Session 03:
Statistische Messgrößen

Dominic Schmitz & Janina Esser

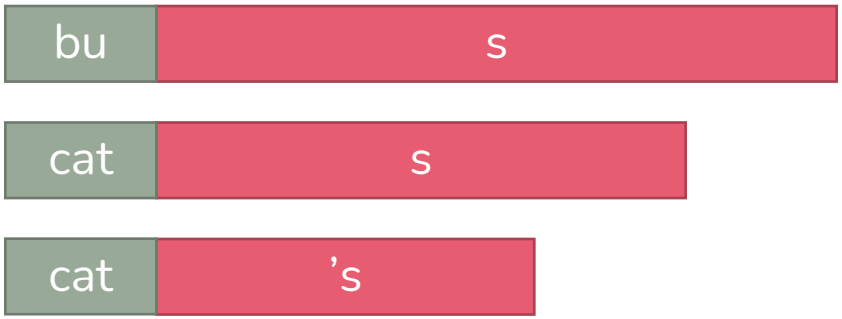
Verein für Diversität in der Linguistik

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[1]

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



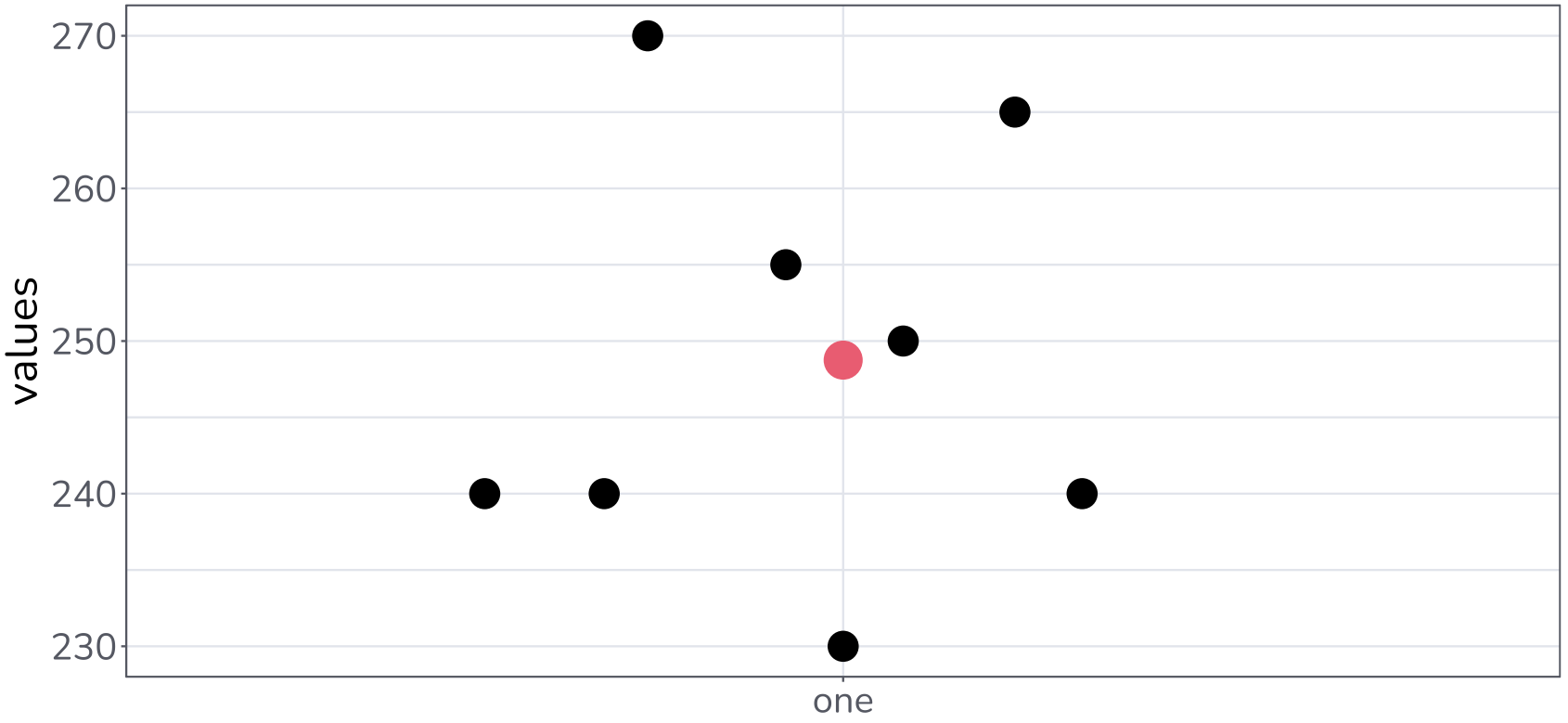
[1] Schmitz, D., Baer-Henney, D., & Plag, I. (2021). The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. Phonetica, 78(5-6), 571-616. doi: 10.1515/phon-2021-2013

Statistische Messgrößen

- Measures of Central Tendency
  - **Mean / Durchschnitt, Arithmetisches Mittel**
  - **Median**
  - **Mode / Modus**
- Measures of Dispersion
  - **Range / Spannweite**
  - **Interquartile Range / Interquartilspannweite**
  - **Sample Covariance / Stichprobenvarianz**
  - **Standard Deviation / Standardabweichung**
  - **Standard Error / Standardfehler**
- Shape of Distribution
  - **Skewness / Schiefe**

Measures of Central Tendency

- **Mean / Durchschnitt, Arithmetisches Mittel**
  The sum of all values divided by the number of values



