Vorlesungsskript Statistik (DSBA)

Erich Neuwirth

21. Jänner 2021

Contents

Begleitende Literatur

- R for Data Science (https://r4ds.had.co.nz)
- Online Statistics Education: An Interactive Multimedia Course of Study (http://onlinestatbook.com)
- Statistical Thinking for the 21st Century (http://thinkstats.org)
- Fahrmeir et al.: Statistik - der Weg zur Datenanalyse
- Cheatsheets for R and RStudio (https://www.rstudio.com/resources/cheatsheets/)

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 R

`data.frame`

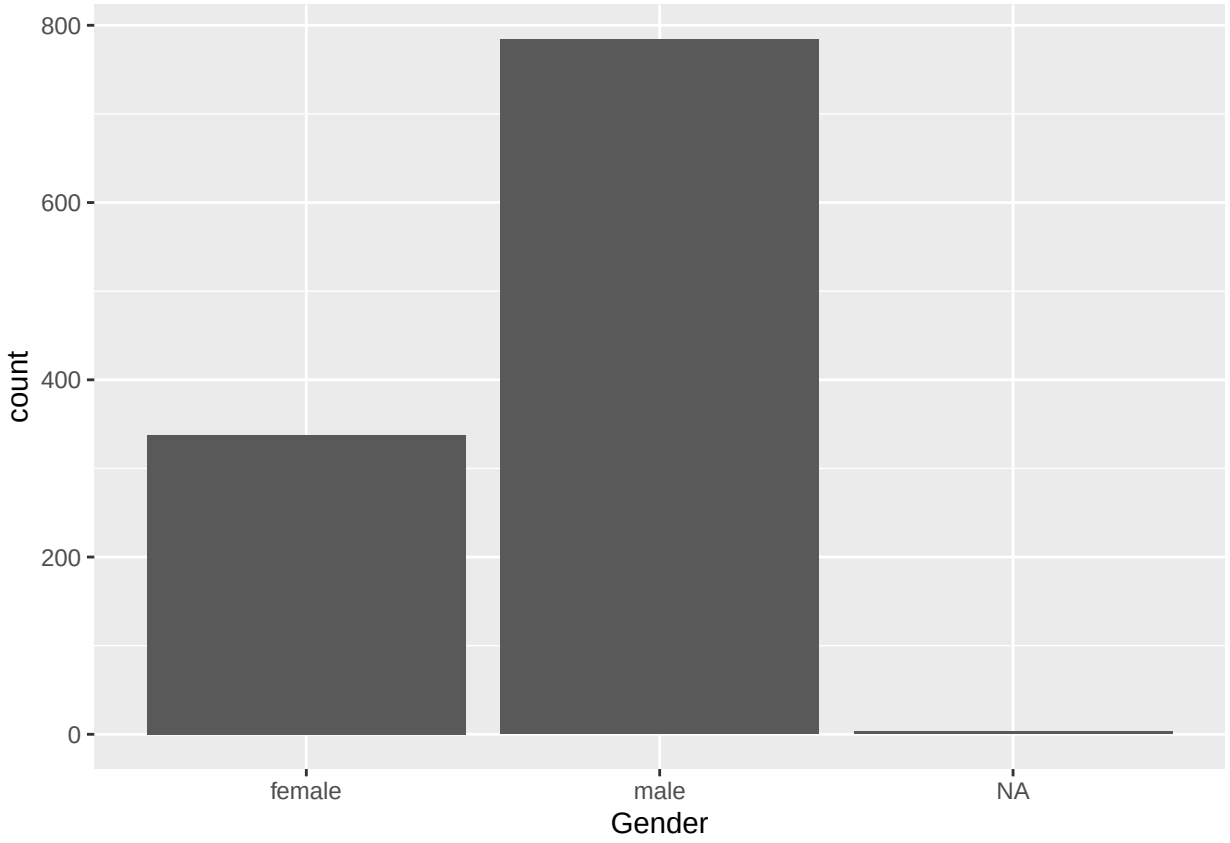
In `tidyverse` gibts

`tibble`

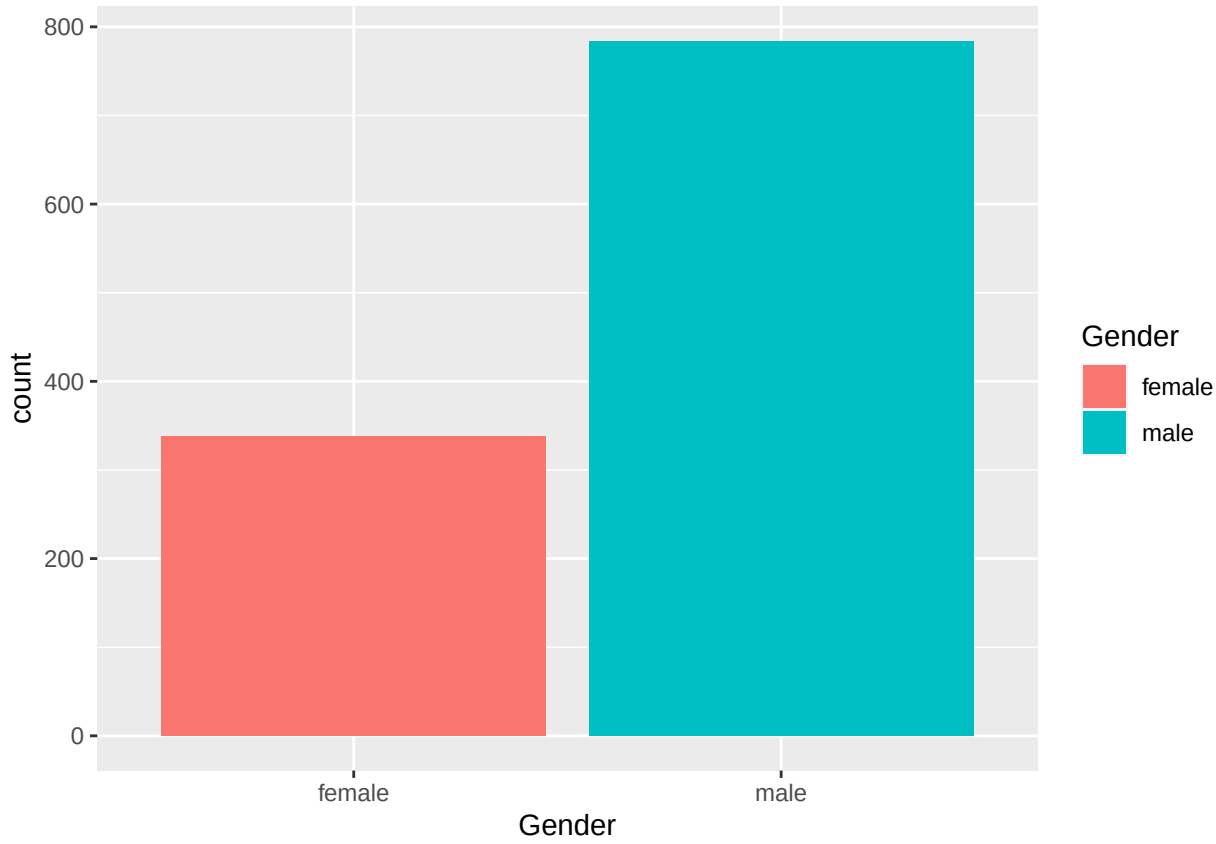
das ist ein „angereicherter“ `data.frame`

```
studdat <- read_excel(file.path("Daten",  "StudierendenDaten","StudDat.xlsx"))
```

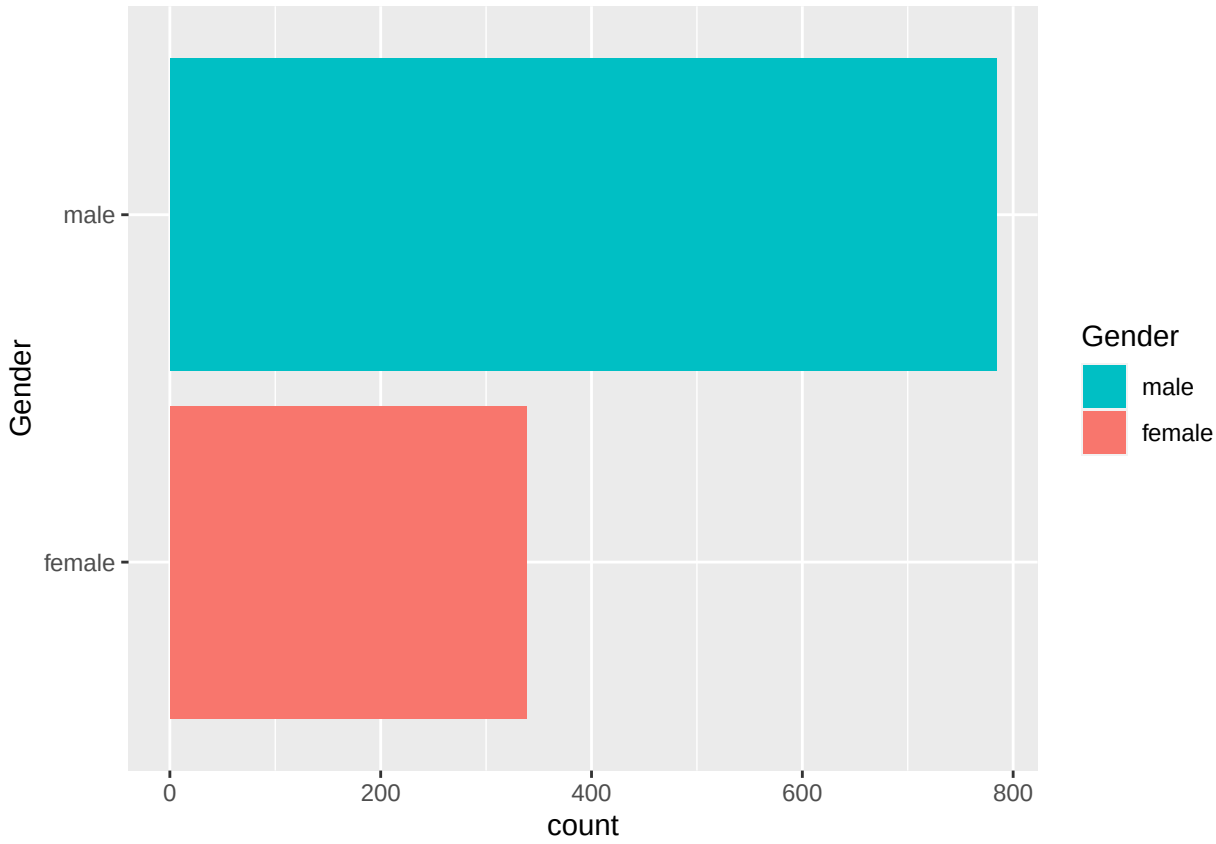
```
studdat %>%
  ggplot(aes(x=Gender)) +
  geom_bar()
```




```
studdat %>%
filter(!is.na(Gender)) %>%
  select(Gender) %>%
  drop_na() %>%
  ggplot(aes(x=Gender,fill=Gender)) +
  geom_bar()
```




```
studdat %>%
filter(!is.na(Gender)) %>%
  select(Gender) %>%
  drop_na() %>%
  ggplot(aes(x=Gender,fill=Gender)) +
  geom_bar() +
  coord_flip() +
    guides(fill = guide_legend(reverse = TRUE))
```

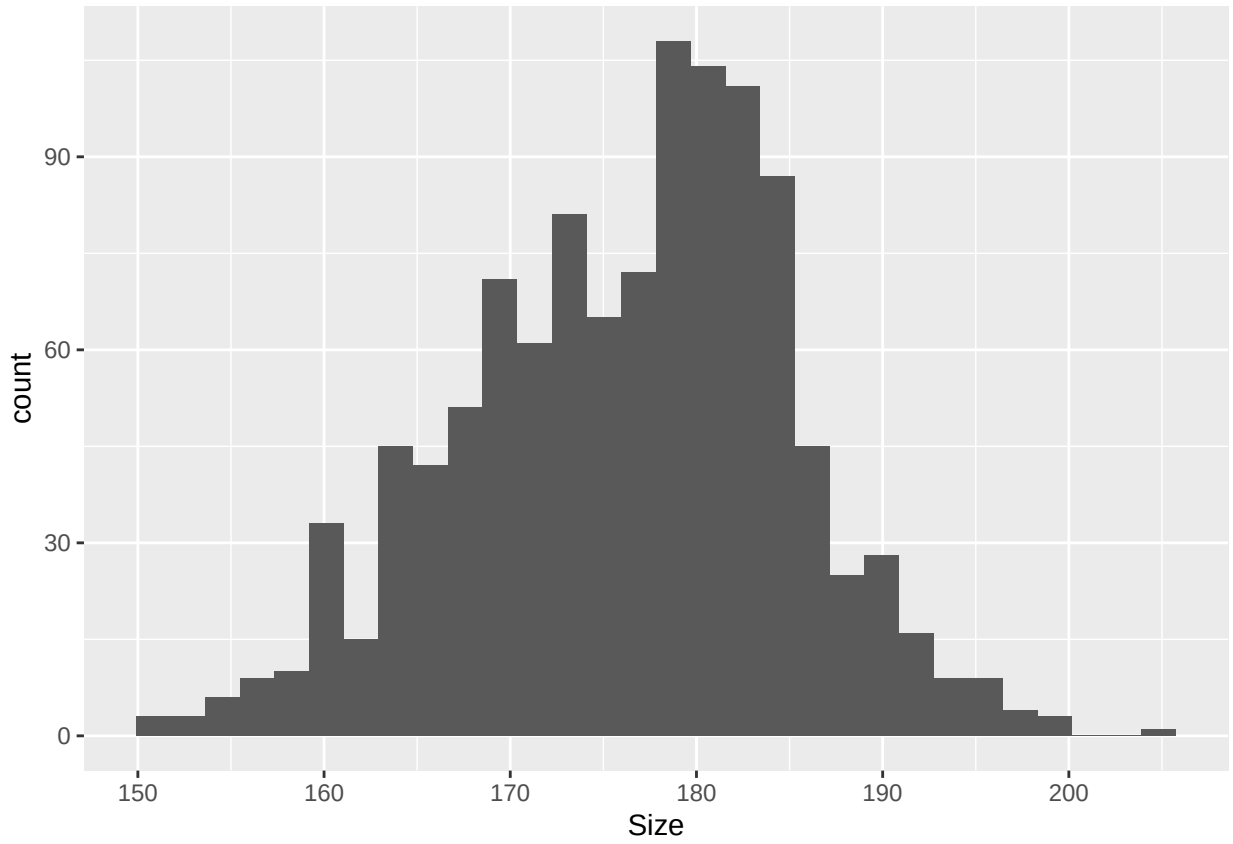


Man unterscheidet bei Intervall- und Verhältnisskalen und zusätzlich noch diskrete und stetige (kontinuierliche) Skalen.

