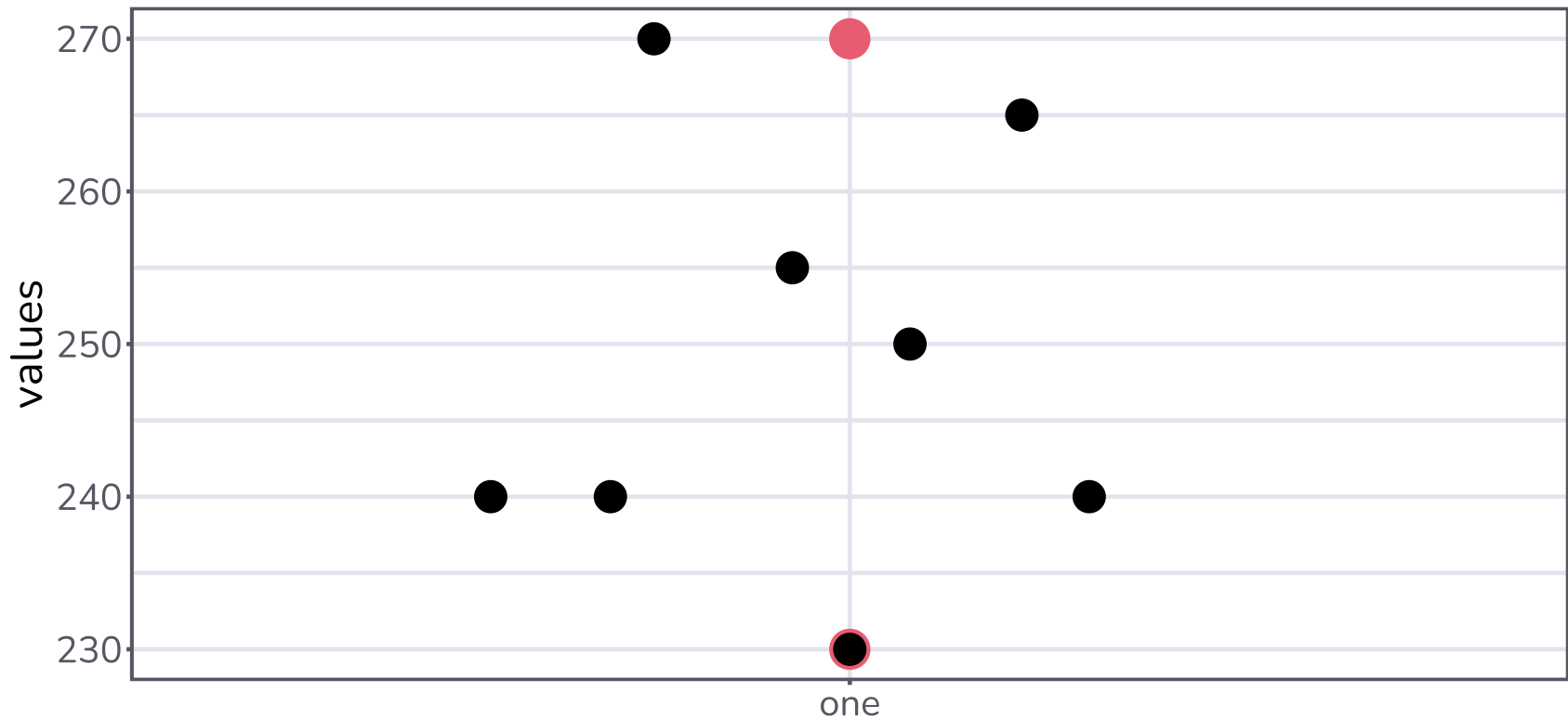
```
mode_stat(data_s$baseDur)
```

```
## [1] 0.25162
```

```
mode_stat(data_s$speakingRate)
```

```
## [1] 2.94
```

Die Differenz zwischen dem kleinsten und dem größten Wert in einer Gruppe von Werten



Die Differenz zwischen dem kleinsten und dem größten Wert in einer Gruppe von Werten

$$R = x_{max} - x_{min}$$

Beispiel:

**230, 240, 240, 240, 250, 255, 265, 270**

$$R = 270 - 230 = 40$$

Die Differenz zwischen dem kleinsten und dem größten Wert in einer Gruppe von Werten

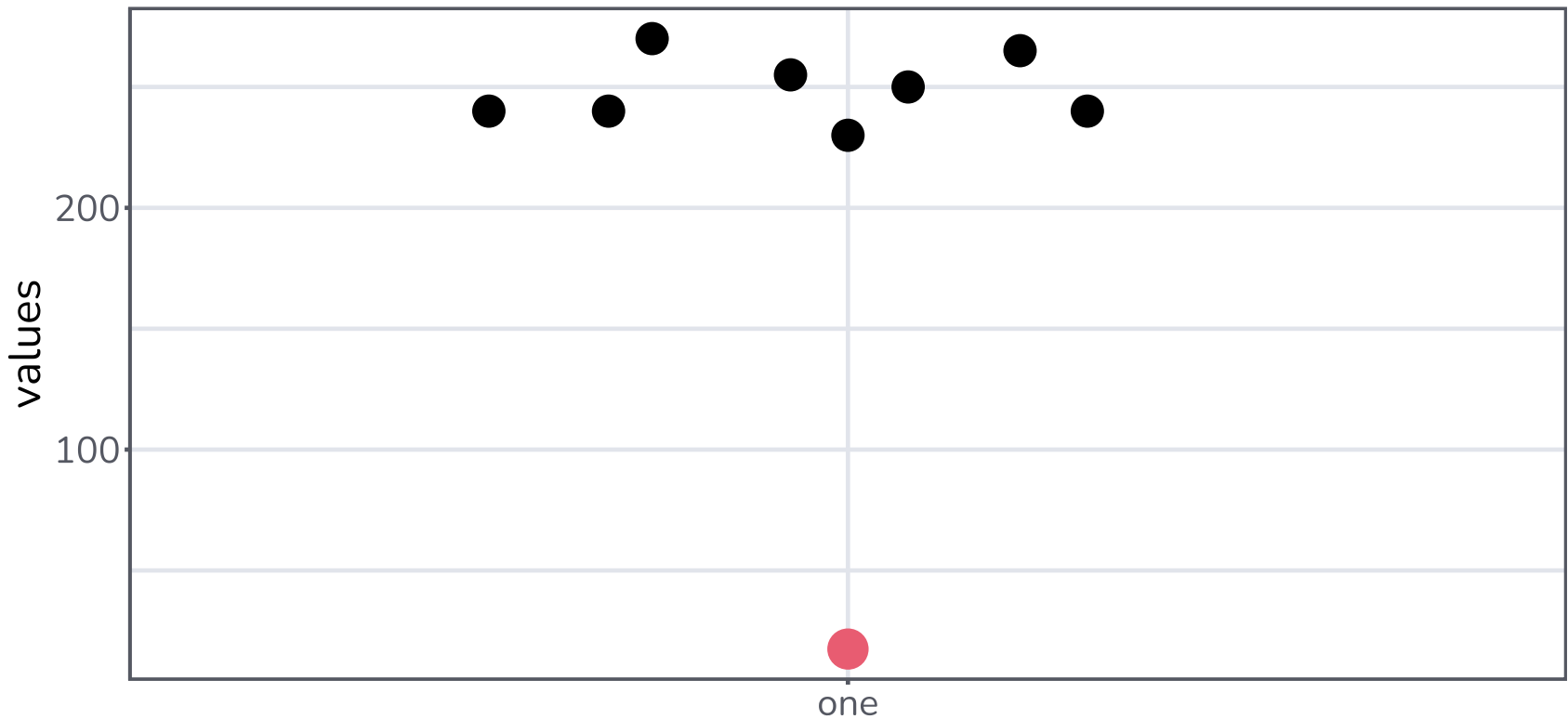
```
range(data_s$sDur)  
## [1] 0.04176 0.32750
```

```
range(data_s$baseDur)  
## [1] 0.17995 0.68749
```

```
range(data_s$speakingRate)  
## [1] 1.52 6.94
```

