# Vorlesungsskript Statistik - Version 1

Erich Neuwirth

12. Jänner 2021

## Contents



# Statistische Daten

## Skalenninveau

- Nominalskala
- Ordinalskala (diskret oder stetig)
- Intervallskala (diskret oder stetig)
- Verhältnissskala (diskret oder stetig)

Skalen können diskret (isolierte Werte) oder stetig (zwischen je 2 Werten ist ein weiterer Wert möglich) sein

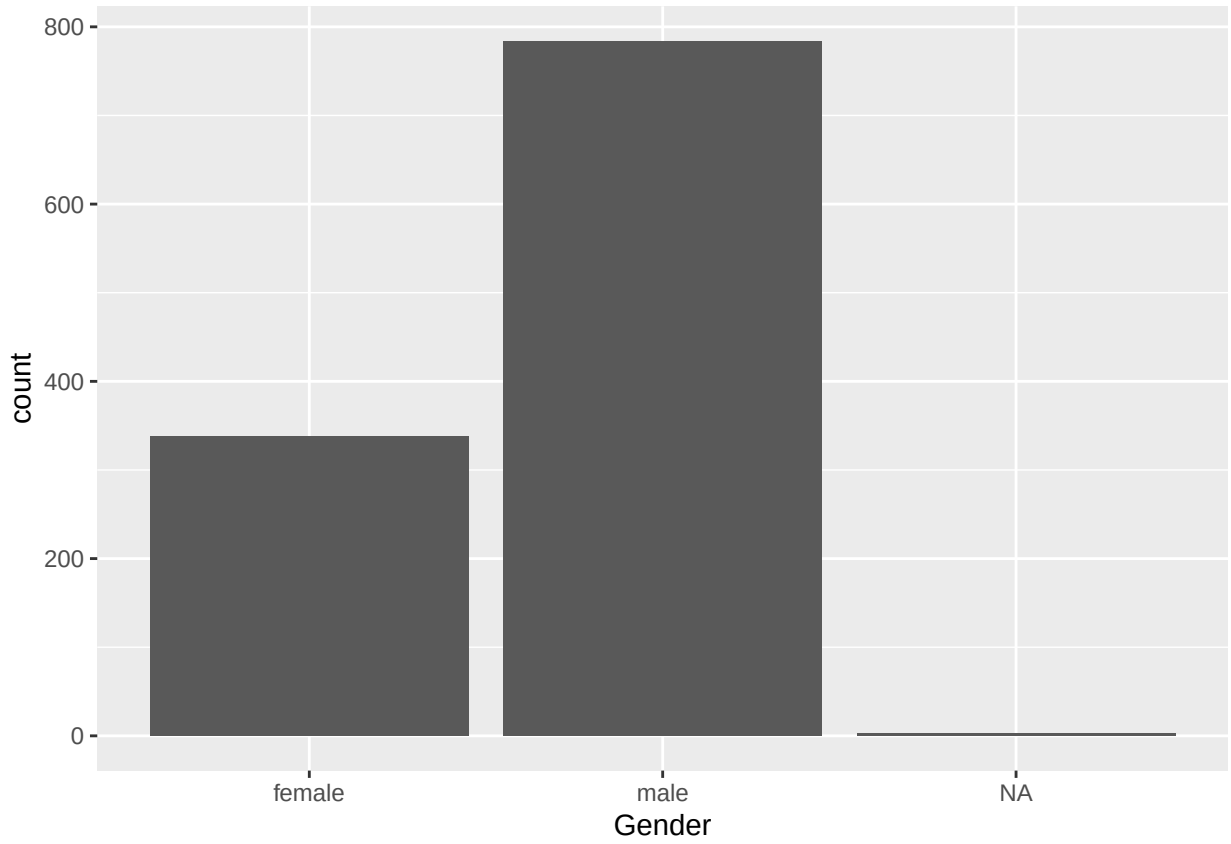
## Darstellung Häufigkeiten (univariat)