Measures of Central Tendency

- **Mean / Durchschnitt, Arithmetisches Mittel**
  The sum of all values divided by the number of values

$$A = \frac{1}{n}\sum_{i=1}^{n} a_1 = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Example:

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

Measures of Central Tendency

- **Mean / Durchschnitt, Arithmetisches Mittel**
  The sum of all values divided by the number of values

```
mean(data$sDur)
## [1] 0.1315305

mean(data$baseDur)
## [1] 0.3190967

mean(data$speakingRate)
## [1] 3.449667
```

Measures of Central Tendency

- **MEAN / DURCHSCHNITT, ARITHMETISCHES MITTEL**
  The sum of all values divided by the number of values

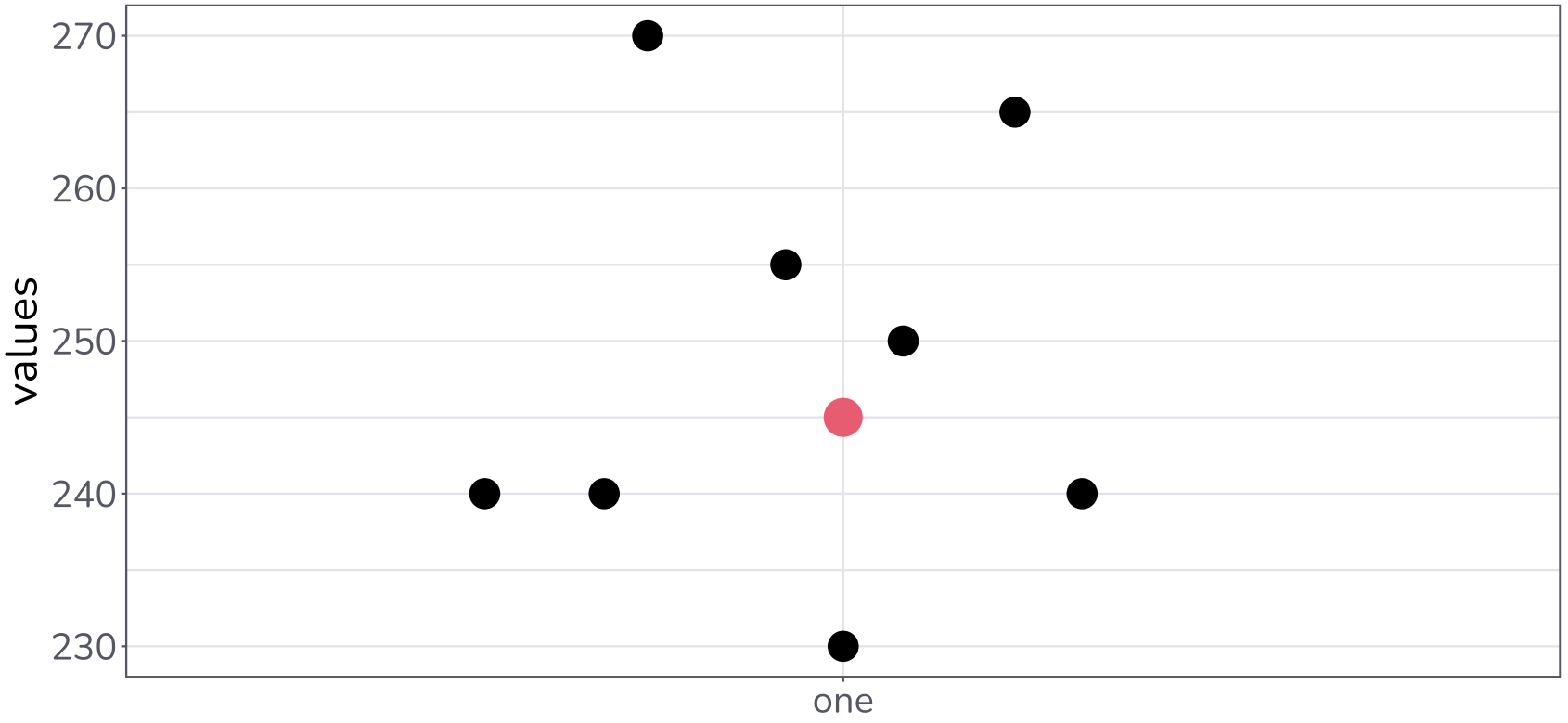
```
mean(data$sDur[data$typeOfS == "nm"])
## [1] 0.156608

mean(data$sDur[data$typeOfS == "pl"])
## [1] 0.1317052

mean(data$sDur[data$typeOfS == "is"])
## [1] 0.1062782
```

Measures of Central Tendency

- **MEDIAN**
  The middle value in a series of values ordered from the smallest to the largest



Measures of Central Tendency

- **MEDIAN**
  The middle value in a series of values ordered from the smallest to the largest

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

Example:

245

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

240

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

Measures of Central Tendency

- **MEDIAN**
  The middle value in a series of values ordered from the smallest to the largest

```
median(data$sDur)
## [1] 0.118175

median(data$baseDur)
## [1] 0.306315

median(data$speakingRate)
## [1] 3.355
```

Measures of Central Tendency

- **MEDIAN**
  The middle value in a series of values ordered from the smallest to the largest