# Interquartile Range / Interquartilsabstand

Beschreibt die mittleren 50% der Werte wenn sie vom kleinsten zu größten Wert geordnet sind





Beschreibt die mittleren 50% der Werte wenn sie vom kleinsten zu größten Wert geordnet sind

$$IQR = Q_{0.75} - Q_{0.25}$$

## Beispiel:

1. 270, 240, 240, 255, 250, 265, 230, 240    ursprüngliche Daten
2. 230, 240, 240, 240, 250, 255, 265, 270    aufsteigend sortiert
3. Quantile berechnen  
 $Q_{75} = \frac{1}{2} * (255 + 265) = 260$  &  $Q_{25} = \frac{1}{2} * (240 + 240) = 240$
4.  $IQR = Q_{0.75} - Q_{0.25} = 20$

# Interquartile Range / Interquartilsabstand

Beschreibt die mittleren 50% der Werte wenn sie vom kleinsten zu größten Wert geordnet sind

```
IQR(data_s$sDur)
```

```
## [1] 0.06783
```

```
IQR(data_s$baseDur)
```

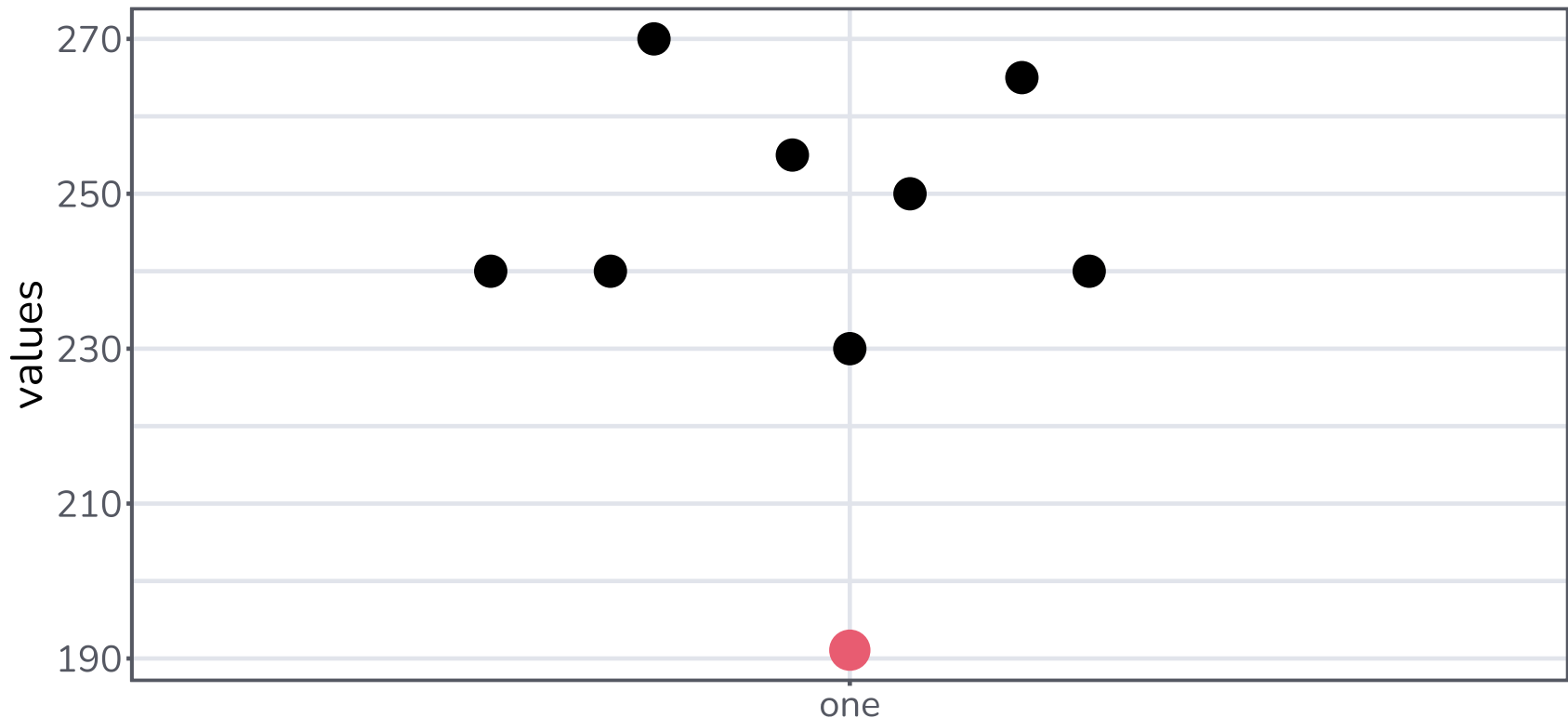
```
## [1] 0.1067575
```

```
IQR(data_s$speakingRate)
```

```
## [1] 1.125
```