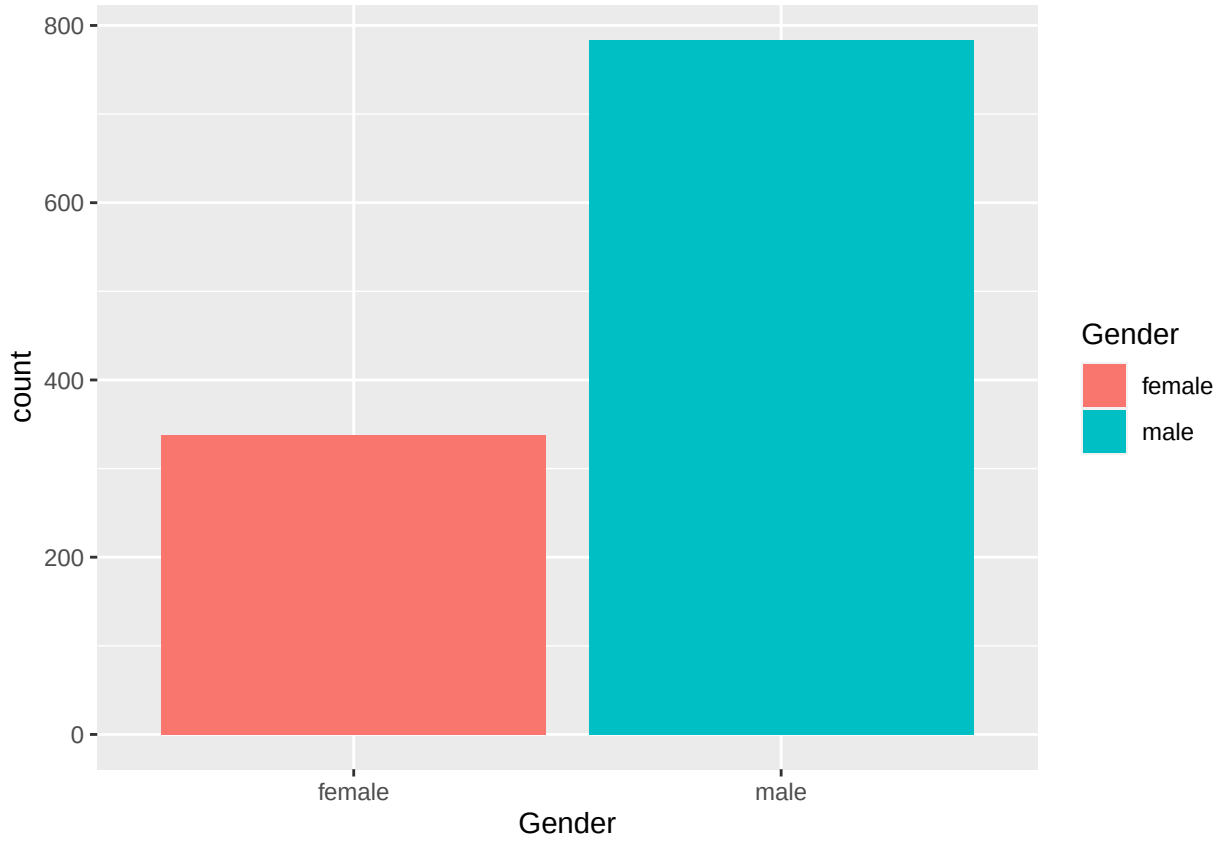
### Nominalskala und Ordinal

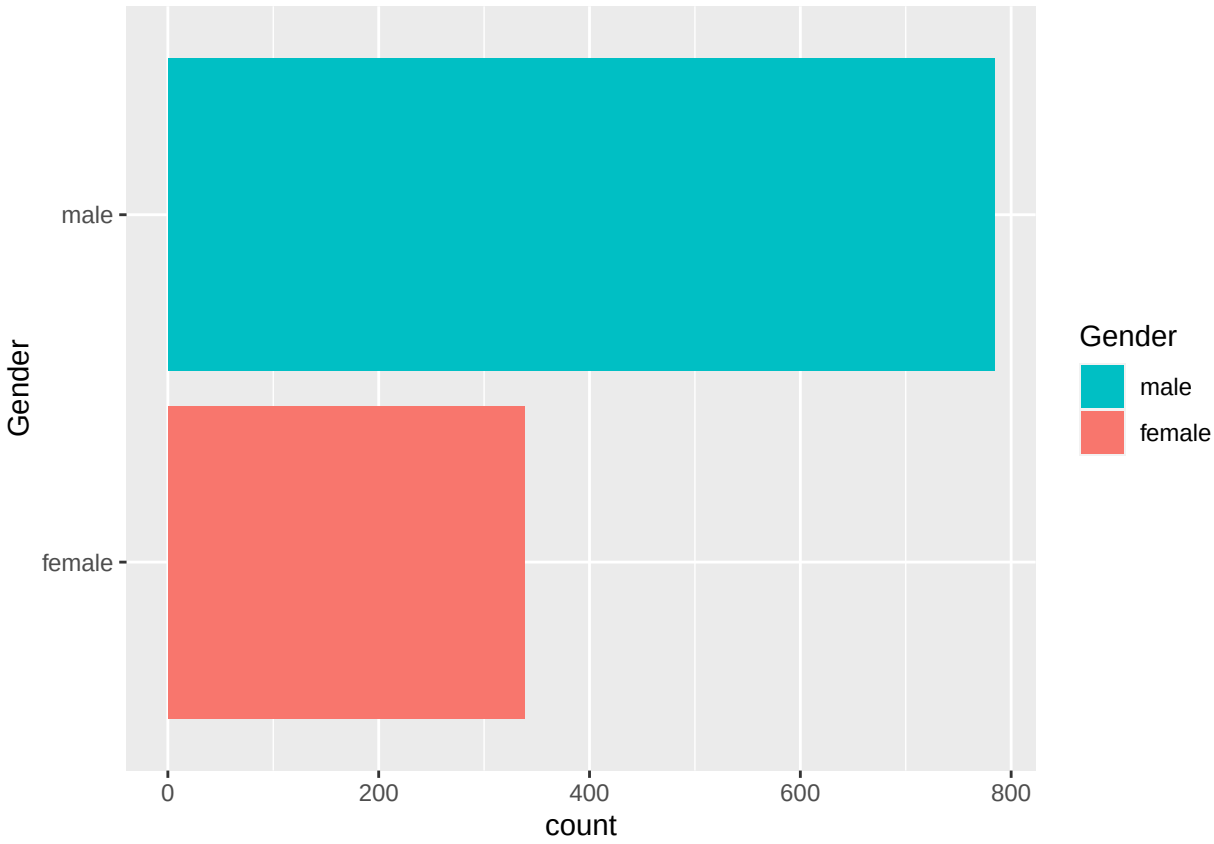
- Stabdiagramme
- Balkendiagramme

Zentraler Begriff in der statistischen Auswertung

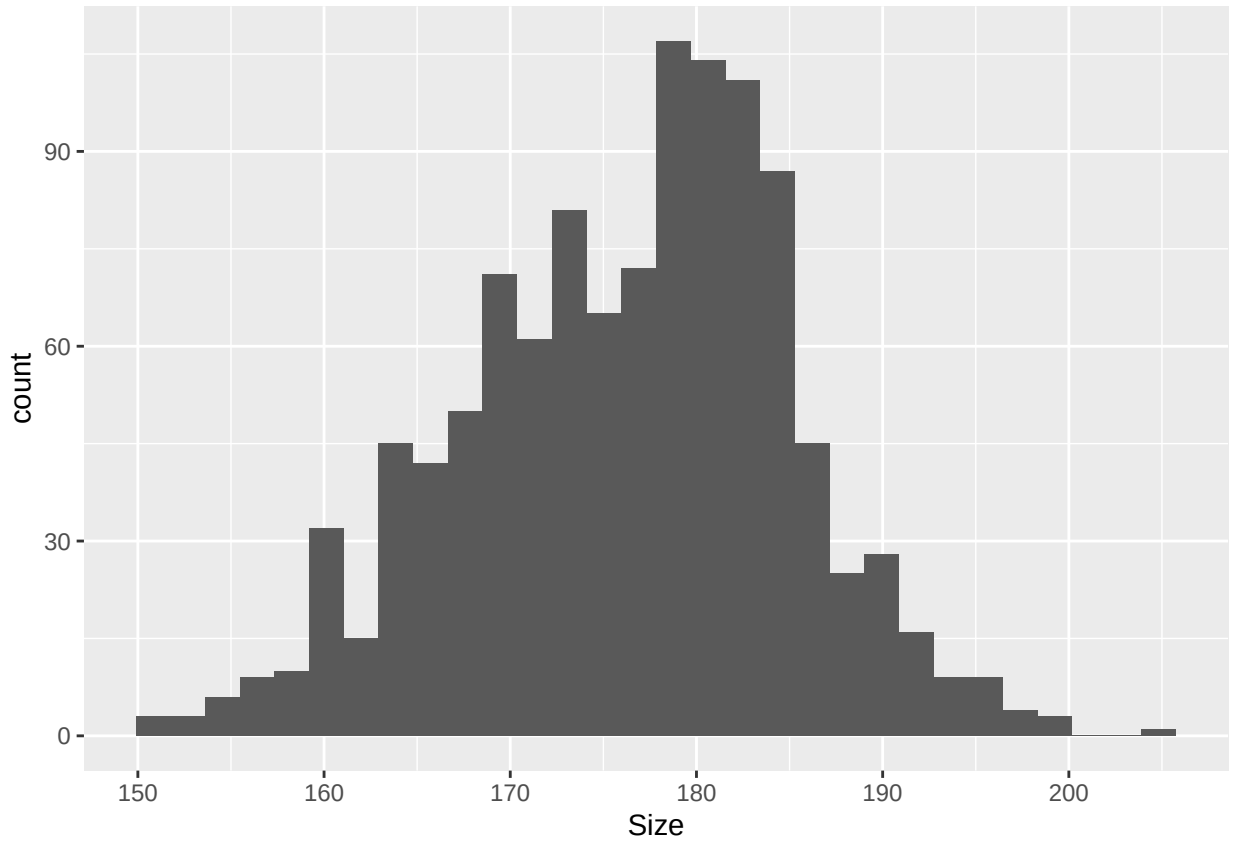
`data.frame` oder Datentabelle

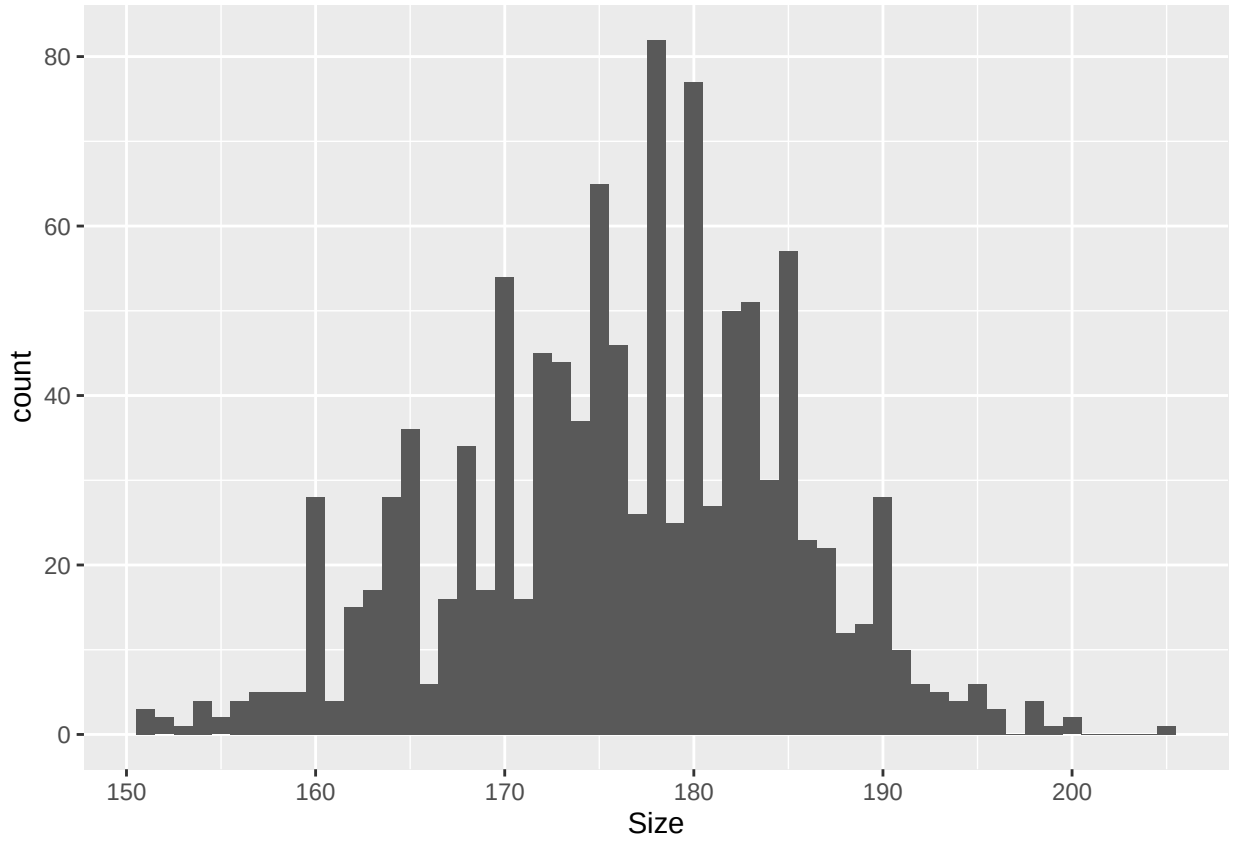