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_histogram()
```

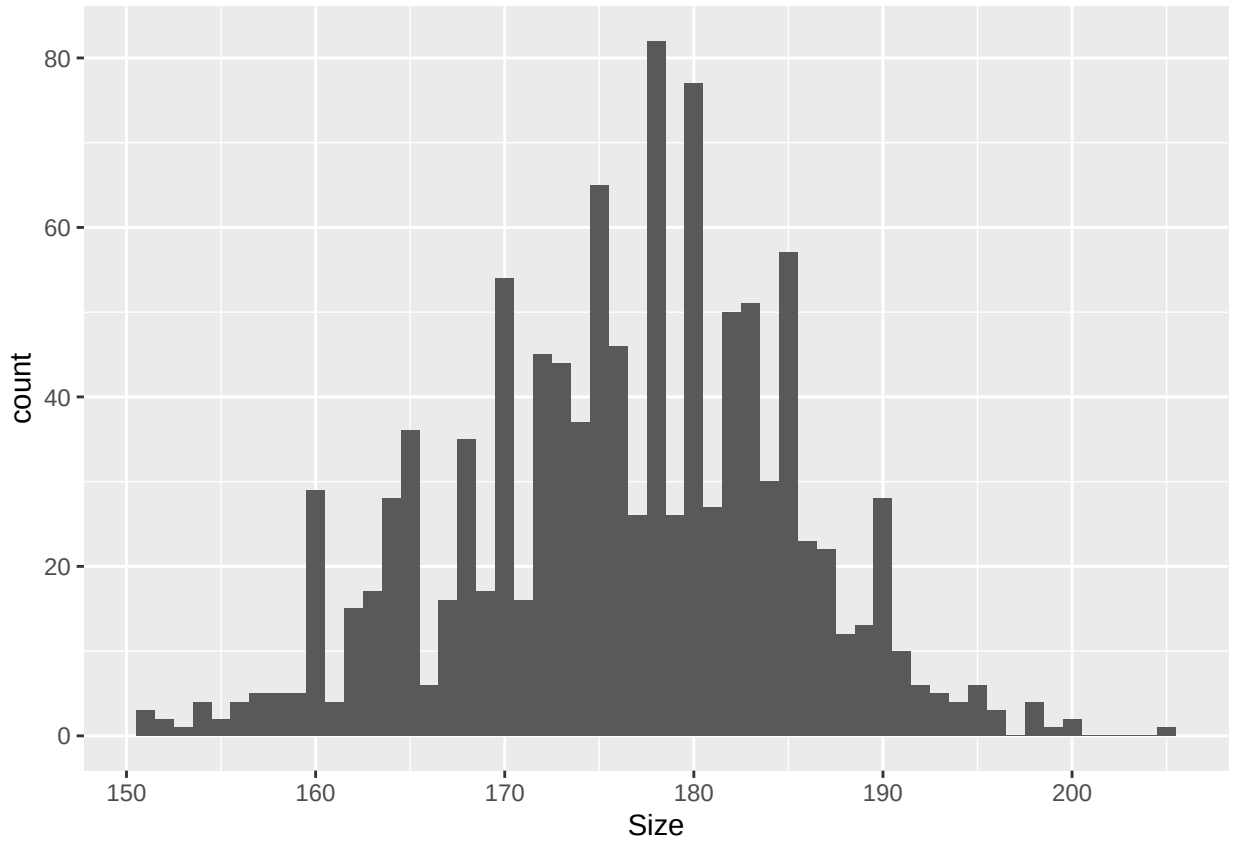
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 18 rows containing non-finite values (stat_bin).
```



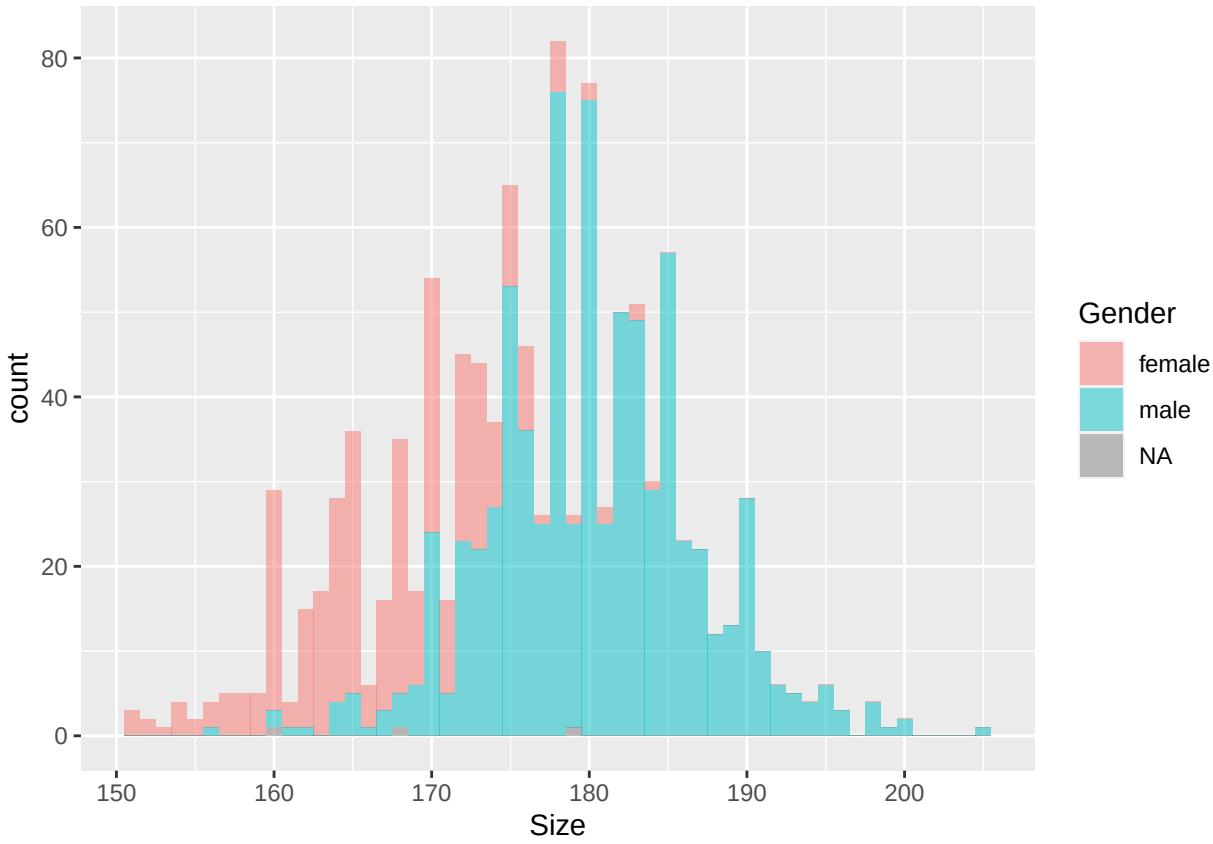

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_histogram(center=0,binwidth=1)
```

```
## Warning: Removed 18 rows containing non-finite values (stat_bin).
```




```
studdat %>%
  ggplot(aes(x=Size,fill=Gender)) +
  geom_histogram(center=0,binwidth=1,alpha=0.5)
```

```
## Warning: Removed 18 rows containing non-finite values (stat_bin).
```



Überlappungen nicht besonders sauber dargestellt.

Besser Kerndichteschätzer

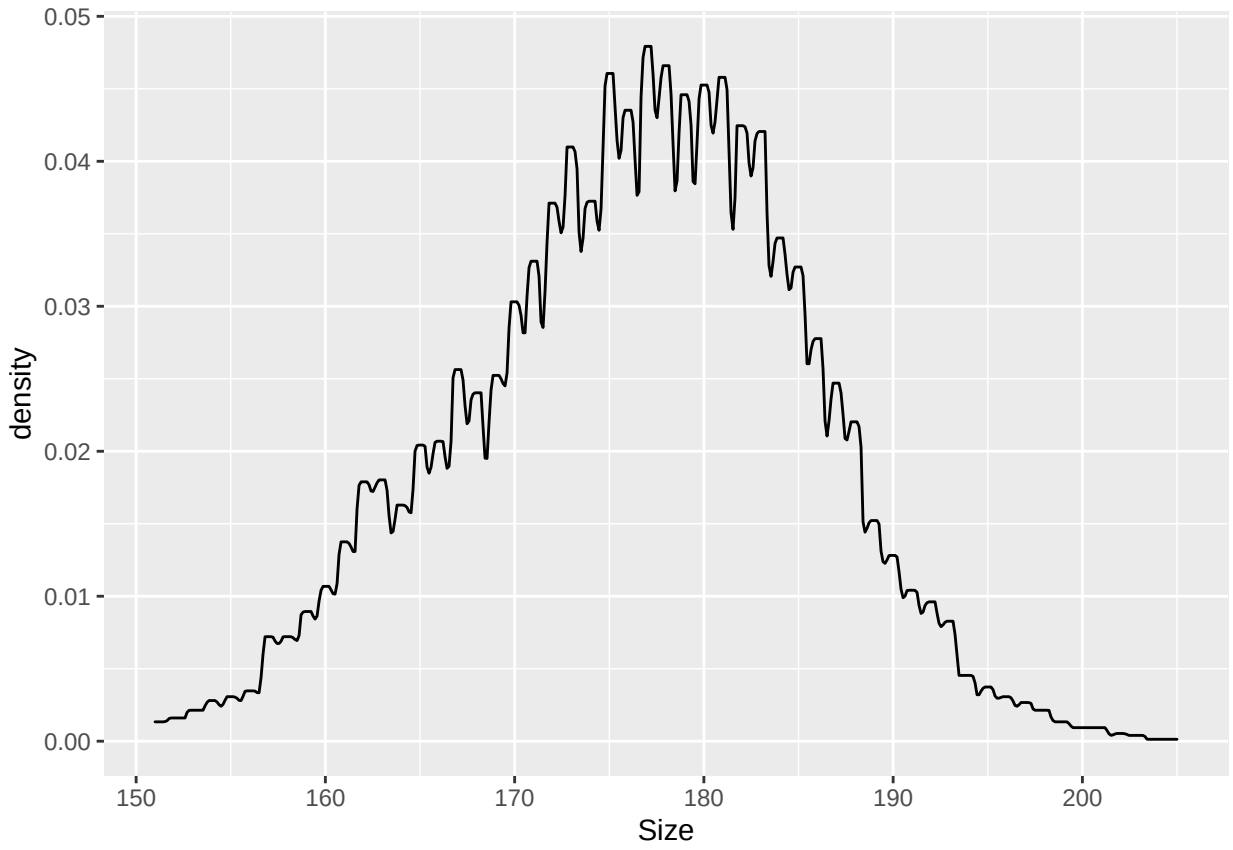
$$f(x|x_1 \dots x_n) = \frac{1}{nw} \sum_{i=1}^{n} k\left(\frac{x - x_i}{w}\right)$$

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

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_density(kernel="rectangular")
```