# Sample Variance / Stichprobenvarianz

Beschreibt die mittlere quadratische Abweichung der einzelnen Werte vom Mittelwert



Beschreibt die mittlere quadratische Abweichung der einzelnen Werte vom Mittelwert

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

Beispiel:

230, 240, 240, 240, 250, 255, 265, 270

$$s^2 = \frac{1}{8-1} \sum_{i=1}^8 (x_i - \bar{x})^2 = \frac{1337.5}{7} \approx 191.07$$

Beschreibt die mittlere quadratische Abweichung der einzelnen Werte vom Mittelwert

```
var(data_s$sDur)
## [1] 0.002990366
```

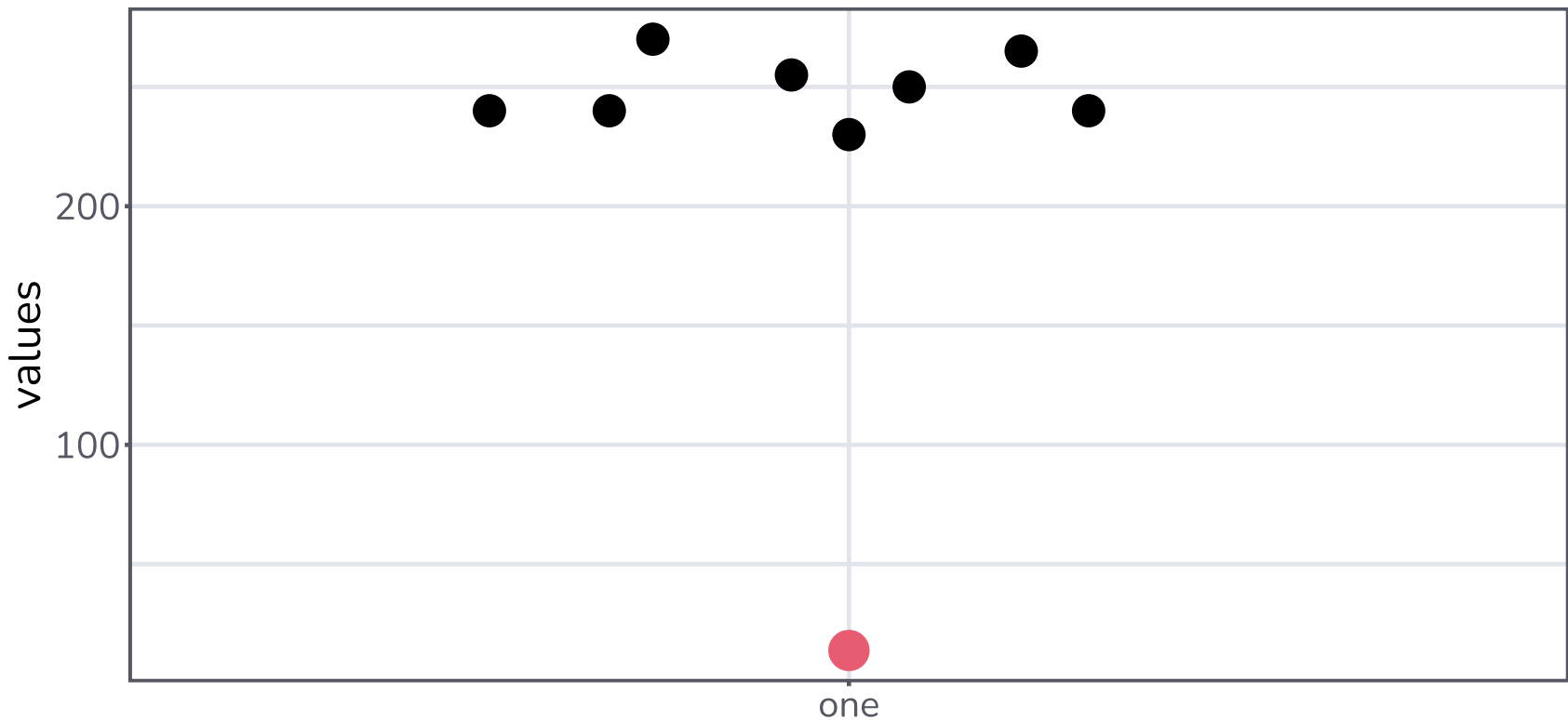
```
var(data_s$baseDur)
## [1] 0.007913081
```

```
var(data_s$speakingRate)
## [1] 0.8649482
```