800
600
400
200
0
count
female
male
NA
Gender

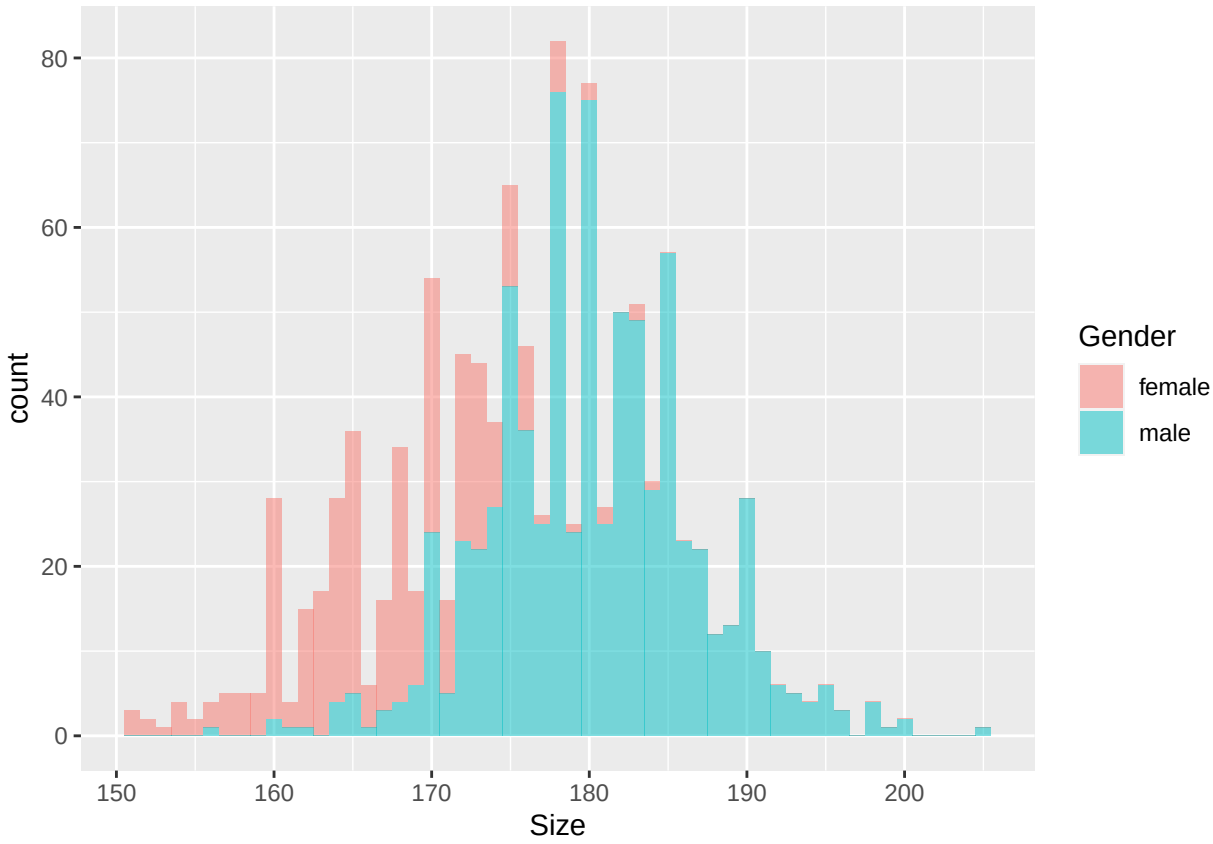

800
600
400
200
0
count
female
male
Gender
Gender
female
male




Man unterscheidet bei Intervall- und Verhältnisskalen und zusätzlich noch diskrete und stetige (kontinuierliche) Skalen.