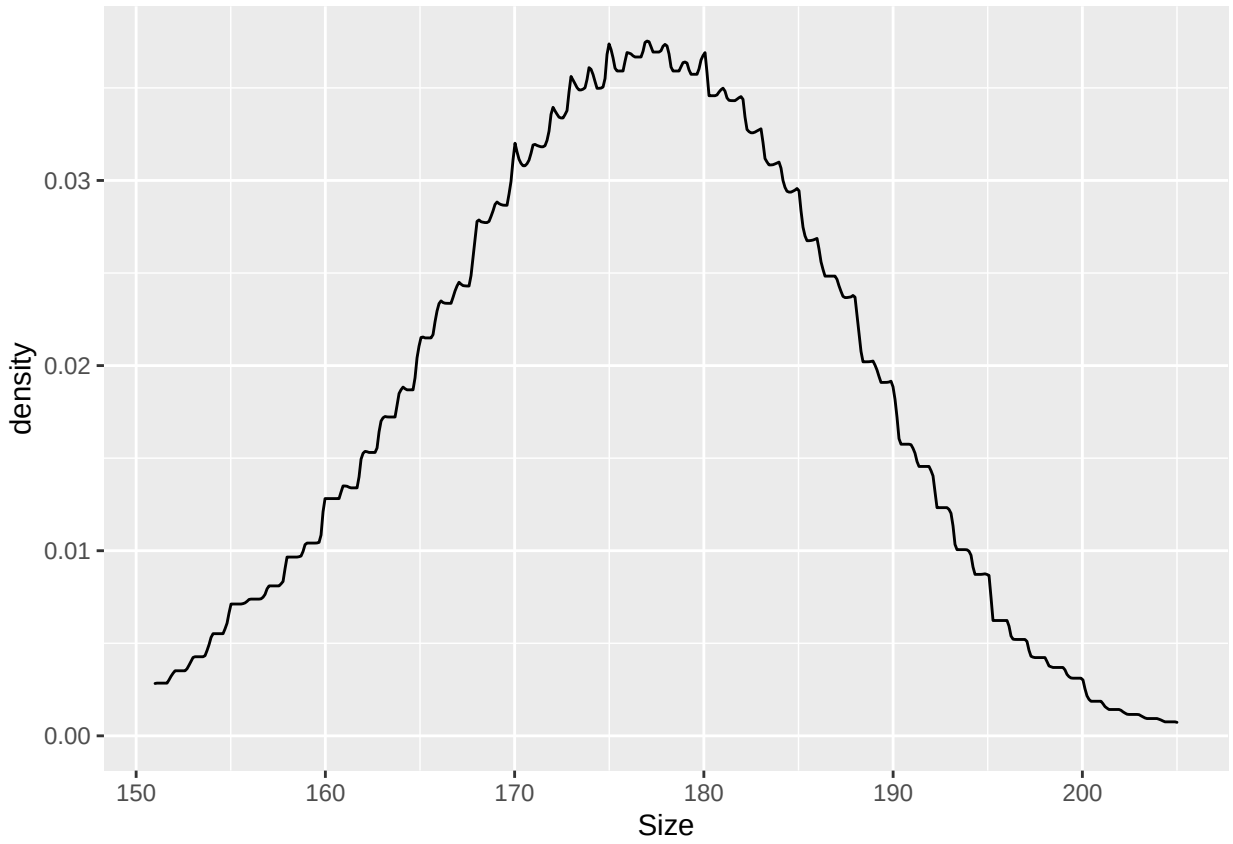
```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Fensterbreite verdreifacht

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_density(kernel="rectangular",adjust=3)
```

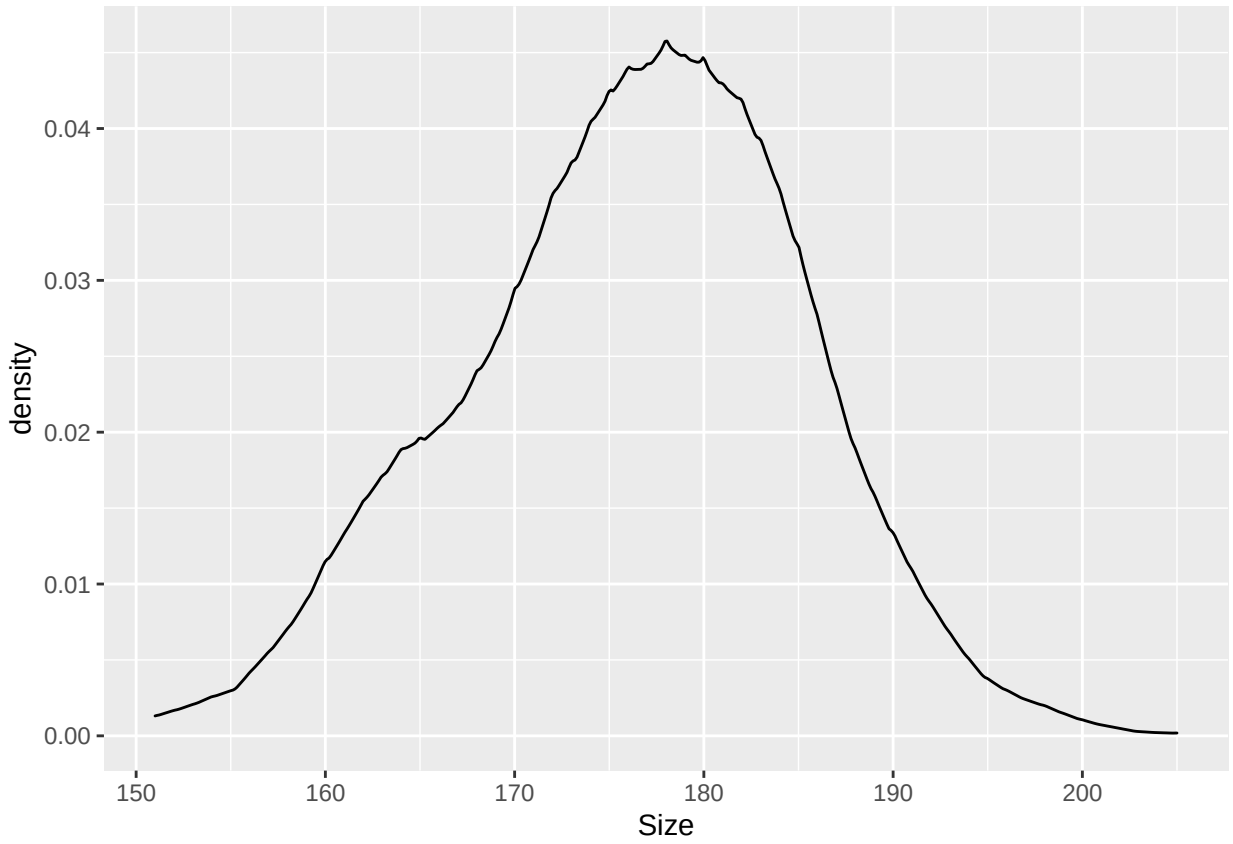
```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Dreieckskern

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_density(kernel="triangular")
```

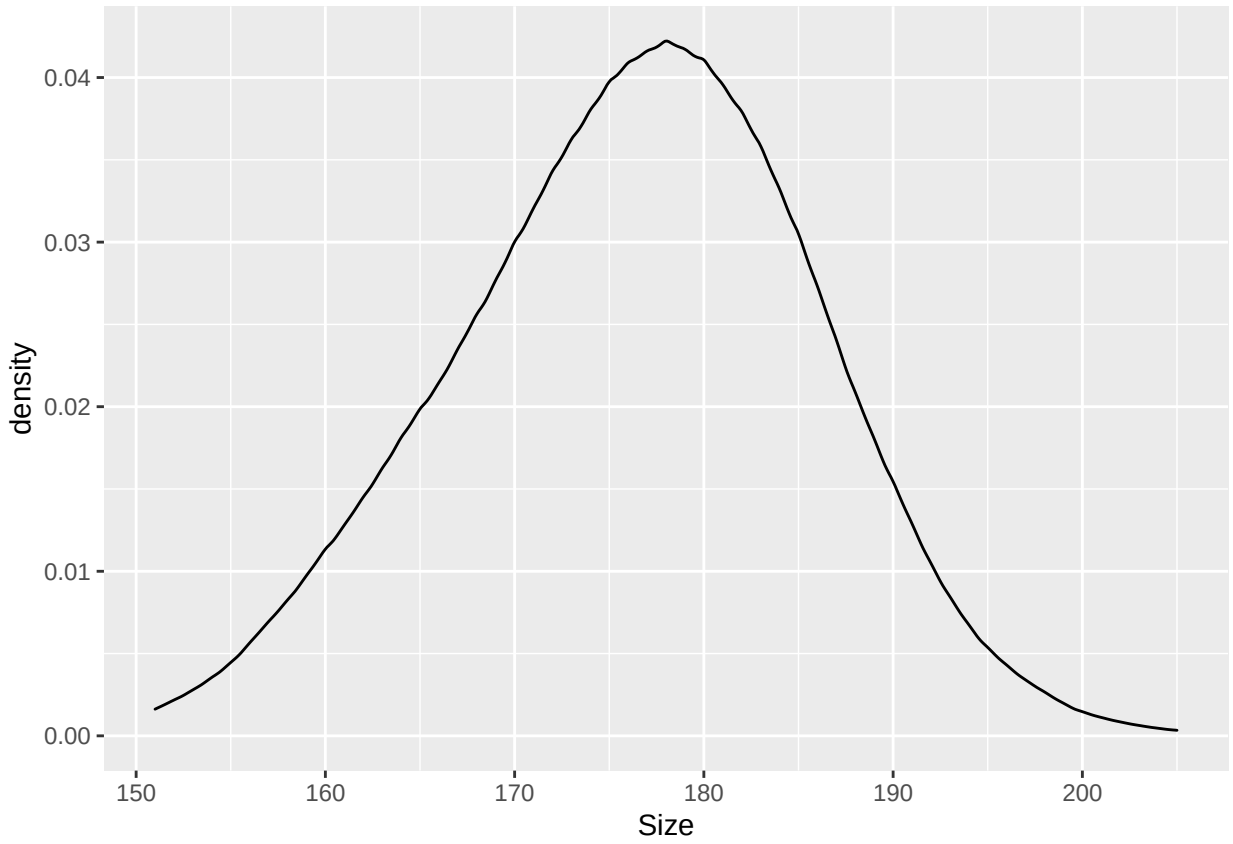
```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Dreieckskern

```
studdat %>%
  ggplot(aes(x=Size)) +
  geom_density(kernel="triangular",adjust=2)
```

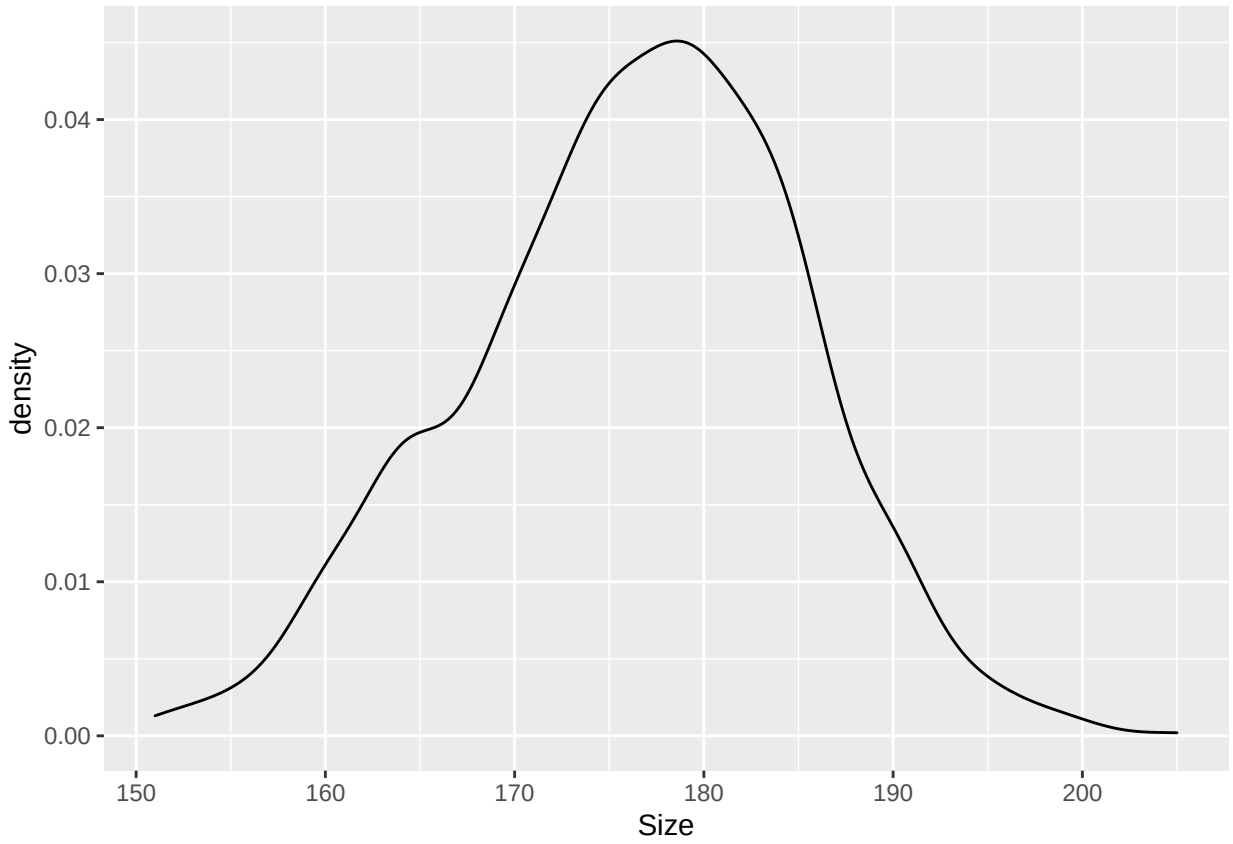
```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Kern Normalverteilungsdichte

```
studdat %>%
  filter(!is.na(Gender)) %>%
  ggplot(aes(x=Size)) +
  geom_density()
```