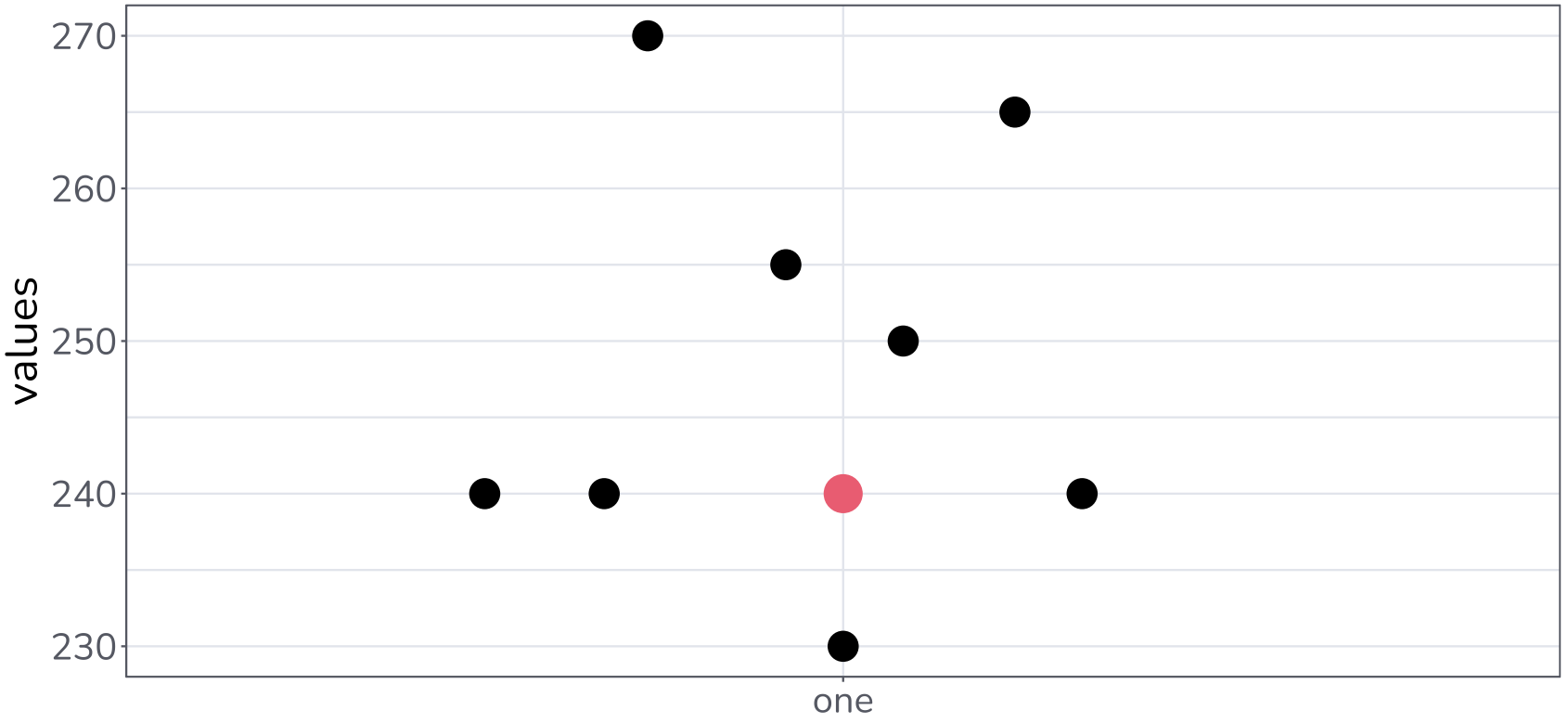
```
median(data$sDur[data$typeOfS == "nm"])
## [1] 0.15425

median(data$sDur[data$typeOfS == "pl"])
## [1] 0.121815

median(data$sDur[data$typeOfS == "is"])
## [1] 0.101505
```

Measures of Central Tendency

- **Mode**
  The value which appears most often in a set of values



Measures of Central Tendency

- **MODE**
  The value which appears most often in a set of values

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

Example:

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

Measures of Central Tendency

- **Mode**
  The value which appears most often in a set of values

```
mode_stat(data$sDur)
## [1] 0.1311

mode_stat(data$baseDur)
## [1] 0.25162

mode_stat(data$speakingRate)
## [1] 2.94
```