count
90
60
30
0
150
160
170
180
190
200
Size


count
80
60
40
20
0
150
160
170
180
190
200
Size



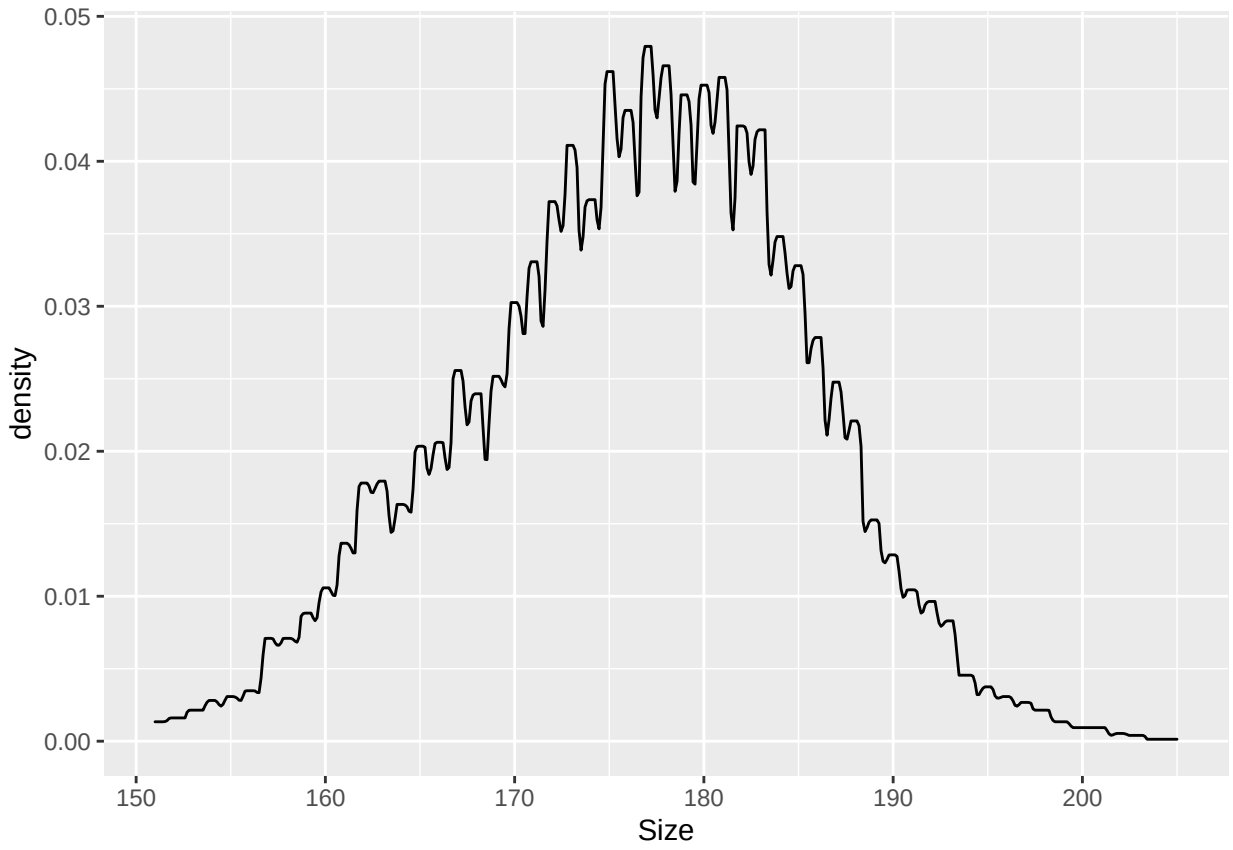

Überlappungen nicht besonders sauber dargestellt.

Besser Kerndichteschätzer

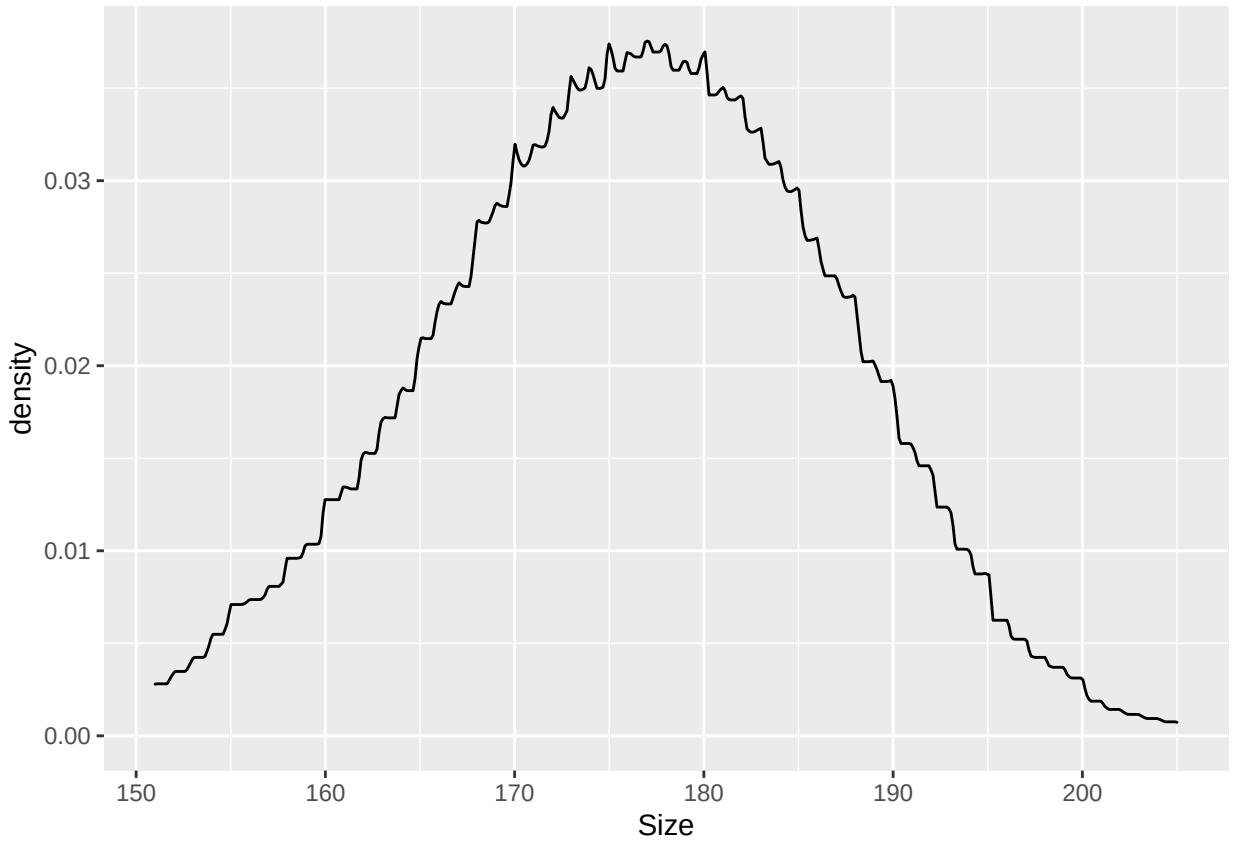
$$f(x|x_1 \dots x_n) = \frac{1}{nw} \sum_{i=1}^{n} k\left(\frac{x - x_i}{w}\right)$$

$k$ ist der Kern, die Funktion $k$ hat Gesamtfläche 1.
$w$ ist die Fensterbreite $n$ ist die Anzahl der Datenpunkte