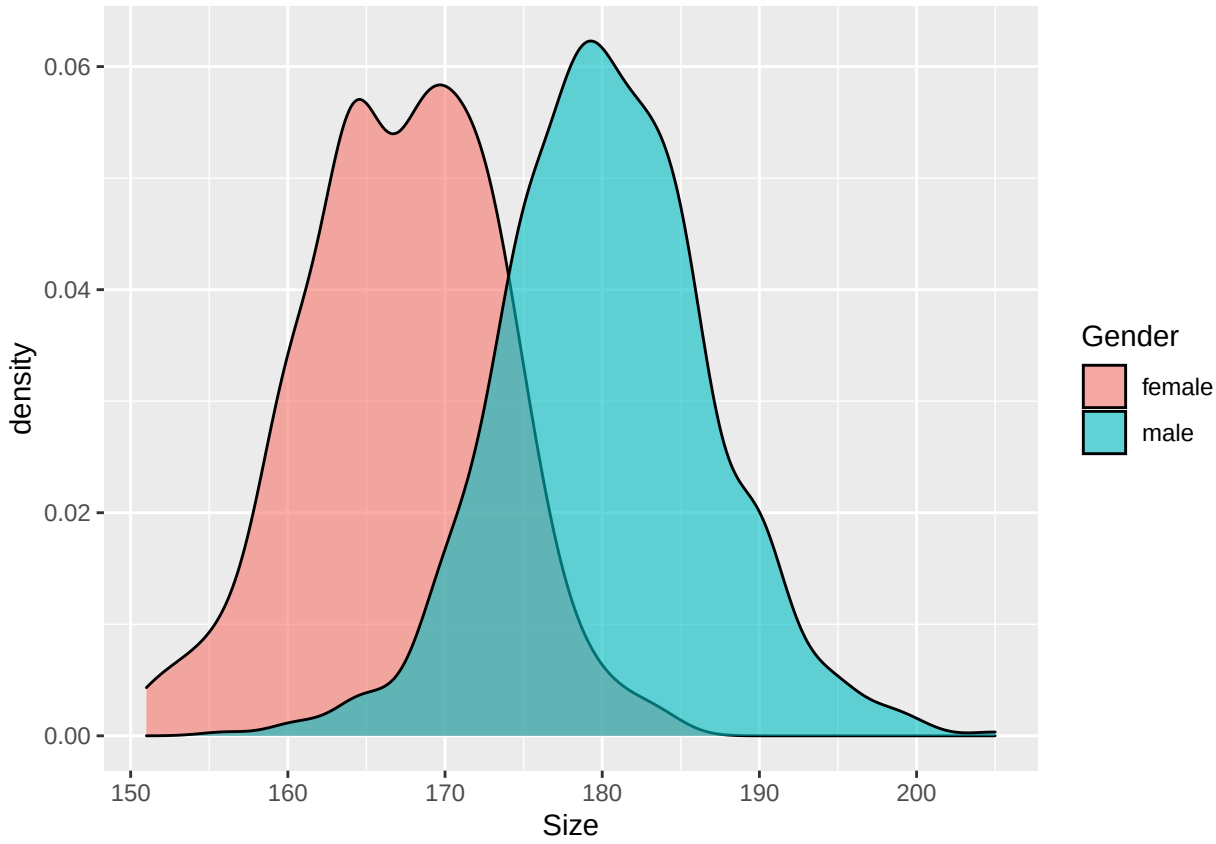
```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



Zweigipfeligkeit oft Hinweis auf 2 zusammengemischte Grundgesamtheiten

```
studdat %>%
  filter(!is.na(Gender)) %>%
  ggplot(aes(x=Size,fill=Gender)) +
  geom_density(alpha=0.6)
```

```
## Warning: Removed 18 rows containing non-finite values (stat_density).
```



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

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

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

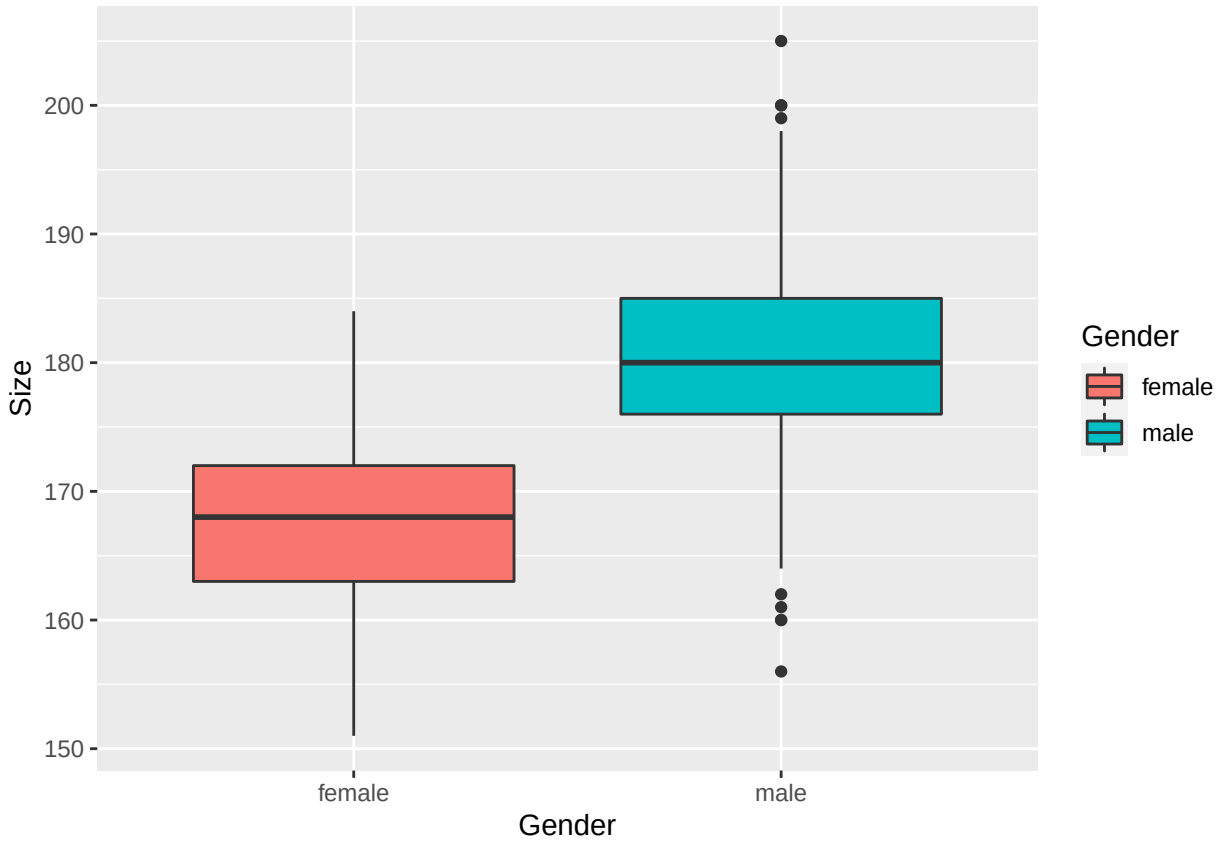
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}$$

Schiefemaß mit Quartilen

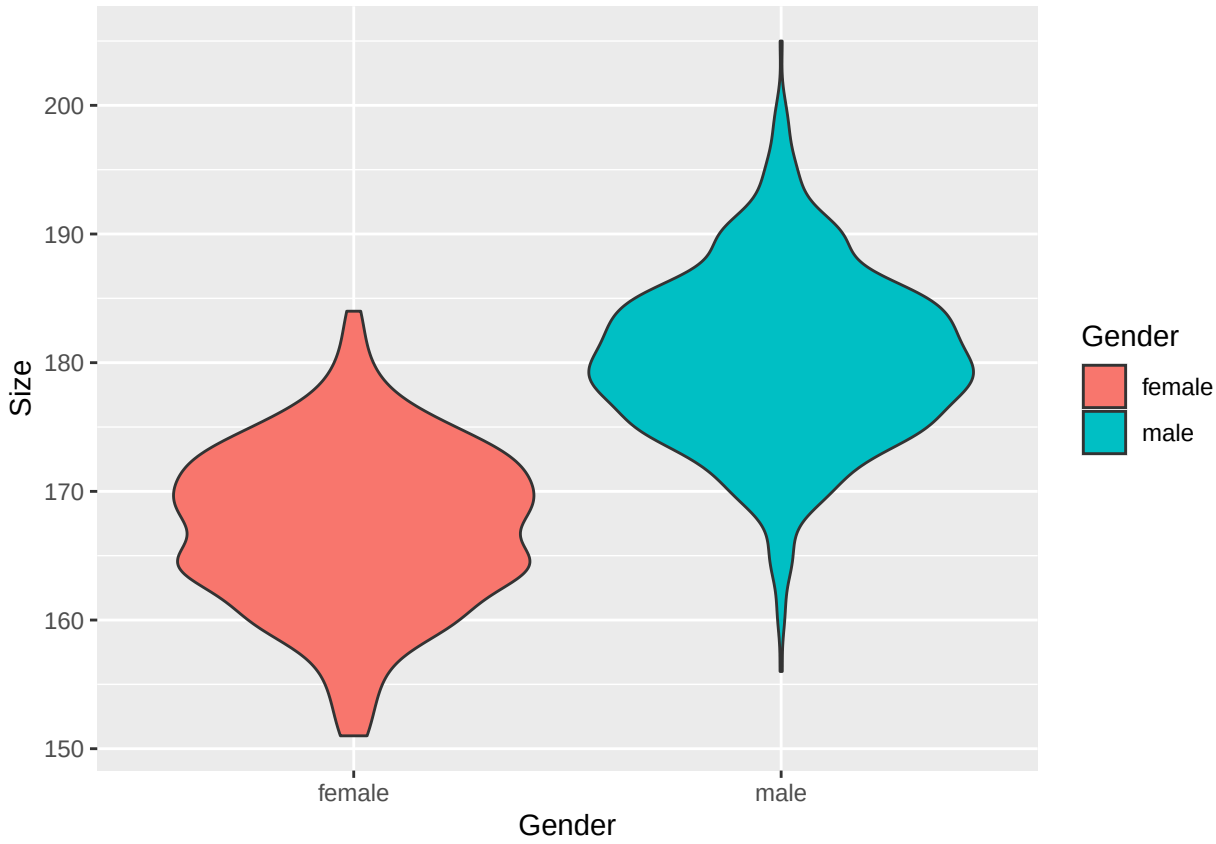
$$\frac{|Q3 - Q2| - |Q2 - Q1|}{Q3 - Q1}$$

Boxplots

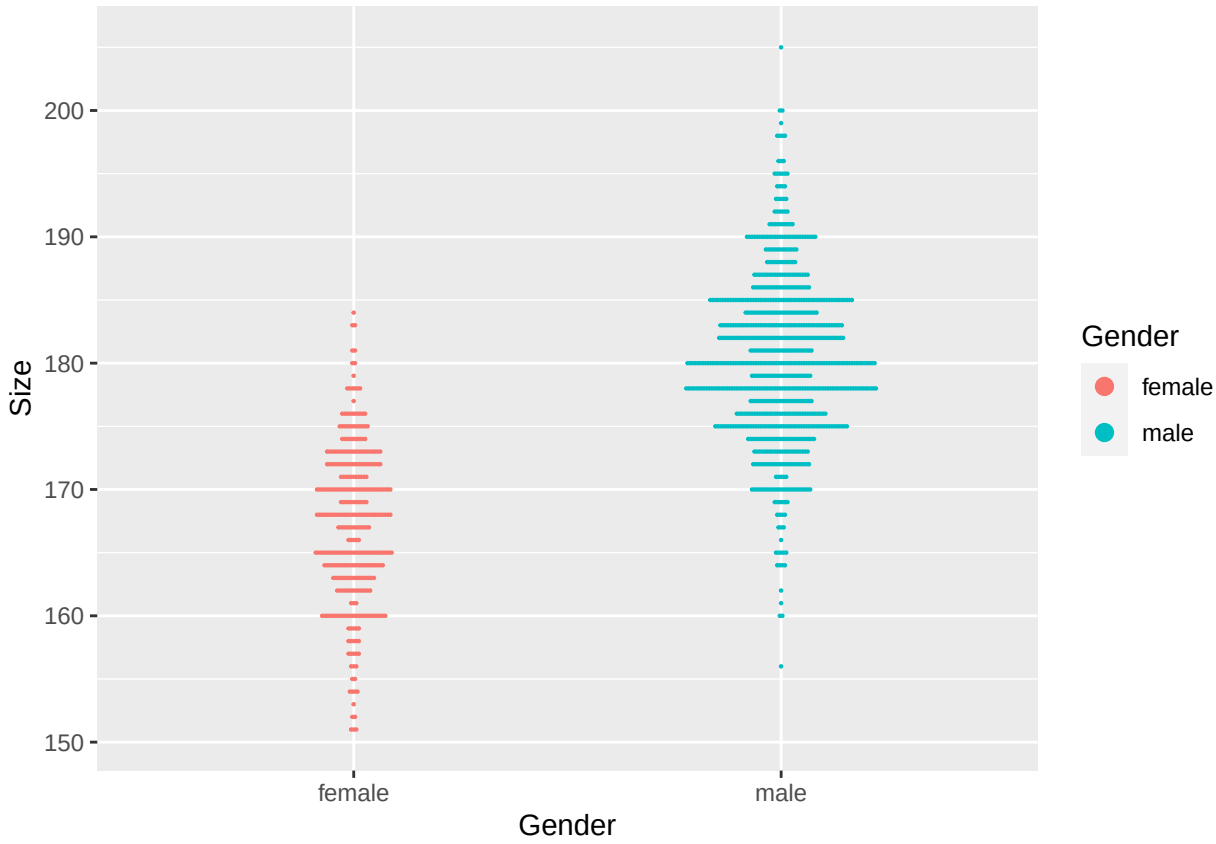
```
studdat %>%
  select(Gender,Size) %>%
  drop_na() %>%
  ggplot(aes(y=Size,x=Gender,fill=Gender)) +
  geom_boxplot()
```



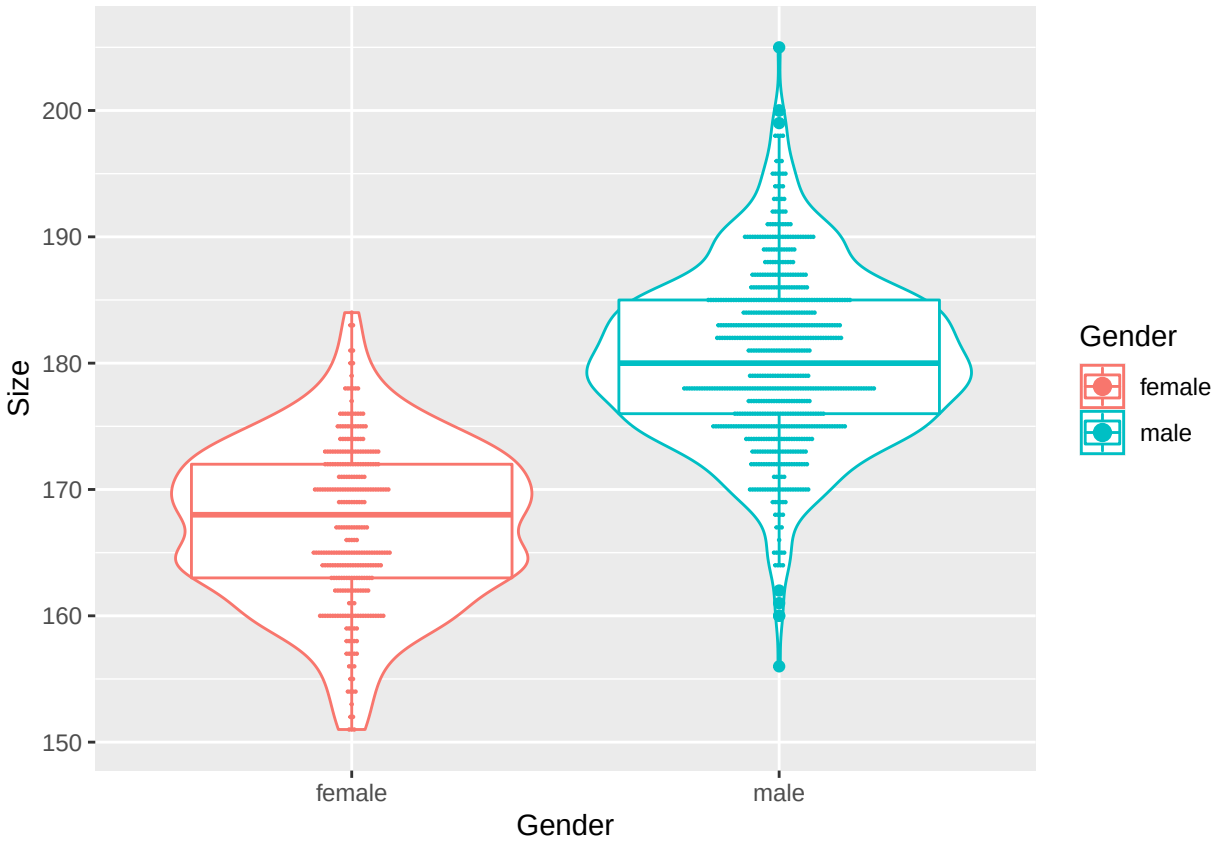