# Standard Deviation / Standardabweichung

Streuungsmaße

Ein Maß dafür, wie weit die Werte um den Mittelwert gestreut sind



Ein Maß dafür, wie weit die Werte um den Mittelwert gestreut sind

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

**Wurzel der  
Varianz!**

Beispiel:

230, 240, 240, 240, 250, 255, 265, 270

$$s = \sqrt{\frac{1337.5}{7}} \approx 13.82$$

Ein Maß dafür, wie weit die Werte um den Mittelwert gestreut sind

```
sd(data_s$sDur)
```

```
## [1] 0.05468424
```

```
sd(data_s$baseDur)
```

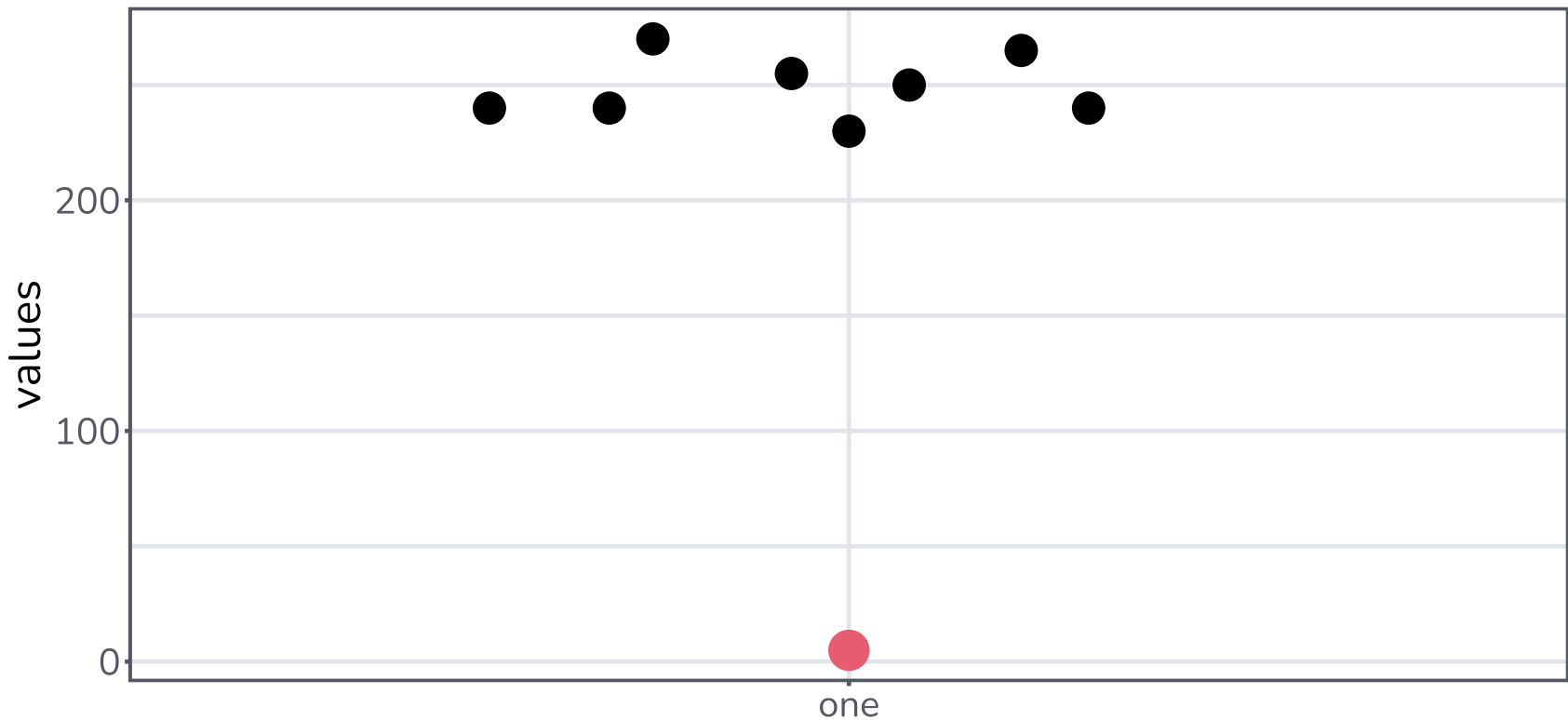
```
## [1] 0.0889555
```

```
sd(data_s$speakingRate)
```

```
## [1] 0.9300259
```