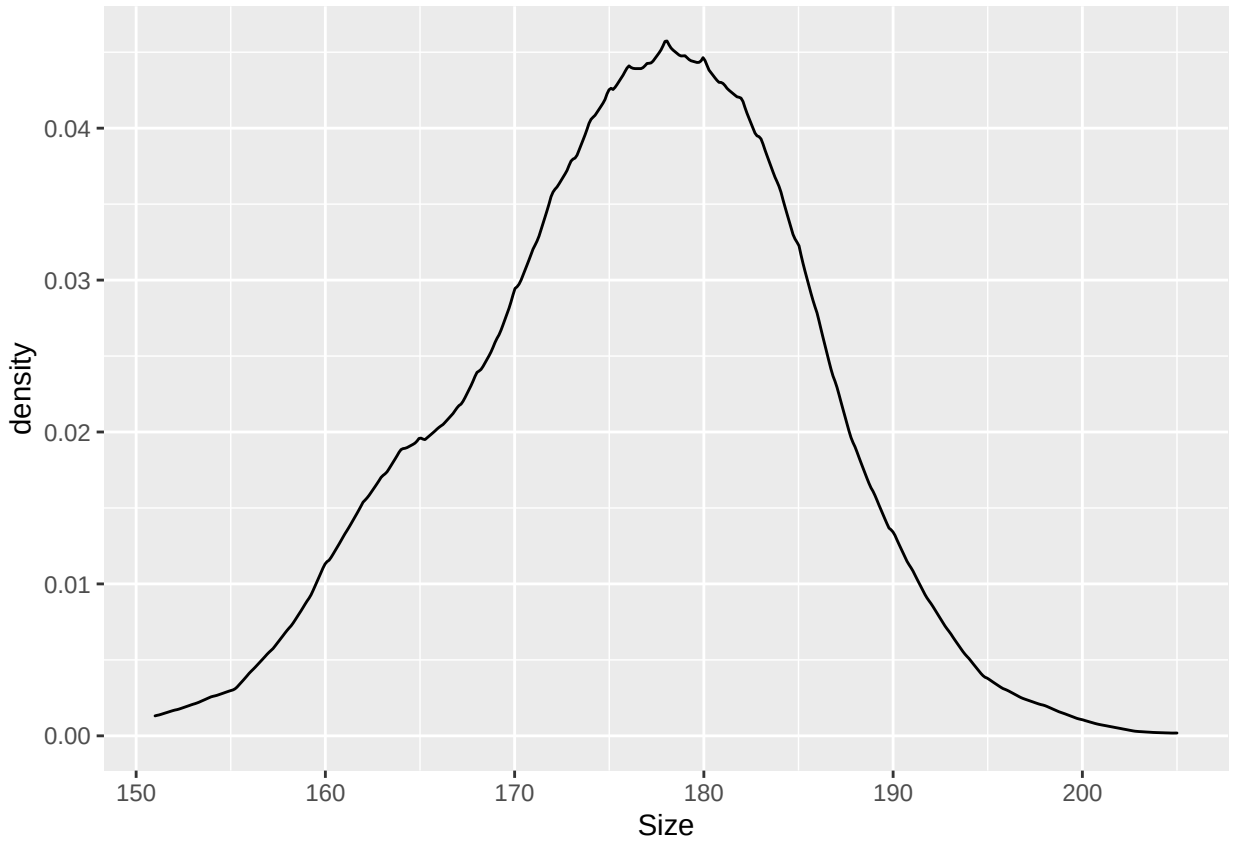
Typische Kerne * Rechteckskern * Dreieckskern * Gauß-Kern