```
studdat %>%
  select(Gender,Size) %>%
  drop_na() %>%
  ggplot(aes(y=Size,x=Gender,fill=Gender)) +
  geom_violin()
```




```
studdat %>%
  select(Gender,Size) %>%
  drop_na() %>%
  ggplot(aes(y=Size,x=Gender,color=Gender)) +
  geom_dotplot(aes(fill=Gender),binaxis="y",stackdir="center",dotsize=0.2,binwidth=1)
```




```
studdat %>%
  select(Gender,Size) %>%
  drop_na() %>%
  ggplot(aes(y=Size,x=Gender,color=Gender)) +
  geom_violin() +
  geom_boxplot() +
  geom_dotplot(aes(fill=Gender),binaxis="y",stackdir="center",dotsize=0.2,binwidth=1)
```



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?

Elementare Wahrscheinlichkeitsrechnung

Binomialkoeffizienten

Aufsteigend geordnete (Zahlen-)Folgen der Länge k wobei alle Zahlen $\leq n$ sein müssen.

$$\begin{aligned} B(n,1) &= n && \text{für } n \geq 1 \\ B(1,k) &= 0 && \text{für } k \geq 2 \\ B(n,k) &= B(n-1,k-1) + B(n-1,k) && \text{für } n \geq 2 \text{ und } k \geq 2 \end{aligned}$$

$B(n,k)$ wird auch als $\binom{n}{k}$ geschrieben und es gilt $B(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Braucht man zur Berechnung von Lottowahrscheinlichkeiten

n-stellige k-ziffrige Zahlen

$$\begin{aligned} Z(n,1) &= 10 && \text{für } n \geq 1 \\ Z(1,k) &= 0 && \text{für } k \geq 2 \\ Z(n,k) &= (10-(k-1))Z(n-1,k-1) + kZ(n-1,k) && \text{für } n \geq 2 \text{ und } k \geq 2 \end{aligned}$$

Wahrscheinlichkeiten und Zufallsvariable

Diskrete Wahrscheinlichkeitsverteilung

Ereignisraum

$$\{x_1, x_2, \ldots\}$$

x_i Elementarereignisse, können endlich oder unendlich viele sein.

p_i Wahrscheinlichkeit des Ergebnisses x_i.

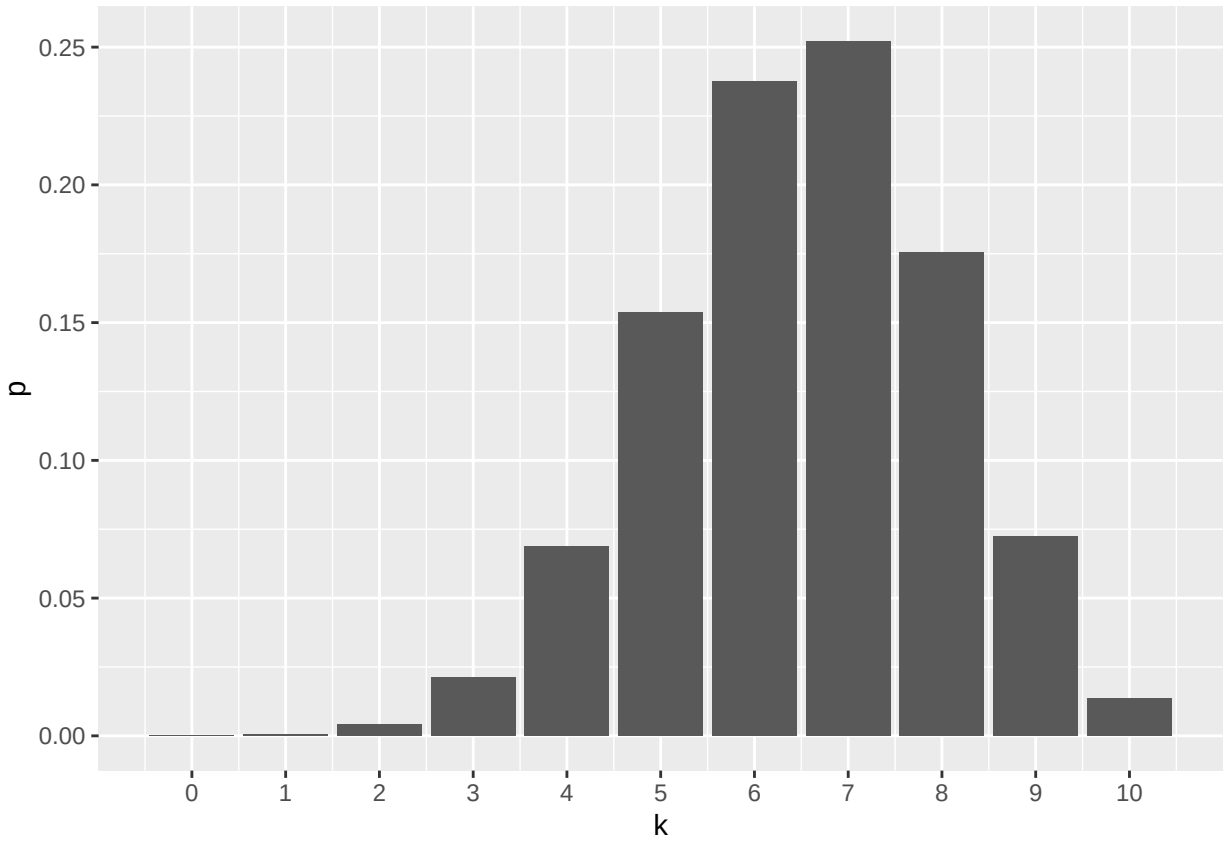
Es muss gelten $p_i \geq 0$ und $\sum_i p_i = 1$

Ereignisse sind Mengen von Elementarereignissen. Die Wahrscheinlichkeit eines Ereignisses ist die Summe der Wahrscheinlichkeiten seiner Elementarereignisse.

Wichtige diskrete Verteilungen

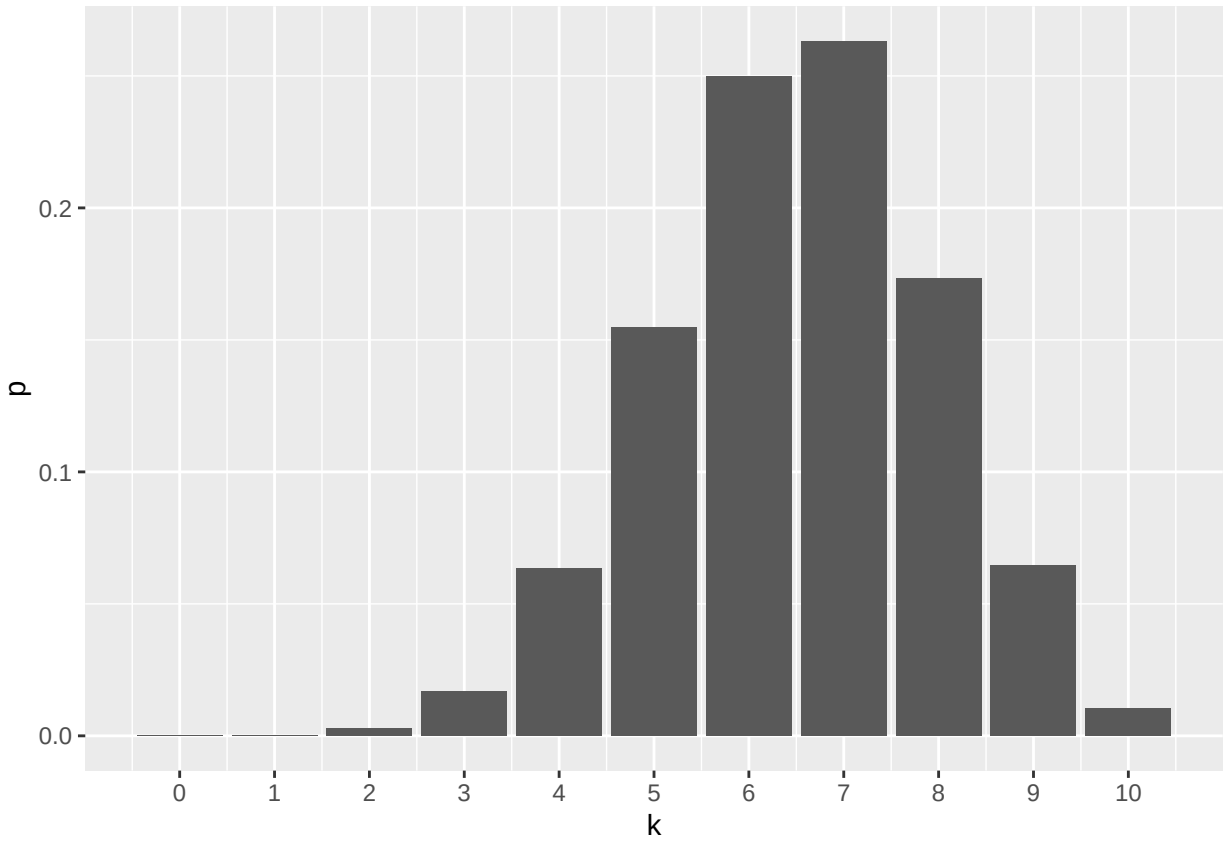
Binomialverteilung
$B(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$
Modell für Ziehen mit Zurücklegen