Measures of Central Tendency

- **Mode**
  The value which appears most often in a set of values

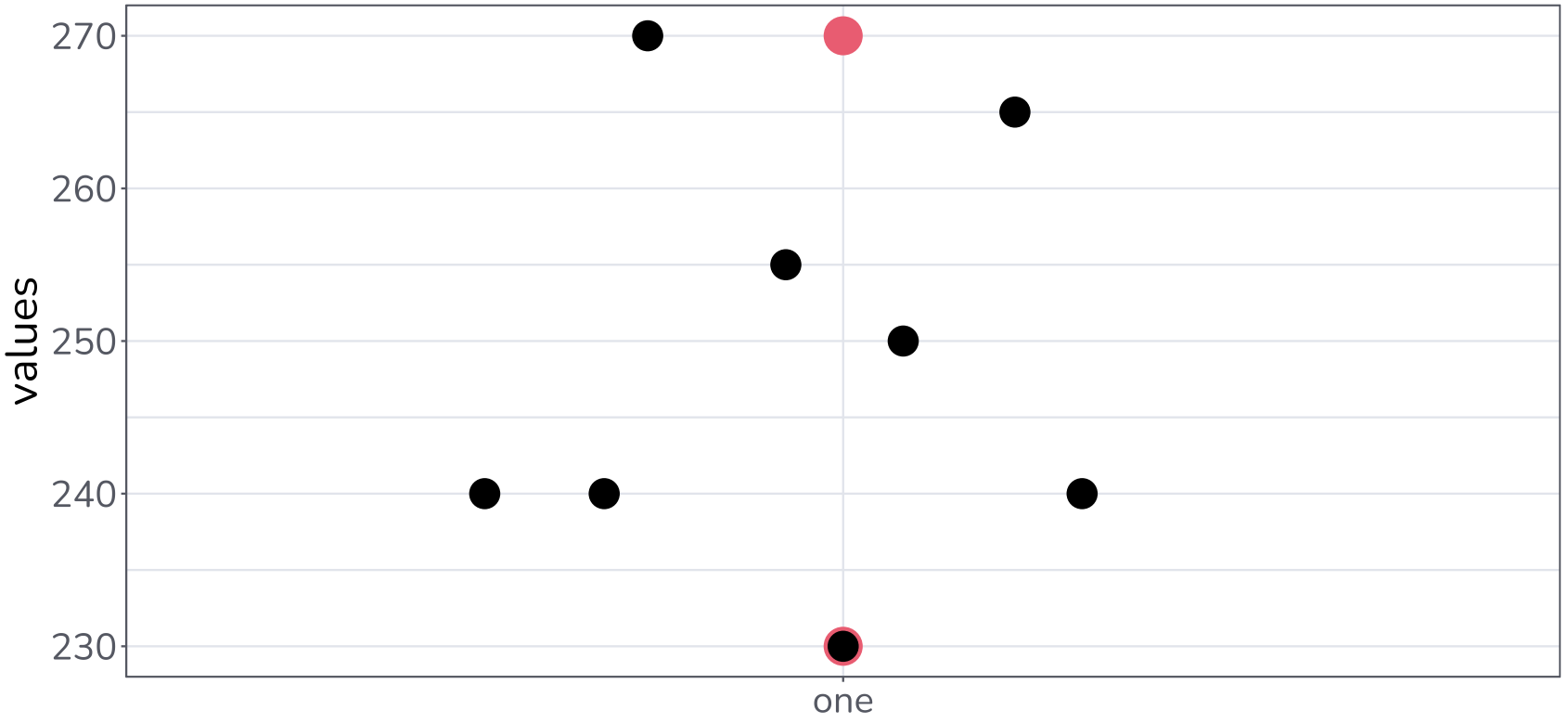
```
mode_stat(data$sDur[data$typeOfS == "nm"])
## [1] 0.096

mode_stat(data$sDur[data$typeOfS == "pl"])
## [1] 0.04176

mode_stat(data$sDur[data$typeOfS == "is"])
## [1] 0.1605
```

Measures of Dispersion

- **RANGE**
  The value between the smallest and the largest value in a set of values



Measures of Dispersion

- **RANGE**
  The value between the smallest and the largest value in a set of values

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

Example:

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

$$R = 280 - 230 = 50$$

Measures of Dispersion

- **RANGE**
  The value between the smallest and the largest value in a set of values

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

range(data$baseDur)
## [1] 0.17995 0.68749

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

Measures of Dispersion

- **RANGE**
  The value between the smallest and the largest value in a set of values

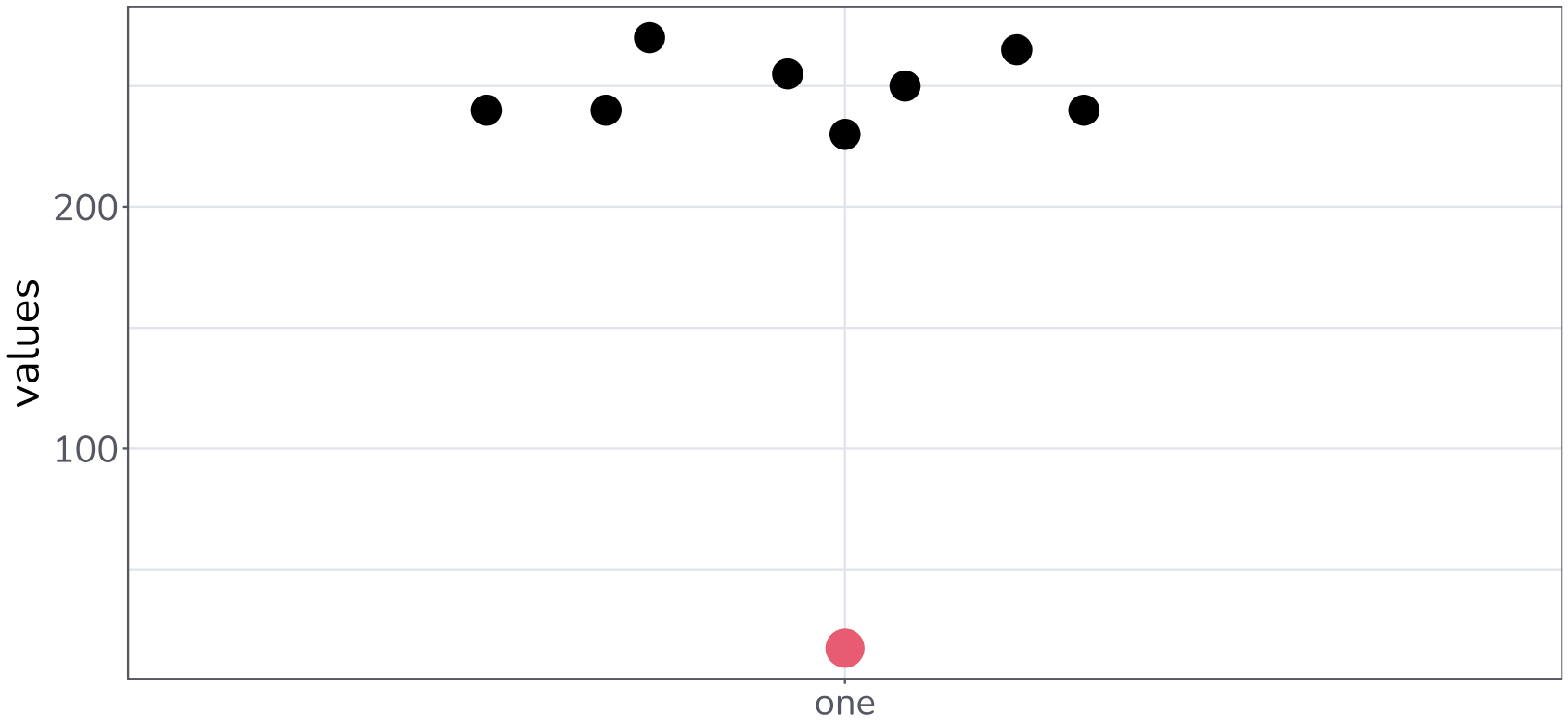
```
range(data$sDur[data$typeOfS == "nm"])
## [1] 0.05202 0.32750

range(data$sDur[data$typeOfS == "pl"])
## [1] 0.04176 0.25289

range(data$sDur[data$typeOfS == "is"])
## [1] 0.04435 0.22428
```