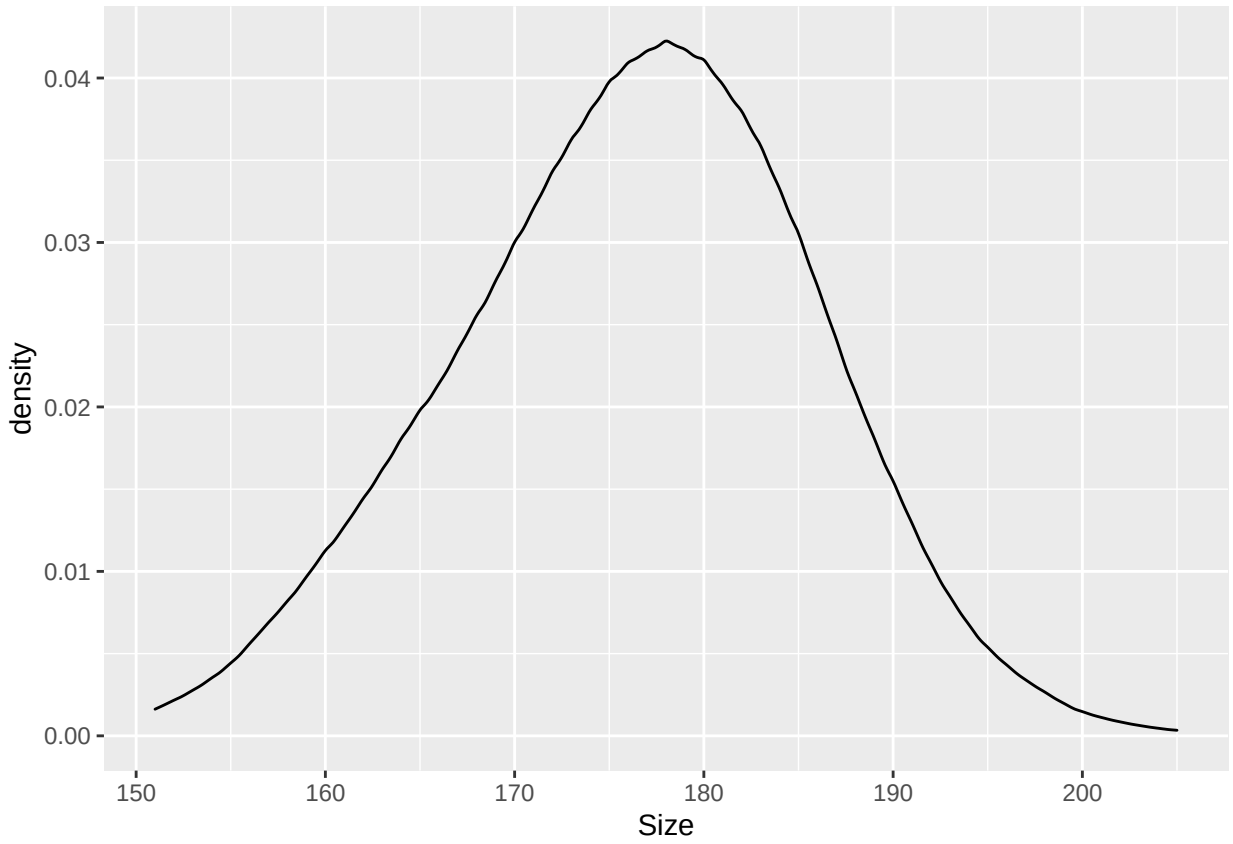

Fensterbreite verdreifacht



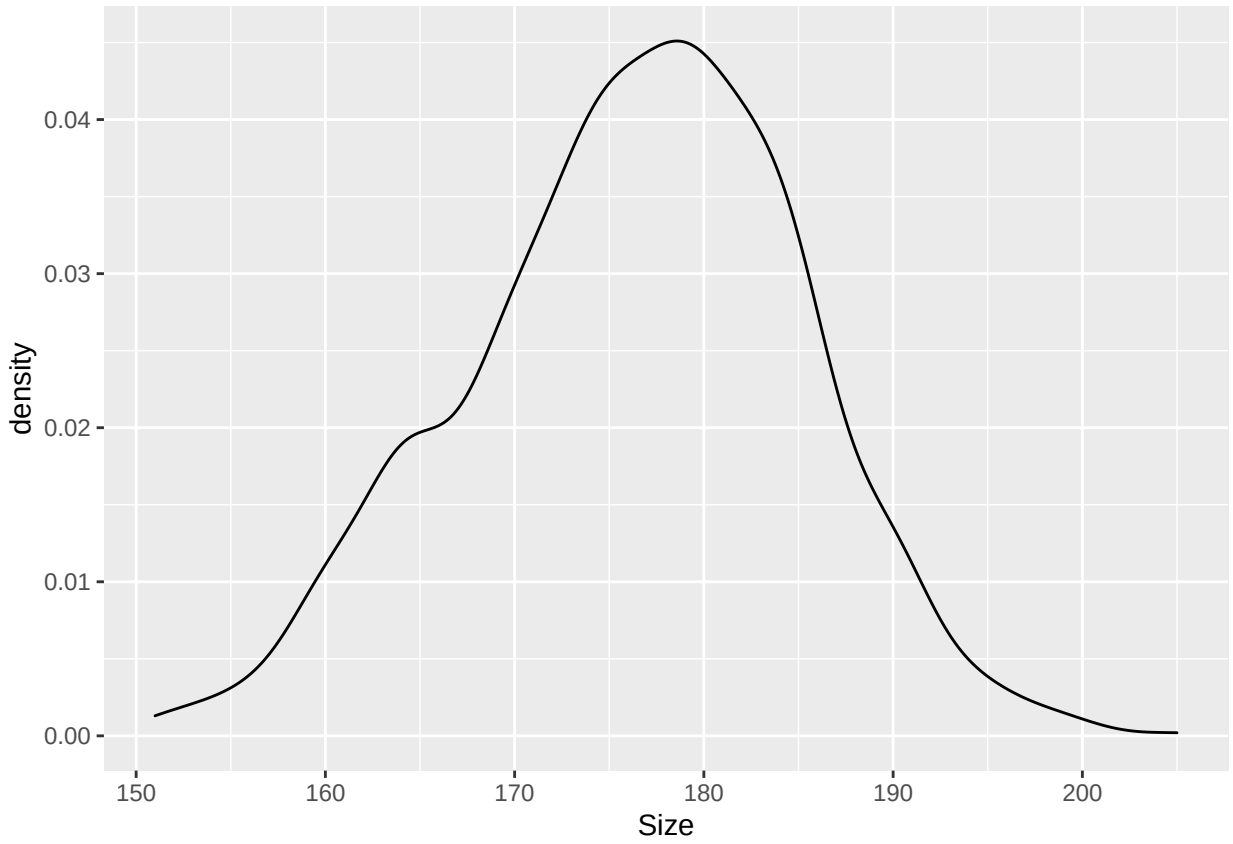

Dreieckskern



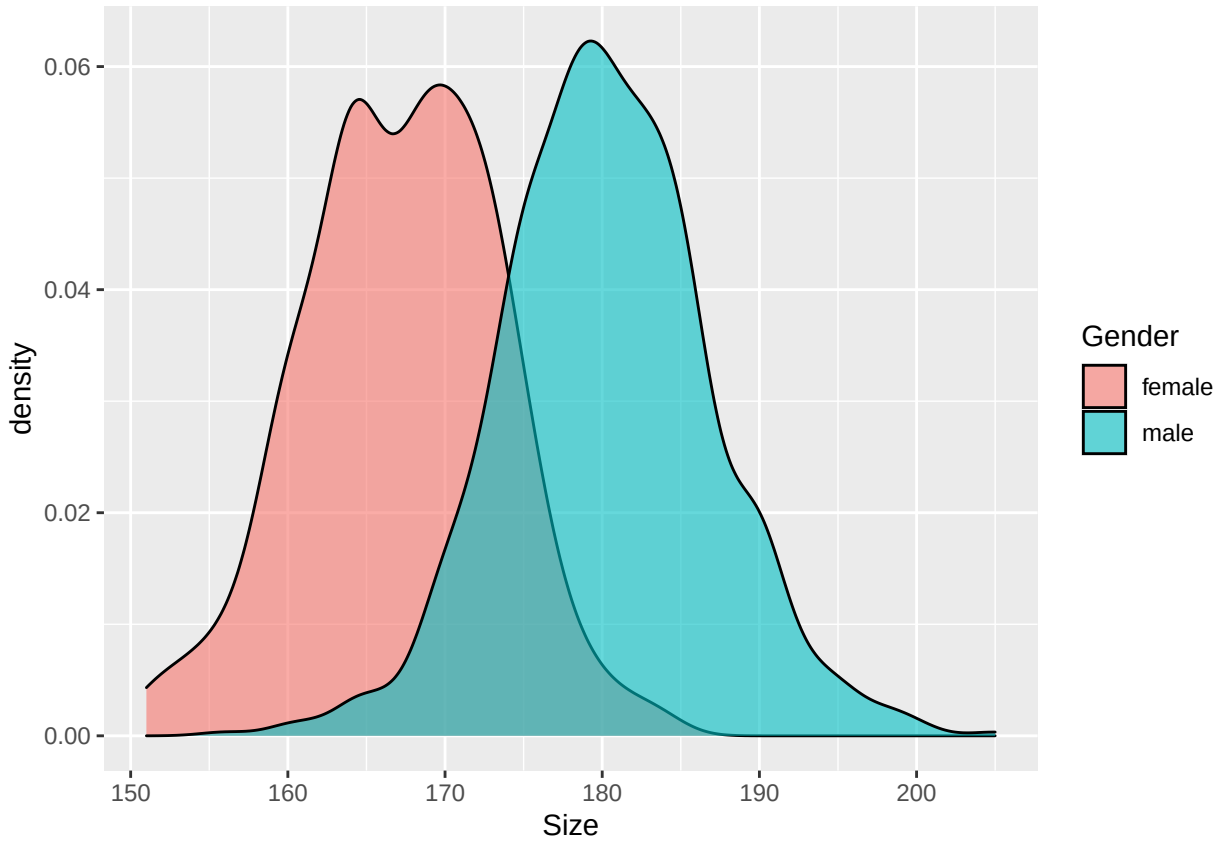

Dreieckskern




Kern Normalverteilungsdichte




Zweigipfeligkeit oft Hinweis auf 2 zusammengemischte Grundgesamtheiten




# Statistische Maßzahlen

## Lagemaßzahlen

### Arithmetisches Mittel

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

$\bar{x}$ löst die Minimumsaufgabe $\sum_{i=1}^{n}(x_i - c)^2$, dieser Ausdruck ist kleinstmöglich für $c = \bar{x}$

Arithmetisches Mittel sinnvoll bei Verhältnisskala, Intervallskala und stark eingeschränkt auch bei Ordinalskala

### Median

$x_{(1)}, x_{(2)} \dots x_{(n)}$ sind die der Größe nach geordneten Werte

Medien $\tilde{x}$ ist der Wert „in der Mitte“

$n$ ungerade $\tilde{x} = x_{(\frac{n+1}{2})}$

$n$ gerade $\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$

$\tilde{x}$ löst die Minimumsaufgabe $\sum_{i=1}^{n} |x_i - c|$, dieser Ausdruck ist kleinstmöglich für $c = \tilde{x}$

### Modus

Häufigster Wert

Sinnvoll bei allen Skalen

Lagemaße bei quantitativen Merkmalen

$L(x_1 + a, x_2 + a, \dots x_n + a) = L(x_1, x_2, \dots x_n) + a$

Wird zu allen Zahlen derselbe Wert addiert, dann ist der neue Mittelwert gleich dem alten Mittelwert plus dieser Zahl (das nennt man equivariant unter Translation)

$L(cx_1, cx_2, \dots cx_n) = cL(x_1, x_2, \dots x_n)$

Werden allen Zahlen mit demselben Wert multipliziert, dann ist der neue Mittelwert gleich dem alten Mittelwert multipliziert mit diesem Wert (das nennt man equivariant unter Reskalierung)

Es gibt noch anderer Mittelwerte

**Geometrisches Mittel** $\sqrt[n]{\prod_{i=1}^{n} x_i}$

Wird verwendet z.B. bei Wachstumsraten

**Harmonisches Mittel** $\frac{1}{\frac{1}{n}\sum_1^n \frac{1}{x_i}} = \frac{n}{\sum_1^n \frac{1}{x_i}}$

Verwendet bei Mittelung von Geschwindigkeit für vorgegebene Teilstrecken

Allgemeiner $p$-Mittelwert

$$x_p = \sqrt[p]{\sum_{i=1}^{n} {x_i}^p}$$

Nur bei positiven $x_i$ sinnvoll

$p = 1$ arithmetisches Mittel
$p = 2$ quadratisches Mittel
$p = -1$ harmonisches Mittel
$p = 0$ $\lim_{p\to 0}$ geometrisches Mittel
$p = \infty$ $\lim_{p\to\infty} \max x_i$
$p = \infty$ $\lim_{p\to-\infty} \min x_i$

Die $p$-Mittelwerte mit $p \neq 1$ sind keine Lagemaße, sie sind nicht translationsequivariant.