```
tibble(k=0:10,p=dbinom(k,10,0.65)) -> binom_df
binom_df %>%
  ggplot(aes(x=k,y=p)) +
  geom_col() +
  scale_x_continuous(breaks=0:10)
```



Hypergeometrische Verteilung
$H(k|n,M,N) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ Modell für Ziehen ohne Zurücklegen.

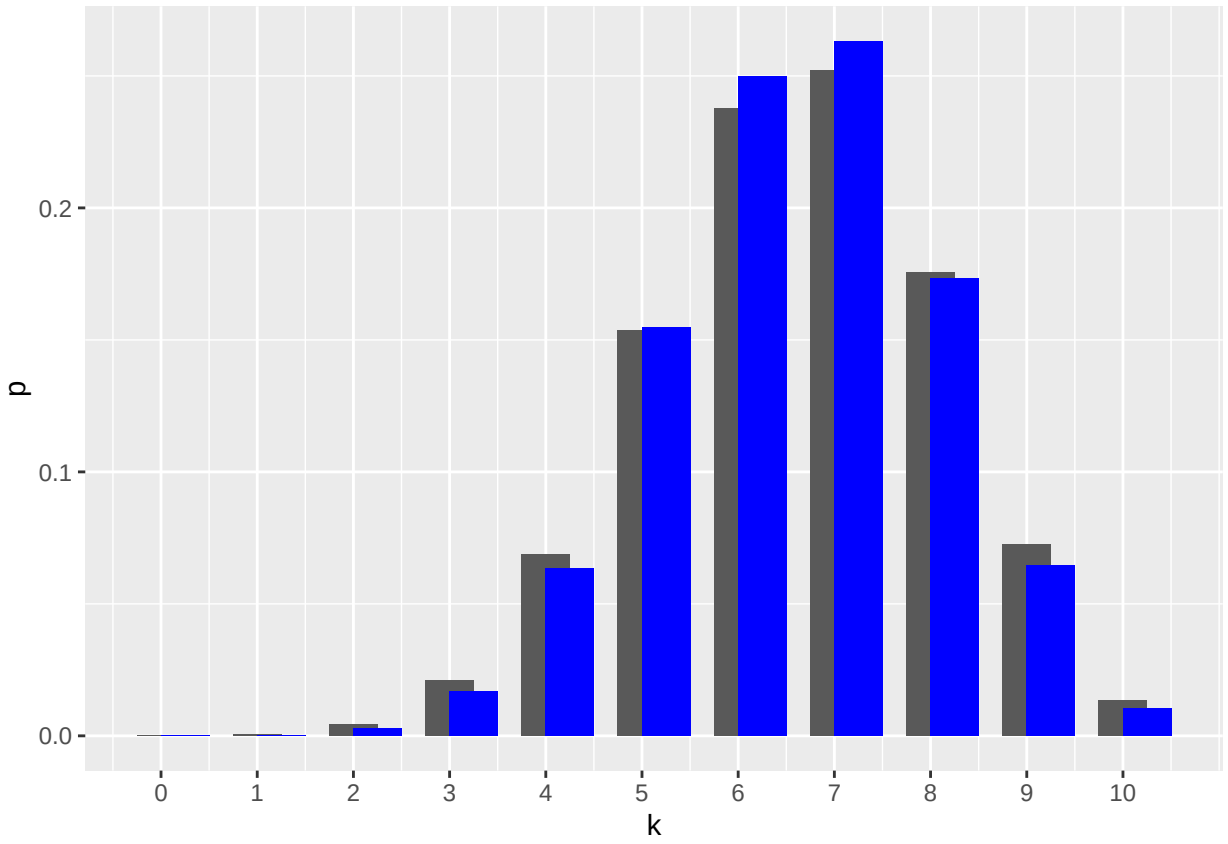
```
tibble(k=0:10,p=dhyper(k,65,35,10)) -> hyper_df
hyper_df %>%
  ggplot(aes(x=k,y=p)) +
  geom_col() +
  scale_x_continuous(breaks=0:10)
```



Wenn n viel kleiner als N ist $B(k|n, \frac{M}{n})$ eine gute Näherung und einfach zu berechnen.

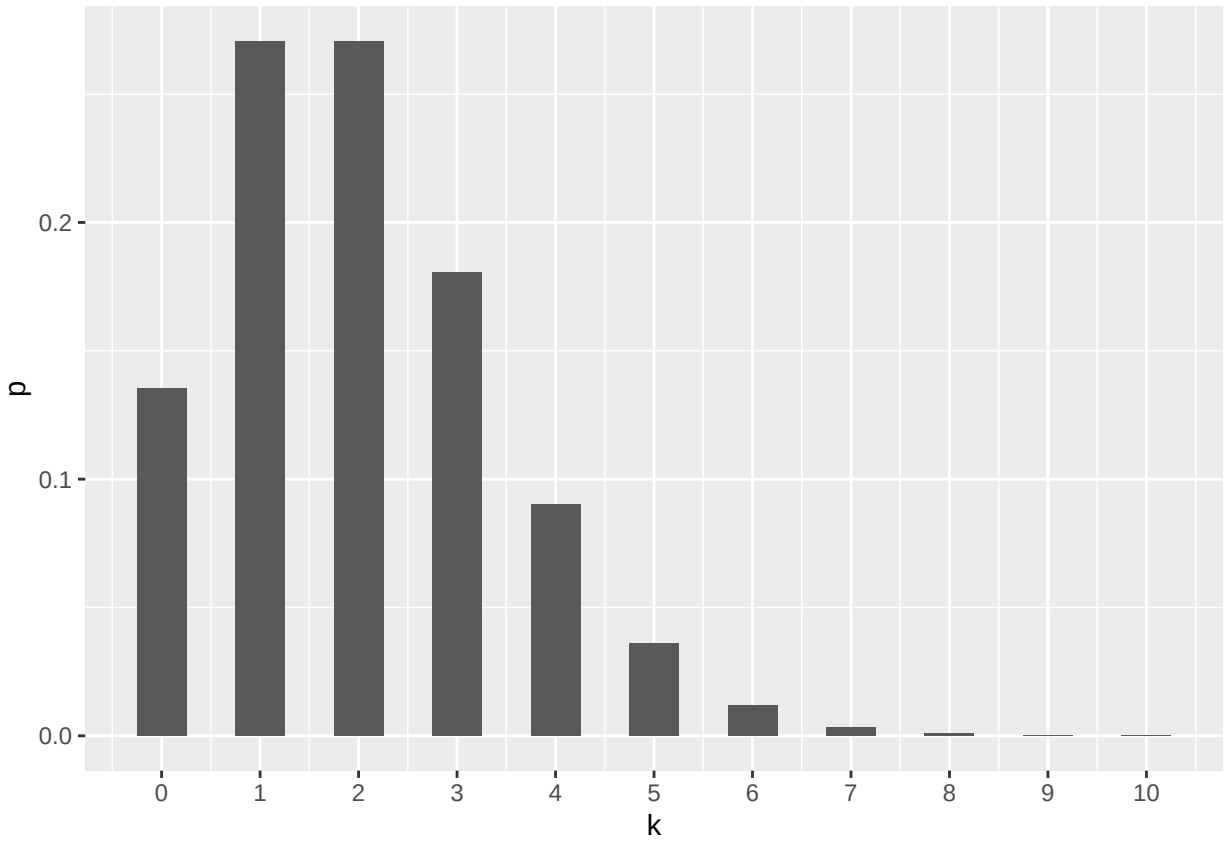
Wenn n viel kleiner als N ist, dann ist $B(k|n, \frac{M}{n})$ eine gute Näherung und einfach zu berechnen.

```
binom_df %>%
  ggplot(aes(x=k,y=p)) +
  geom_col(width=0.5) +
  geom_col(aes(x=k+0.25),data=hyper_df,width=0.5,fill="blue") +
  scale_x_continuous(breaks=0:10)
```



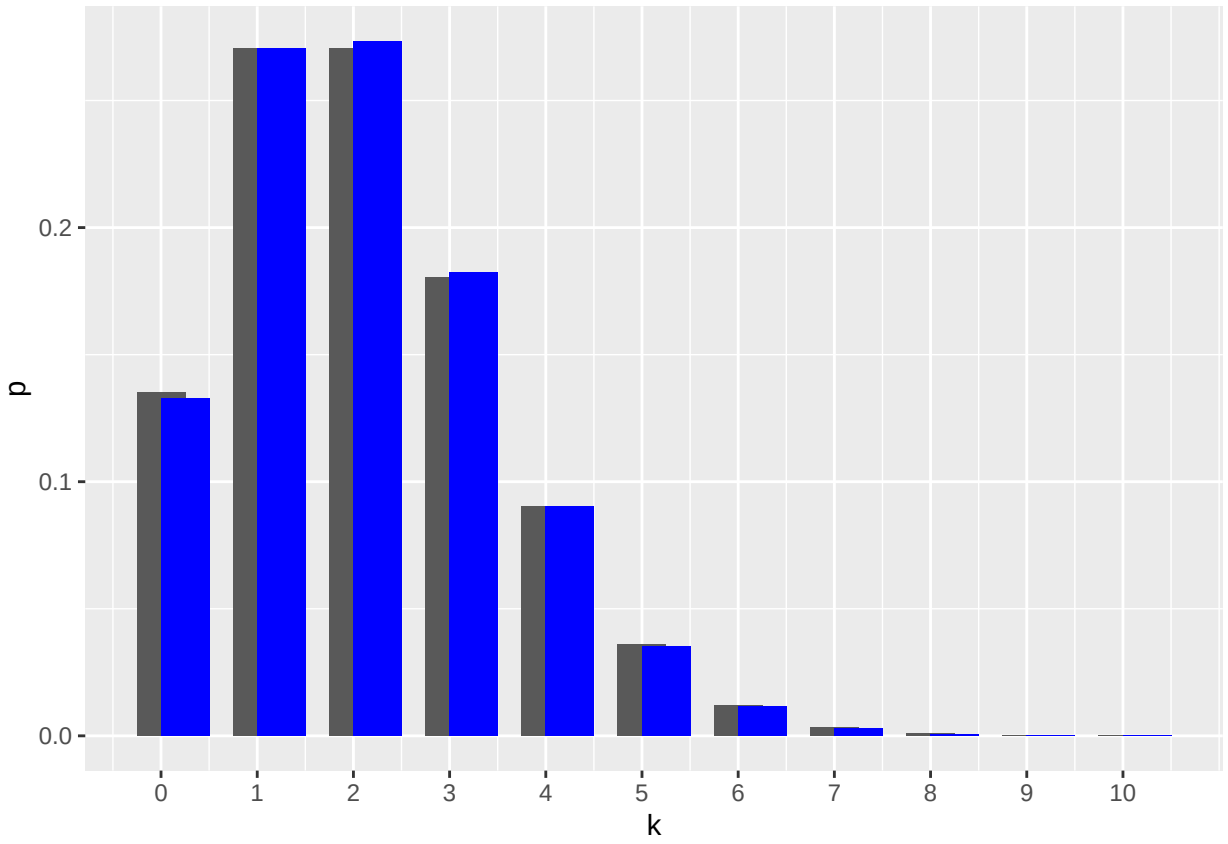
Poisson-Verteilung
$P(k|\lambda) = e^{-\lambda}\frac{x^k}{k!}$