Angabe der Genauigkeit, mit der eine Stichprobe eine Grundgesamtheit repräsentiert



Angabe der Genauigkeit, mit der eine Stichprobe eine Grundgesamtheit repräsentiert

$$\sigma(\bar{X}) = \frac{s}{\sqrt{n}}$$

**Standardabweichung geteilt  
durch die Wurzel der  
Stichprobengröße**

Beispiel:

230, 240, 240, 240, 250, 255, 265, 270

$$\sigma(\bar{X}) = \frac{\frac{1}{8-1} \sum_{i=1}^8 (x_i - \bar{x})^2}{\sqrt{8}} \approx 4.89$$

Angabe der Genauigkeit, mit der eine Stichprobe eine Grundgesamtheit repräsentiert

```
se(data_s$sDur)
```

```
## [1] 0.004464949
```

```
se(data_s$baseDur)
```

```
## [1] 0.007263186
```

```
se(data_s$speakingRate)
```

```
## [1] 0.0759363
```

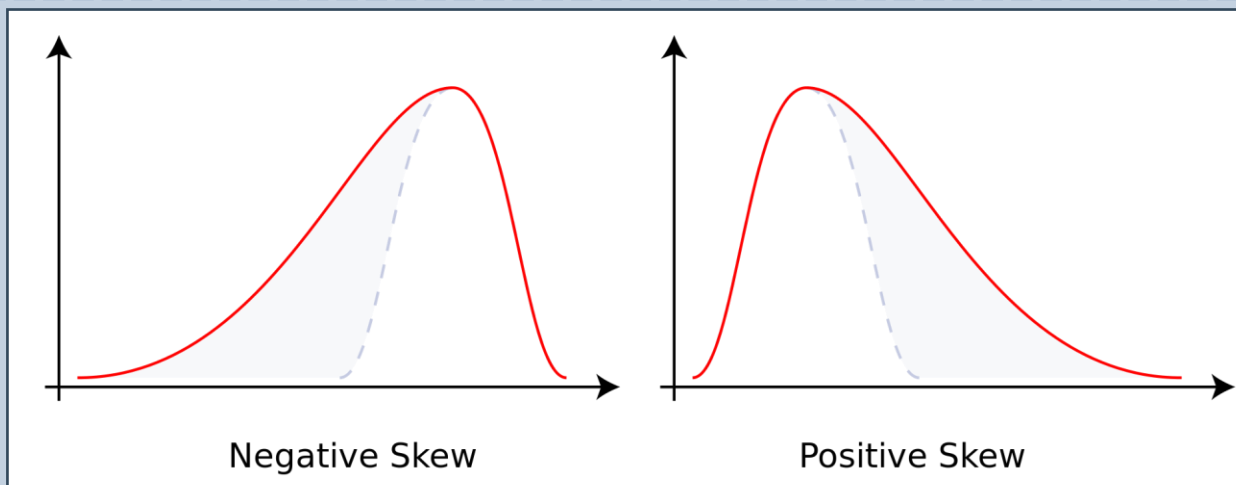
Asymmetrie in einer statistischen Verteilung, bei der die Kurve entweder nach links oder nach rechts verzerrt erscheint