Measures of Dispersion

- **Interquartile Range**
  The range of the interval between the lower and the upper quartile



Measures of Dispersion

- **Interquartile Range**
  The range of the interval between the lower and the upper quartile

$$x_{IQM} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

Example:

1. 270, 240, 240, 255, 250, 265, 230, 240 > sort
2. 230, 240, 240, 240, 250, 255, 265, 270 > quartiles
3. ~~230, 240,~~ 240, 240, 250, 255~~, 265, 270~~ > remove 1st + 4th
4. $R = 255 - 240 = 15$ > range

Measures of Dispersion

- **INTERQUARTILE RANGE**
  The range of the interval between the lower and the upper quartile

```
IQR(data$sDur)
## [1] 0.06783

IQR(data$baseDur)
## [1] 0.1067575

IQR(data$speakingRate)
## [1] 1.125
```

Measures of Dispersion

- **Interquartile Range**
  The range of the interval between the lower and the upper quartile

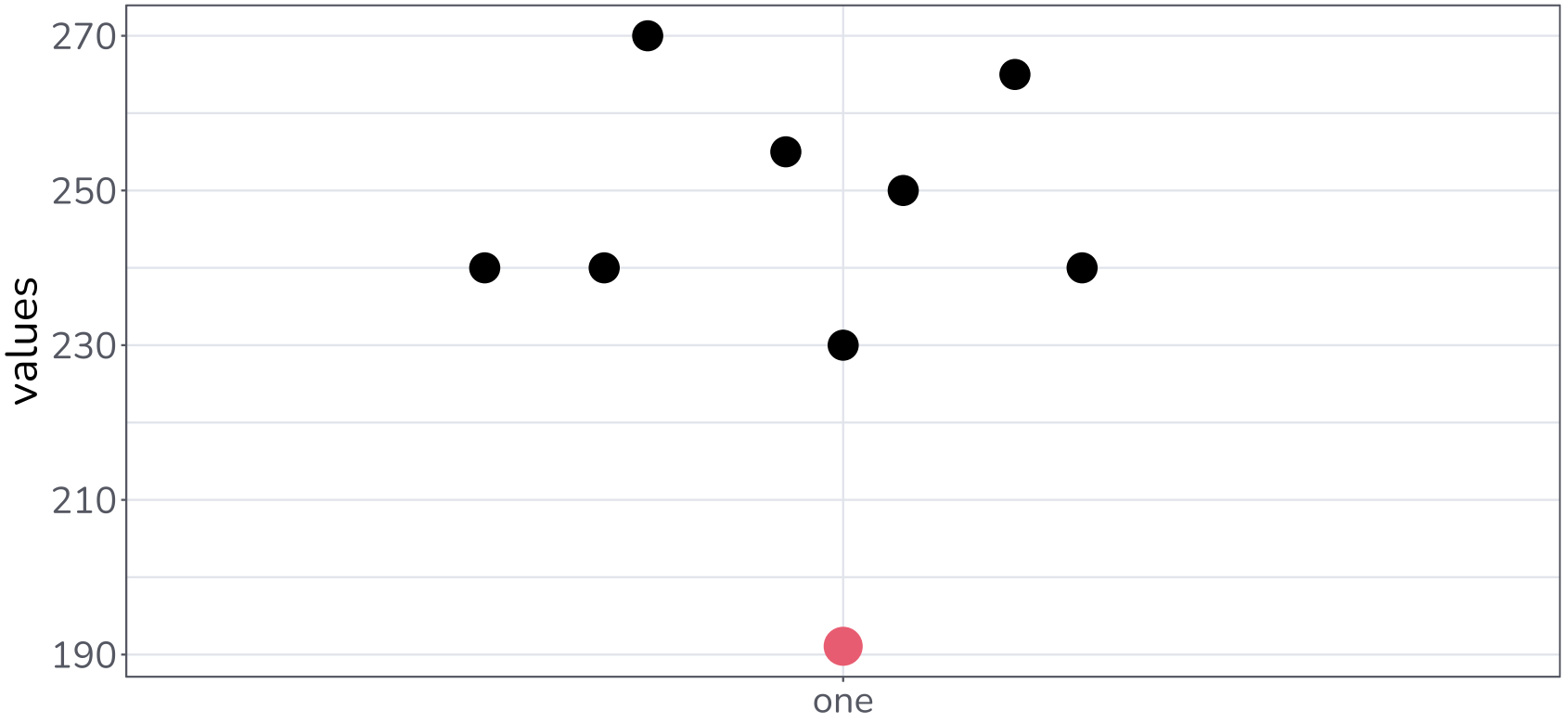
```
IQR(data$sDur[data$typeOfS == "nm"])
## [1] 0.0910275

IQR(data$sDur[data$typeOfS == "pl"])
## [1] 0.072535

IQR(data$sDur[data$typeOfS == "is"])
## [1] 0.0363475
```