```
tibble(k=0:10,p=dpois(k,2)) ->
  poisson_df
poisson_df %>%
  ggplot(aes(x=k,y=p)) +
  geom_col(width=0.5) +
  scale_x_continuous(breaks=0:10)
```



Gute Näherung für Binomialverteilung mit großem n und kleinem p. Dann ist die Poissonverteilung $P(k|np)$ (mit $\lambda = np$) eine gute Näherung für die Binomialverteilung $B(k|n,p)$

```
tibble(k=0:10,p=dbinom(k,100,0.02)) -> binom_df
poisson_df %>%
  ggplot(aes(x=k,y=p)) +
  geom_col(width=0.5) +
  geom_col(aes(x=k+0.25),data=binom_df,width=0.5,fill="blue") +
  scale_x_continuous(breaks=0:10)
```



Stetige Wahrscheinlichkeitsverteilungen Ereignisraum sind alle reellen Zahlen. Wahrscheinlichkeiten gibt es für Intervalle.

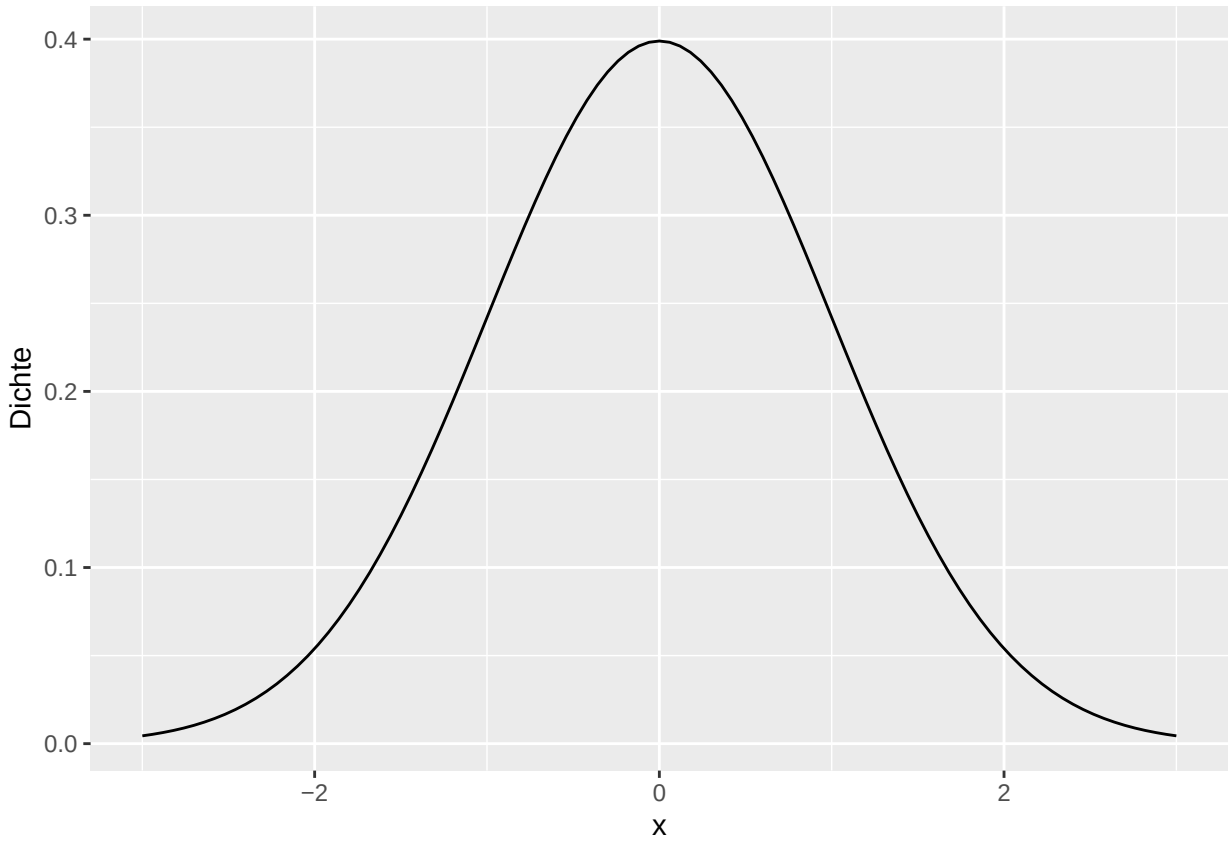
Wahrscheinlichkeiten angegeben durch Dichtefunktion $f(x)$ mit $f(x) \geq 0$ für alle x und $\int_{-\infty}^{\infty} f(x)dx = 1$

Die Wahrscheinlichkeit, dass der Wert x im Intervall $[a, b]$ liegt ist dann
$P(a \leq x \leq b) = \int_a^b f(x)dx$

Wenn F die Verteilungsfunktion zur Dichte f ist, dann ist $F(x) = \int_{-infty}^{x} f(z)dz$ und
$P(a \leq x \leq b) = F(b) - F(a)$

Normalverteilung
$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

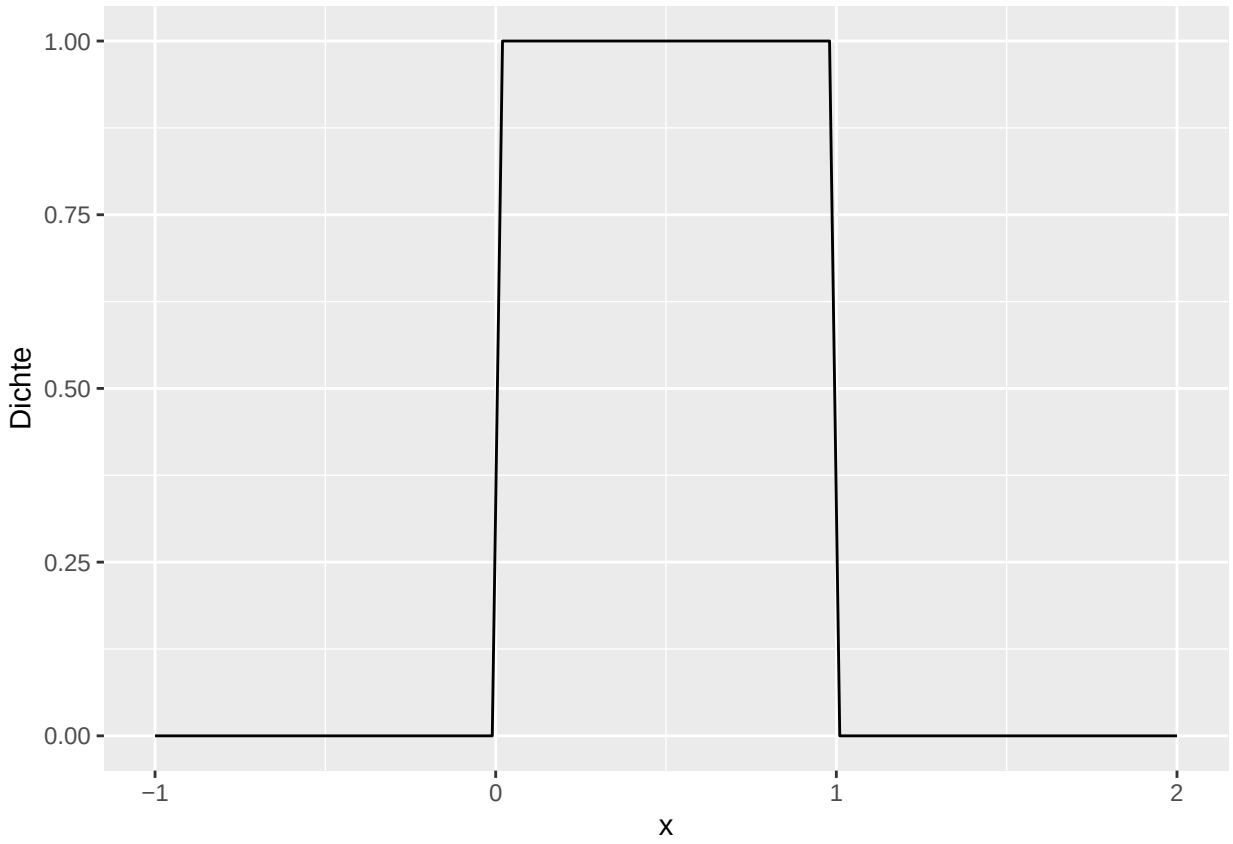
```
tibble(x=c(-3,3),y=c(0,dnorm(0))) %>%
  ggplot(aes(x,y)) +
  stat_function(fun=dnorm) +
  scale_y_continuous("Dichte")
```



Gleichverteilung auf dem Intervall $[a, b]$

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

```
tibble(x=c(-1,2),y=c(0,1)) %>%
  ggplot(aes(x,y)) +
  stat_function(fun=dunif) +
  scale_y_continuous("Dichte")
```



Zufallsvariablen

Zufallsvariablen haben eine Verteilung und können addiert werden. Z.B. Augensumme zweier Würfel. Es gibt diskrete und stetige Zufallsvariable.

Zufallsvariablen haben einen Erwartungswert und eine Varianz. Bei diskreten Zufallsvariablen X ist
$E(X) = \sum_i x_i p_i$
$V(X) = \sum_i (x_i - E(X))^2 p_i$

Bei diskreten Zufallsvariablen X ist
$E(X) = \int_{-\infty}^{\infty} x f(x) dx$
$E(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$

Für 2 Zufallsvariable X_1 und X_2gilt immer $E(X_1 + X_2) = E(X_1) + E(X_2)$
Für 2 *unabhängige* Zufallsvariable X_1 und X_2gilt immer $V(X_1 + X_2) = V(X_1) + V(X_2)$

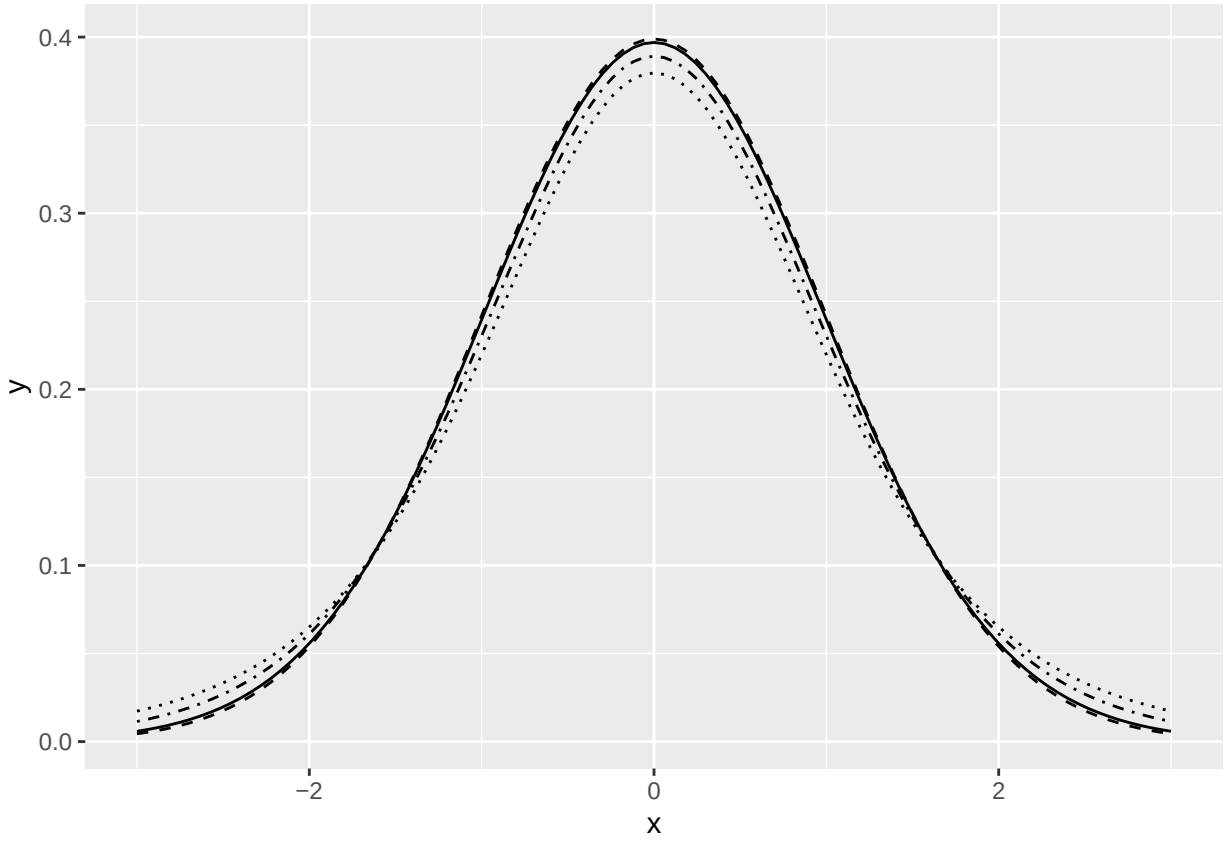
Für Zufallsvariablen und Konstanten a und b gilt $E(aX + b) = aE(X) + b$ und $V(aX + b) = a^2 V(X)$

Aus normalverteilten Zufallsvariablen werden weitere Verteilungen abgeleitet.

Wenn $X_0, X_1, \dots X_n$ normalverteilt mit $\mu = 0$ und $\sigma = 1$ sind, dann heißt die Verteilung von $T = \frac{X_0}{\sum_{i=1}^{n} X_0}$ t-Verteilung mit n Freiheitsgraden. Die Dichtefunktion ist
$f(x|n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + x^2/n)^{-\frac{n+1}{2}}$
Dabei ist $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$

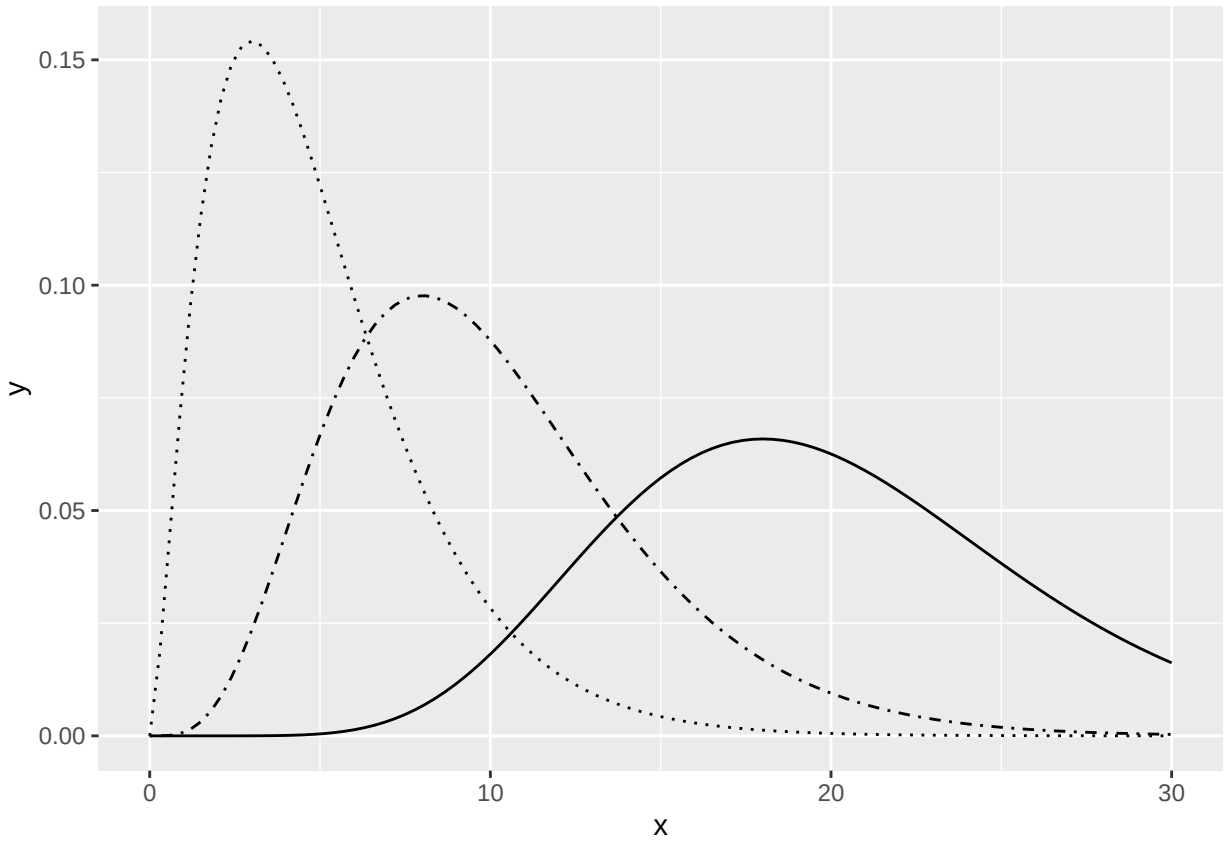
```
tibble(x=c(-3,3),y=c(0,dnorm(0))) %>%
  ggplot(aes(x,y)) +
  stat_function(fun=function(x)dnorm(x),linetype="dashed") +
  stat_function(fun=function(x)dt(x,5),linetype="dotted") +
```