## Verteilungsfunktion

$$F(x) = \frac{1}{n}\#\{x_i | x_i \leq x\}$$

Quartile und Quantile

$Q_1$ definiert durch $F(Q_1) = \frac{1}{4}$ (bzw. präzisiert $F(x) \leq \frac{1}{4}$) für $x \leq Q_1$ und $F(x) \geq \frac{1}{4}$ für $x \geq Q_1$

$Q_3$ definiert durch $F(Q_1) = \frac{3}{4}$ (bzw. präzisiert $F(x) \leq \frac{3}{4}$) für $x \leq Q_3$ und $F(x) \geq \frac{3}{4}$ f¨¨r $x \geq Q_3$

Allgemein $p$-Quantil mit $0 \leq p \leq 1$

$Q_p$ definiert durch $F(Q_1) = p$ (bzw. präzisiert $F(x) \leq p$) für $x \leq Q_p$ und $F(x) \geq p$ für $x \geq Q_p$

$Q_{0.5}$ ist der Median

## Streuungsmaße

Standardabweichung

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

Semiinterquartilsdistanz

$$\frac{Q_3 - Q_1}{2}$$

Spannweite (nicht sehr empfehlenswert als Streuungsmaß weil sehr empfindlich gegen Ausreißer)

$$\max(x_i) - \min(x_i)$$

## Schiefemaßzahlen

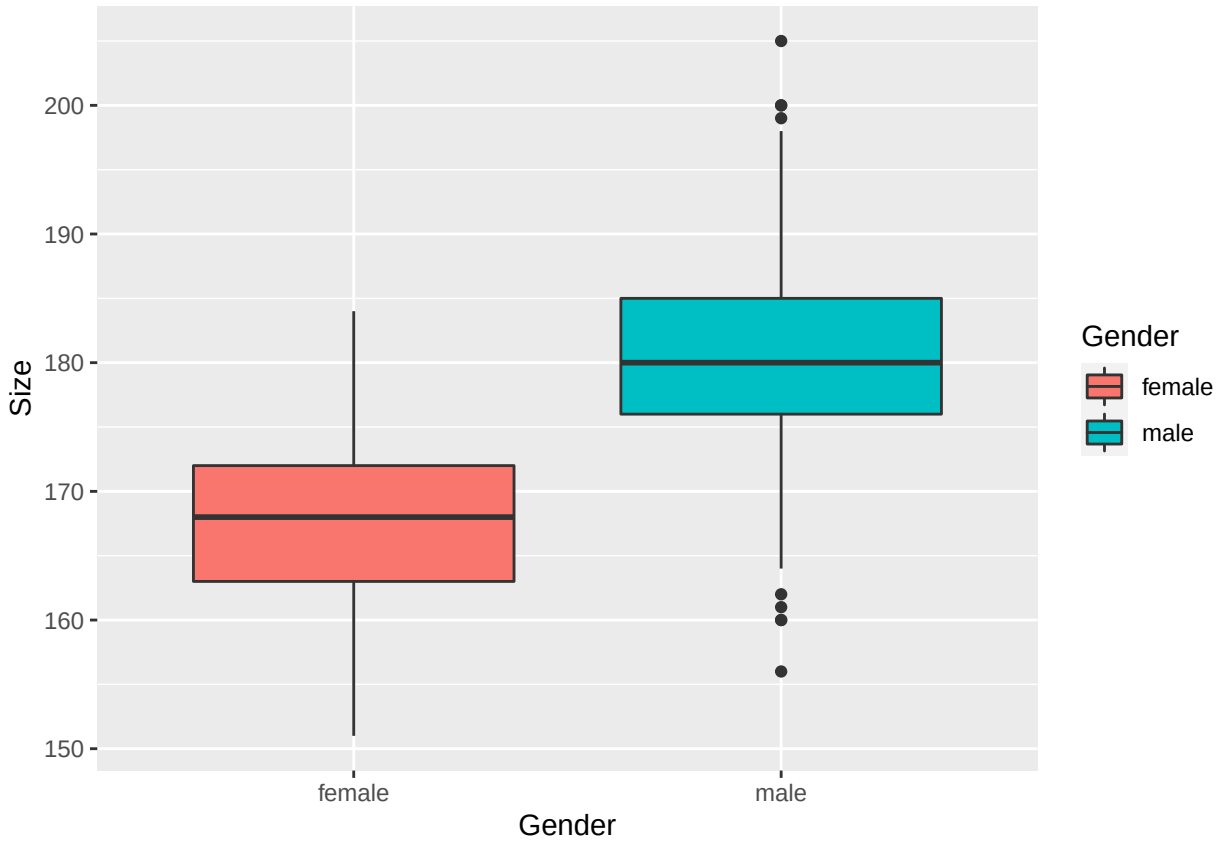
Schiefemaß nach Pearson

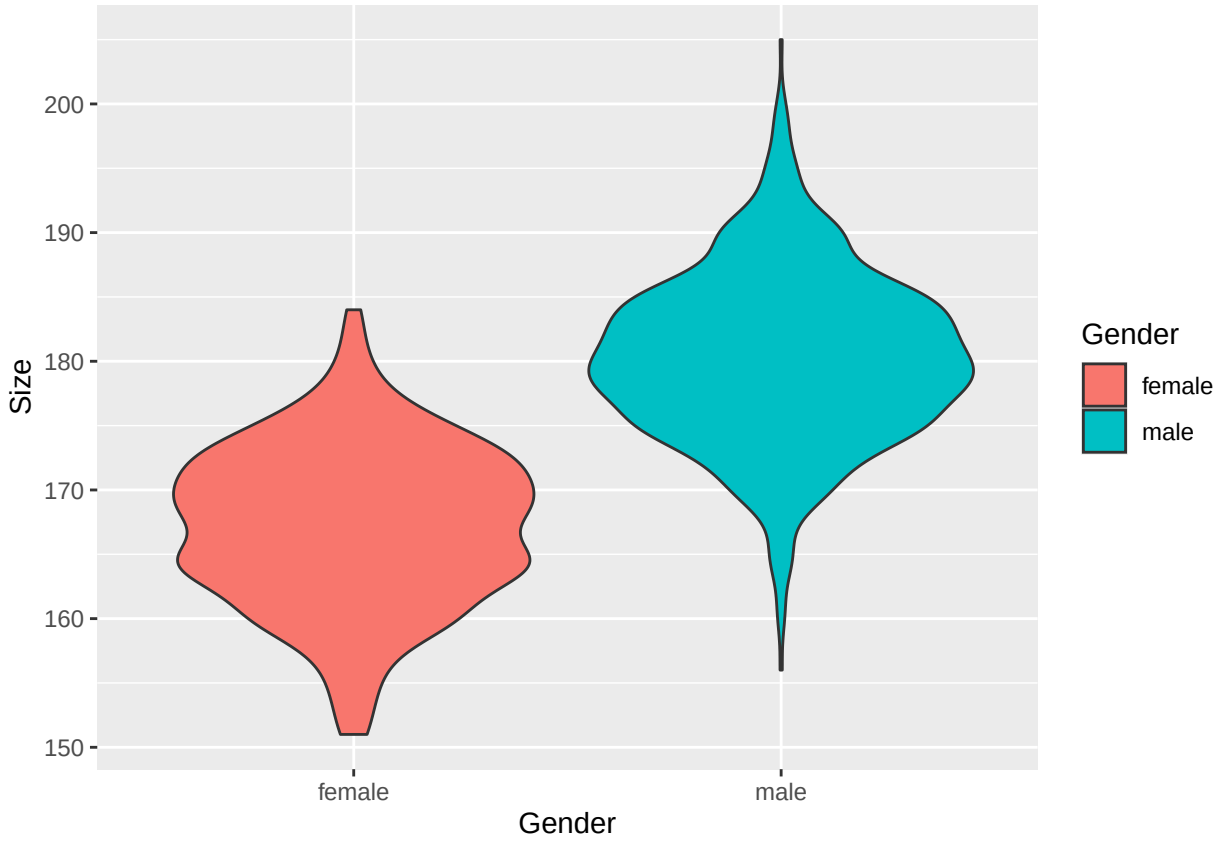
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}$$