Measures of Dispersion

- **SAMPLE VARIANCE**
  A numerical measure of how the data values are dispersed around the mean



Measures of Dispersion

- **Sample Variance**
  A numerical measure of how the data values are dispersed around the mean

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

Example:

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

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

Measures of Dispersion

- **SAMPLE VARIANCE**
  A numerical measure of how the data values are dispersed around the mean

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

var(data$baseDur)
## [1] 0.007913081

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

Measures of Dispersion

- **SAMPLE VARIANCE**
  A numerical measure of how the data values are dispersed around the mean

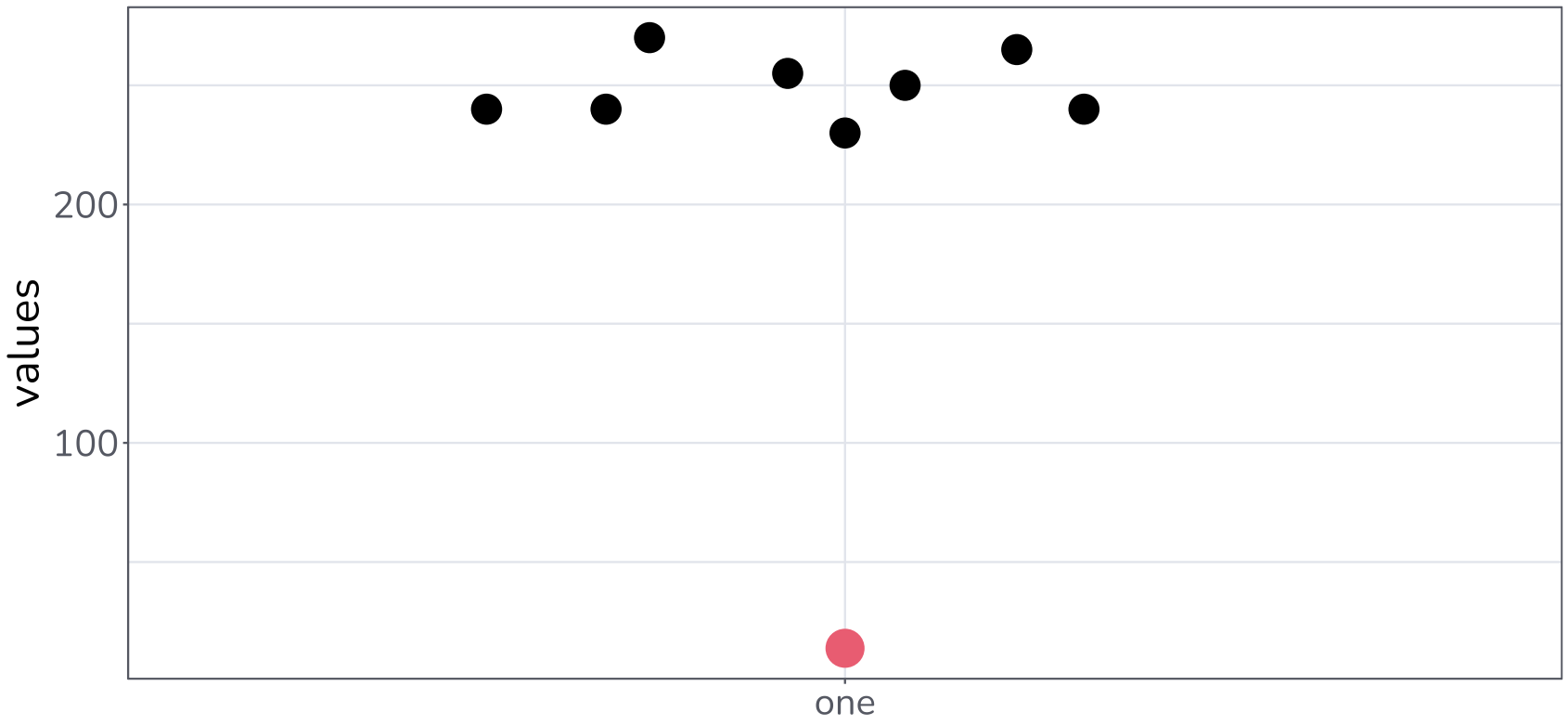
```
var(data$sDur[data$typeOfS == "nm"])
## [1] 0.003943441

var(data$sDur[data$typeOfS == "pl"])
## [1] 0.002601761

var(data$sDur[data$typeOfS == "is"])
## [1] 0.001255514
```

Measures of Dispersion

- **Standard Deviation**
  An indication of the overall distance of individual values from the mean



Measures of Dispersion

- **Standard Deviation**
  An indication of the overall distance of individual values from the mean

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

square root of the variance!