$$v = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

$\bar{x}$  = Durchschnitt

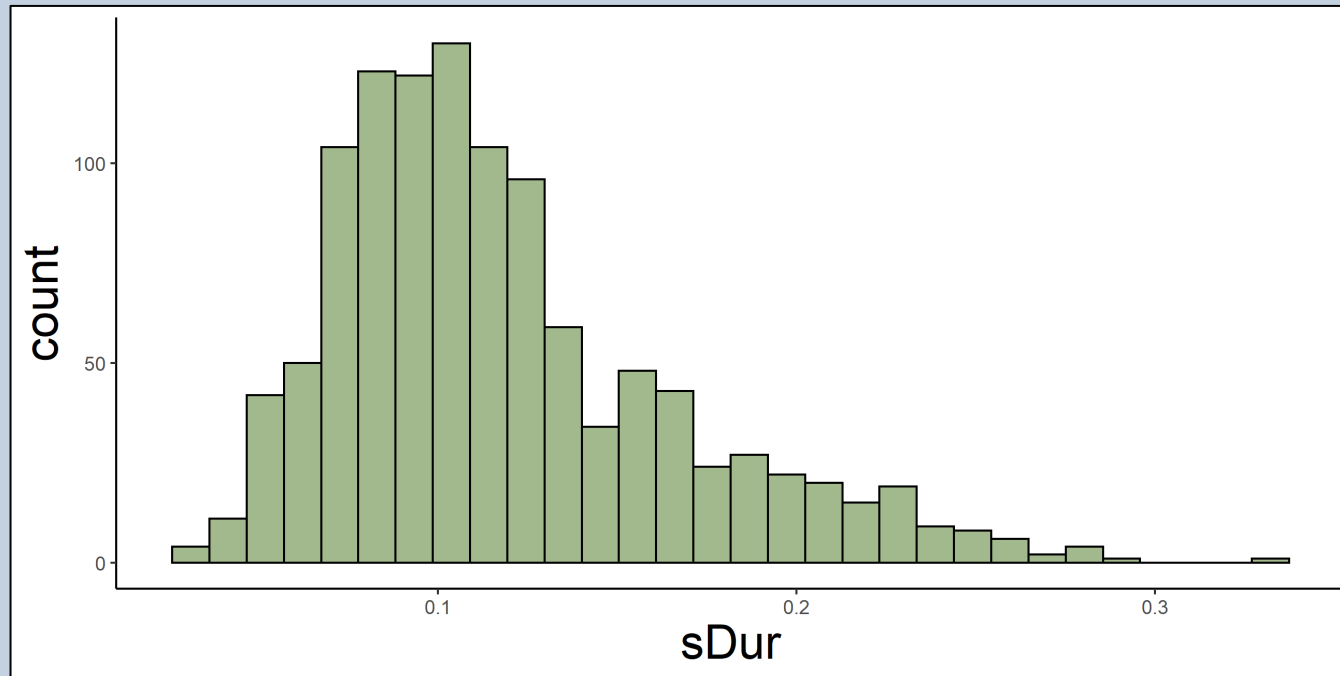
s = Standardabweichung

Beispiel:



Die meisten linguistischen Daten sind positiv schief, d. h. es gibt mehr Daten auf der linken Seite der Verteilung als auf der rechten

## Beispiel:



Asymmetrie in einer statistischen Verteilung, bei der die Kurve entweder nach links oder nach rechts verzerrt oder schief erscheint

```
skewness(data_s$sDur)
```

```
## [1] 0.9483159
```

```
skewness(data_s$baseDur)
```

```
## [1] 1.360664
```

```
skewness(data_s$speakingRate)
```

```
## [1] 0.8348821
```