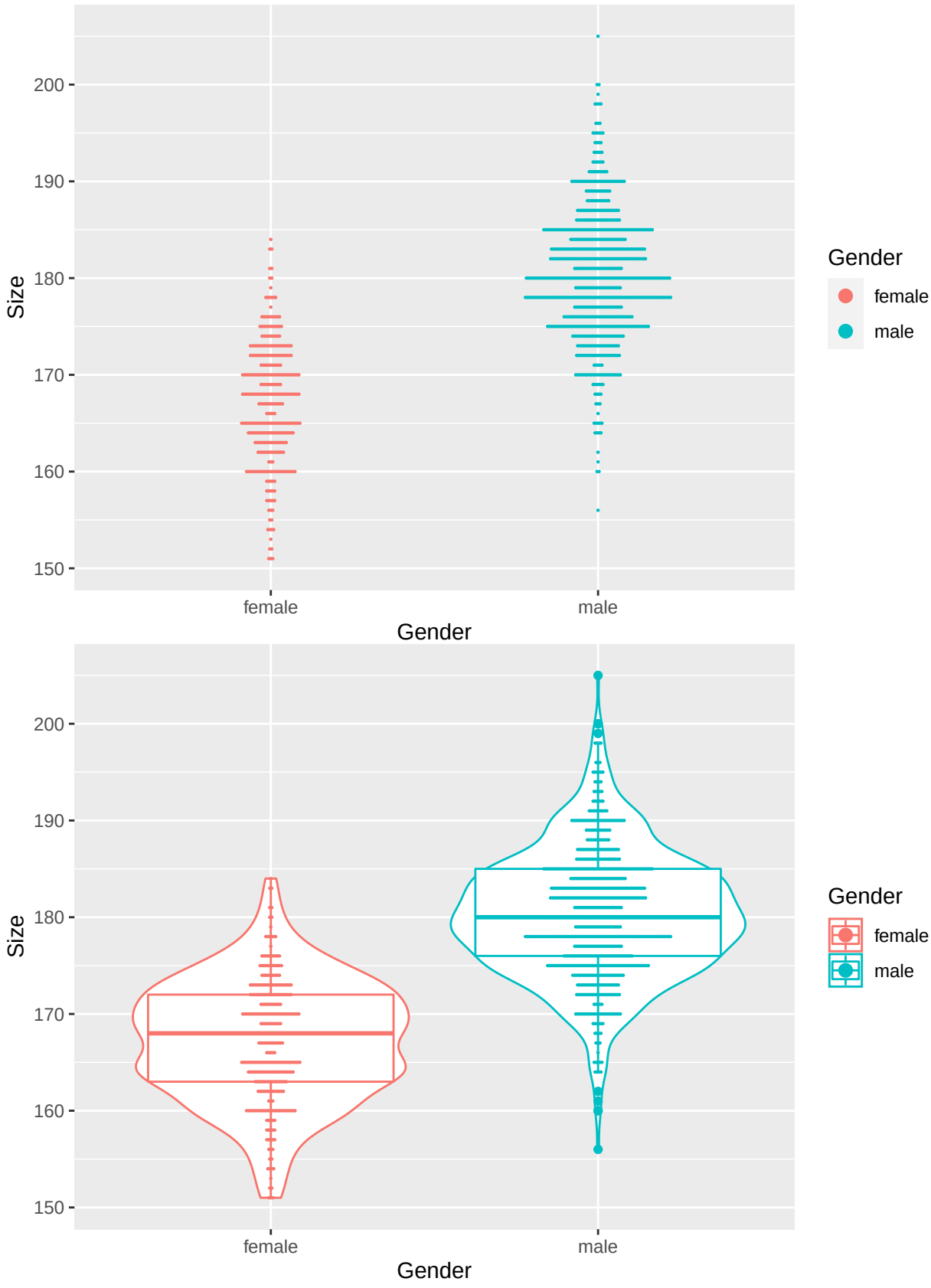
Schiefemaß mit Quartilen

$$\frac{|Q3 - Q2| - |Q2 - Q1|}{Q3 - Q1}$$

Boxplots


200
190
180
170
160
150
Size
female
male
Gender
Gender
female
male


200
190
180
170
160
150
Size
female
male
Gender
Gender
female
male


Size
200
190
180
170
160
150
female
male
Gender
Gender
female
male
Size
200
190
180
170
160
150
female
male
Gender
Gender
female
male

## Konzentrationsmaße

$x_{(i)}$ sind die der Größe nach geordneten $x_i$

Lorenz-Kurve

Welcher Anteil der Merkmalsträger besitzt welchen Anteil an der Merkmalssumme?

Beispiel: Welchen Anteil am Gesamteinkommen verdienen die 10% mit dem niedrigsten Einkommen?

$n$ Datenpunkte, $x_{(j)}$ mit $j \leq i$ sind Anteil $\frac{i}{n}$ der Merkmalsträger und haben $\frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$ der Merkmalssumme.

Man zeichnet die Punkte $\left(\frac{i}{n}, \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}\right)$ für alle $i$ $(1 \leq i \leq n)$ und verbindet sie, das ist die Lorenzkurve

Das Doppelte der Fläche zwischen der Lorenzkurve und der 45°-Geraden heißt Gini-Koeffizient

Herfindahl-Index

Mit $p_i = \frac{x_i}{\sum_{j=1}^{n} x_j}$ (Anteil des Werts eines einzelnen Merkmalsträgers an der Gesamtsumme)

$$H = \sum_{i=1}^{n} {p_i}^2$$

Herfindahl misst absolute Konzentration (wenige haben viel), Gini-Koeffizient misst relative Konzentration (ein relativ kleiner Anteil der Merkmalsträger besitzt einen großen Teil der Gesamtsumme)

Konzentrationsrate

$$CR_g = \text{Summe der } g \text{ größten } p_i$$

Welchen Anteil an der Gesamtmerkmalssumme haben die $g$ Merkmalsträger mit dem jeweils höchsten Einzelanteil an der Gesamtsumme?