Example:

230, 240 [illegible] 235, 265, 270

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

Measures of Dispersion

- **Standard Deviation**
  An indication of the overall distance of individual values from the mean

```
sd(data$sDur)
## [1] 0.05468424

sd(data$baseDur)
## [1] 0.0889555

sd(data$speakingRate)
## [1] 0.9300259
```

Measures of Dispersion

- **Standard Deviation**
  An indication of the overall distance of individual values from the mean

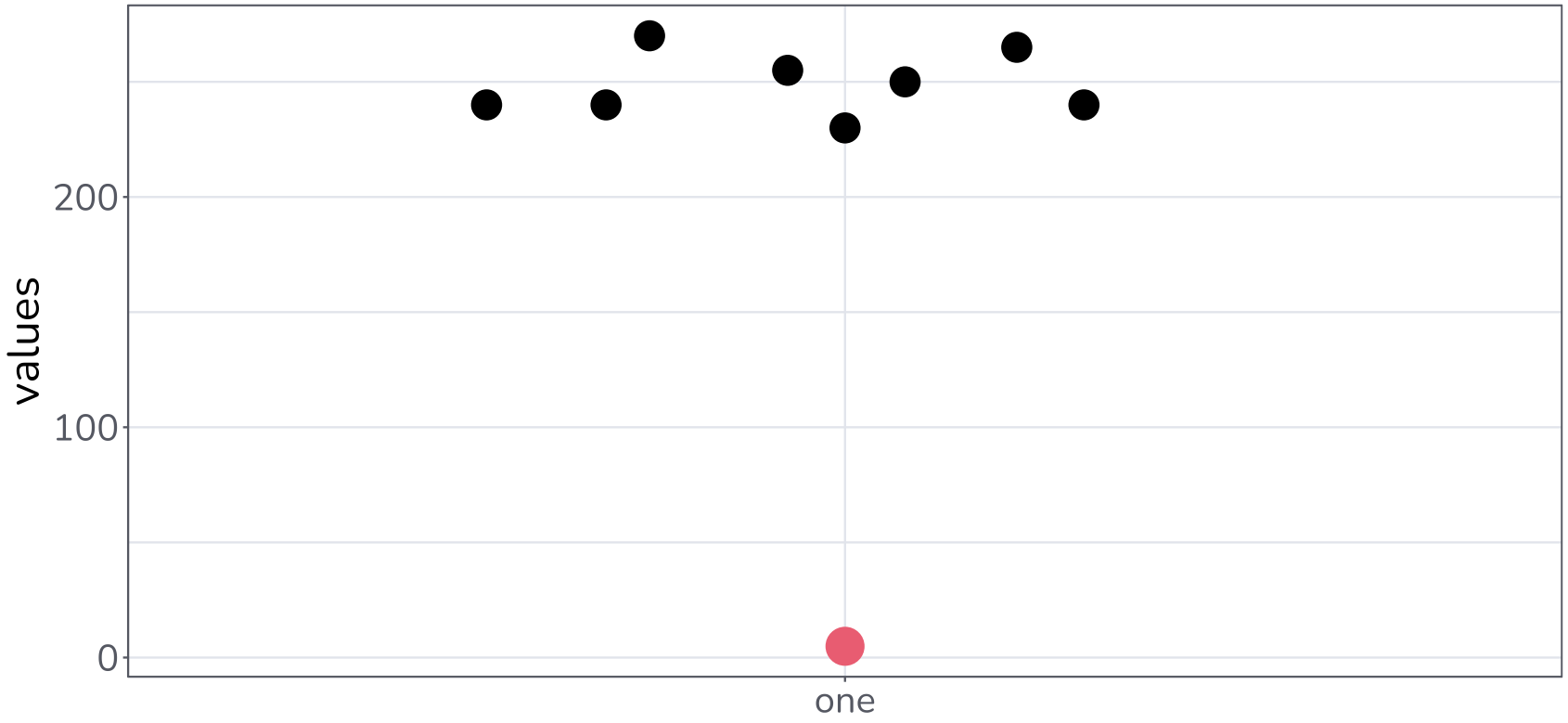
```
sd(data$sDur[data$typeOfS == "nm"])
## [1] 0.06279683

sd(data$sDur[data$typeOfS == "pl"])
## [1] 0.05100746

sd(data$sDur[data$typeOfS == "is"])
## [1] 0.03543323
```

Measures of Dispersion

- **Standard Error**
  A statistical term that measures the accuracy with which a sample represents a population



Measures of Dispersion

- **Standard Error**
  A statistical term that measures the accuracy with which a sample represents a population

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

σ being the standard deviation of [illegible] the population)

standard deviation divided by square root of sample size

Example:

2[illegible], 240, 250, 255, 265, 270

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

Measures of Dispersion

- **Standard Error**
  A statistical term that measures the accuracy with which a sample represents a population

```
std.error(data$sDur)
## [1] 0.004464949

std.error(data$baseDur)
## [1] 0.007263186

std.error(data$speakingRate)
## [1] 0.0759363
```

Measures of Dispersion

- **Standard Error**
  A statistical term that measures the accuracy with which a sample represents a population

```
std.error(data$sDur[data$typeOfS == "nm"])
## [1] 0.008880812

std.error(data$sDur[data$typeOfS == "pl"])
## [1] 0.007213545

std.error(data$sDur[data$typeOfS == "is"])
## [1] 0.005011015
```

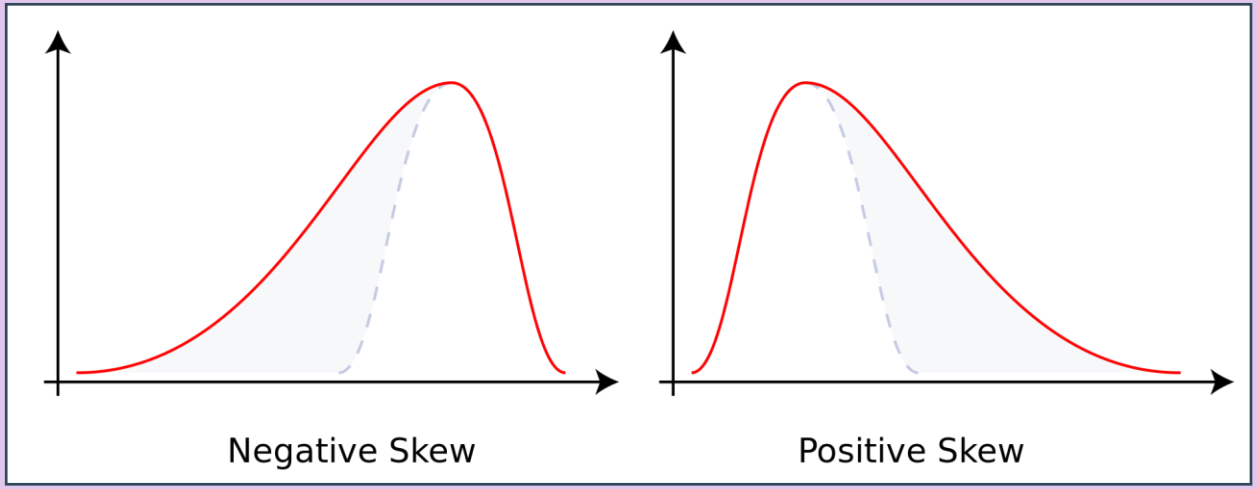
Shape of Distribution

- **SKEWNESS**
  Asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right

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

$\bar{x}$ = mean s = deviation

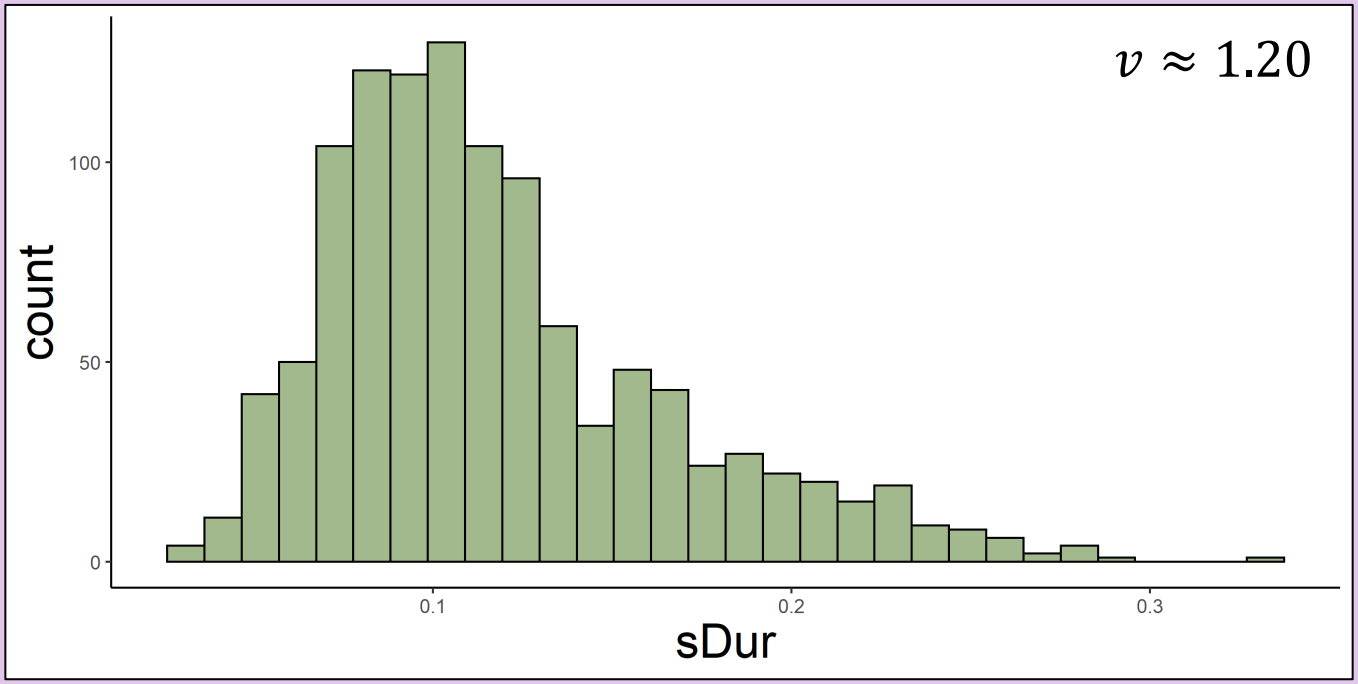
Example:



Shape of Distribution

- **SKEWNESS**
  Most linguistic data is positively skewed, i.e. there is more data to the left of the distribution than to the right

Example:



Shape of Distribution

- **SKEWNESS**
  Asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right

```
skewness(data$sDur)
## [1] 0.9483159

skewness(data$baseDur)
## [1] 1.360664

skewness(data$speakingRate)
## [1] 0.8348821
```

Shape of Distribution

- **SKEWNESS**
  Asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right

```
skewness(data$sDur[data$typeOfS == "nm"])
## [1] 0.5884803

skewness(data$sDur[data$typeOfS == "pl"])
## [1] 0.6259893

skewness(data$sDur[data$typeOfS == "is"])
## [1] 0.8515867
```