```
  stat_function(fun=function(x)dt(x,10),linetype="dotdash") +
  stat_function(fun=function(x)dt(x,50))
```



Wenn $X_1, \ldots X_n$ normalverteilt mit $\mu = 0$ und $\sigma = 1$ sind, dann heißt die Verteilung von $\chi = \sum_{i=1}^{n} X_i^2$ χ^2-Verteilung mit n Freiheitsgraden. Die Dichtefunktion ist
$f(x|n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + x^2/n)^{-\frac{n+1}{2}}$
Dabei ist $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$

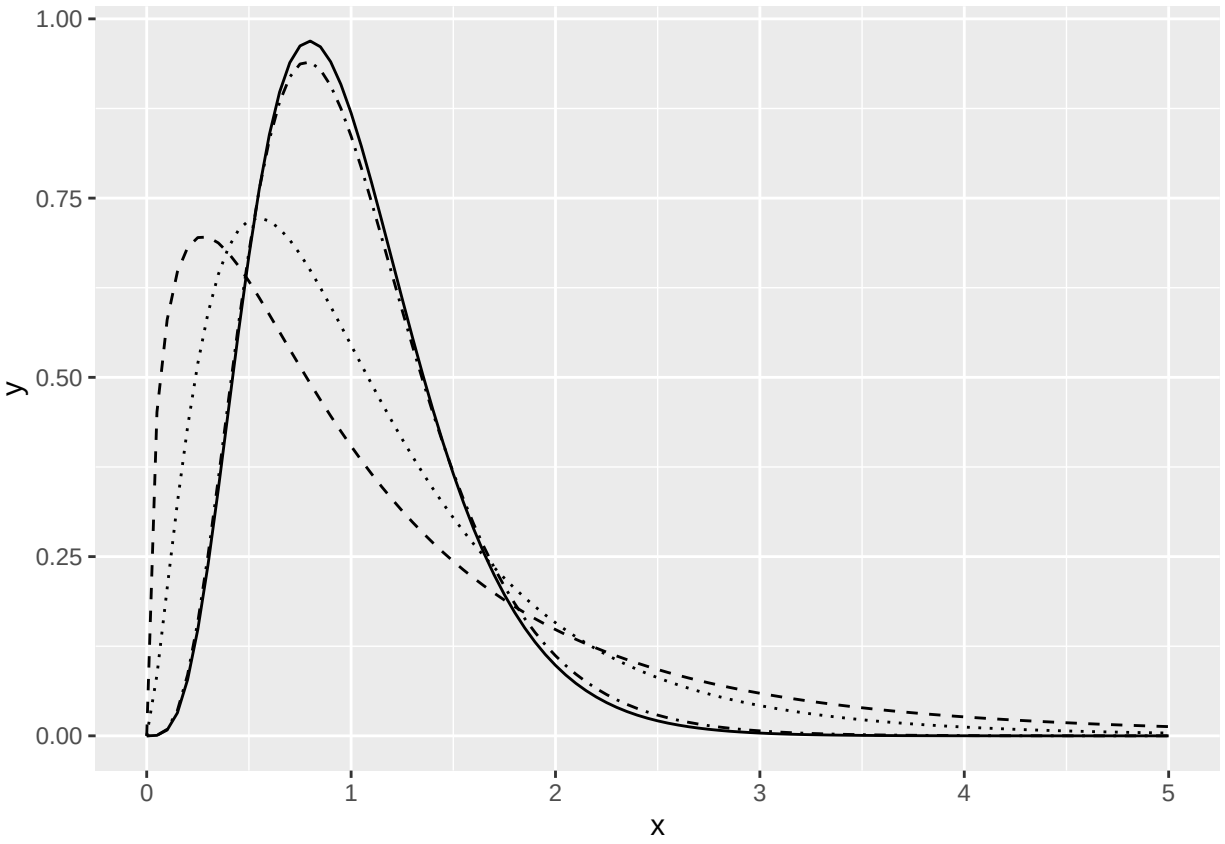
```
tibble(x=c(0,30),y=c(0,0.5)) %>%
  ggplot(aes(x,y)) +
#  stat_function(fun=function(x)dchisq(x,1),linetype="dashed") +
  stat_function(fun=function(x)dchisq(x,5),linetype="dotted") +
  stat_function(fun=function(x)dchisq(x,10),linetype="dotdash") +
  stat_function(fun=function(x)dchisq(x,20),linetype="solid")
```



Wenn $X_1, X_2 \dots X_n$ und $Y_1, Y_2 \dots Y_m$ normalverteilt mit $\mu = 0$ und $\sigma = 1$ sind, dann heißt die Verteilung von $\Xi = \frac{\frac{\sum_{i=1}^{n} X_i^2}{n}}{\frac{\sum_{i=1}^{m} Y_i^2}{m}}$ χ^2-Verteilung. Die Dichtefunktion ist

$f(x|n,m) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{=\frac{x}{2}}$

```
tibble(x=c(0,5),y=c(0,df(0,1,10))) %>%
  ggplot(aes(x,y)) +
  stat_function(fun=function(x)df(x,3,10),linetype="dashed") +
  stat_function(fun=function(x)df(x,5,20),linetype="dotted") +
  stat_function(fun=function(x)df(x,10,100),linetype="dotdash") +
  stat_function(fun=function(x)df(x,10,500),linetype="solid")
```



Gesetz der Großen Zahlen und Zentraler Grenzwertsatz

$X_1, X_2, \ldots, X_n, \ldots$ sind unabhängige Zufallsvariable mit der gleichen Verteilung und mit $E(X_i) = \mu$ und $V(X_i) = \sigma^2$

$S_n = \sum_{i=1}^{n} X_i$ die Summe der Zufallsvariablen und $\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ der Mittelwert dieser Zufallsvariablen.

Der gilt für den Erwartungswert $E(\frac{S_n}{n}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$
und für die Varianz $V(\frac{S_n}{n}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$

Für die Standardabweichung von $\frac{S_n}{n}$ gilt daher $\sigma(\frac{S_n}{n}) = \sqrt{V(\frac{S_n}{n})}) = \frac{\sigma}{\sqrt{n}}$

Je größer n wird, desto kleiner wird die Schwankung, also gilt das
Gesetz der großen Zahlen: Mit zunehmendem n konvergiert eine Zufallsvariable gegen den (konstanten, nicht zufälligen) Erwartungswert der einzelnen Verteilungen, die summiert werden.

Wir bilden jetzt $Z_n = \frac{S_n - \mu}{\sigma\sqrt{n}}$.
Für Z_n gilt $E(X) = 0$ und $V(Z) = 1$. Eine Zufallsvariable mit diesen Eigenschaften nennt man standardisierte Zufallsvariable.

Der Zentrale Grenzwertsatz besagt, dass die Zufallsvariablen Z_n gegen eine Normalverteilung mit $mu = 0$ und $sigma = 1$ konvergieren, dass also die Verteilung von Z_n bei großem n annähernd eine Normalverteilung ist.

Äquivalent ausgedrückt heißt das, dass die Verteilung von S_n annähernd eine Normalverteilung mit Erwartungswert $n\mu$ und Standardabweichung $\sqrt{n}\sigma$ und die Verteilung von $\frac{S_n}{n}$ annähernd eine Normalverteilung mit Erwartungswert μ und Standardabweichung $\frac{\sigma}{\sqrt{n}}$ ist.

Statistische Tests

Die Grundidee eines statistischen Tests:

Wir haben eine Familie von Wahrscheinlichkeitsverteilungen, die für ein Experiment in Frage kommen, beispielsweise Binomialverteilungen $B(x, n, p)$ für festes n und $0 \leq p \leq 1$. Das Experiment modellieren wir mit einer Zufallsvariablen X. Außerdem haben wir eine Nullhypothese, also eine Annahme über das p, z.B. $p = \frac{1}{2}$, der diese Zufallsvariable folgt.

Der Test soll jetzt helfen zu entscheiden, ob die beobachteten Daten mit der Nullhypothese vereinbar sind oder die Annahme rechtfertigen, dass die Daten nicht aus der Verteilung mit der Annahme der Nullhypothese stammen. So ein Verfahren kann niemals fehlerfrei funktionieren. Deshalb geben wir eine α-Irrtumswahrscheinlichkeit vor. Mit dieser Wahrscheinlichkeit lassen wir zu, dass das Testverfahren die Nullhypothese ablehnt, obwohl die dahinterliegende Verteilung der Nullhypothese entspricht.

Dazu konstruieren wir einen Bereich, in dem die Zufallsgröße im Nullhypothesenfall mit einer Wahrscheinlichkeit von mindestes $1 - \alpha$ liegt.

Wenn der beobachtete Wert unserer Zufallsvariablen in diesem Annahmebereich liegt, dann nehmen wir die Nullhypothese an, wenn er außerhalb liegt, dann lehnen wir die Nullhypothese ab.

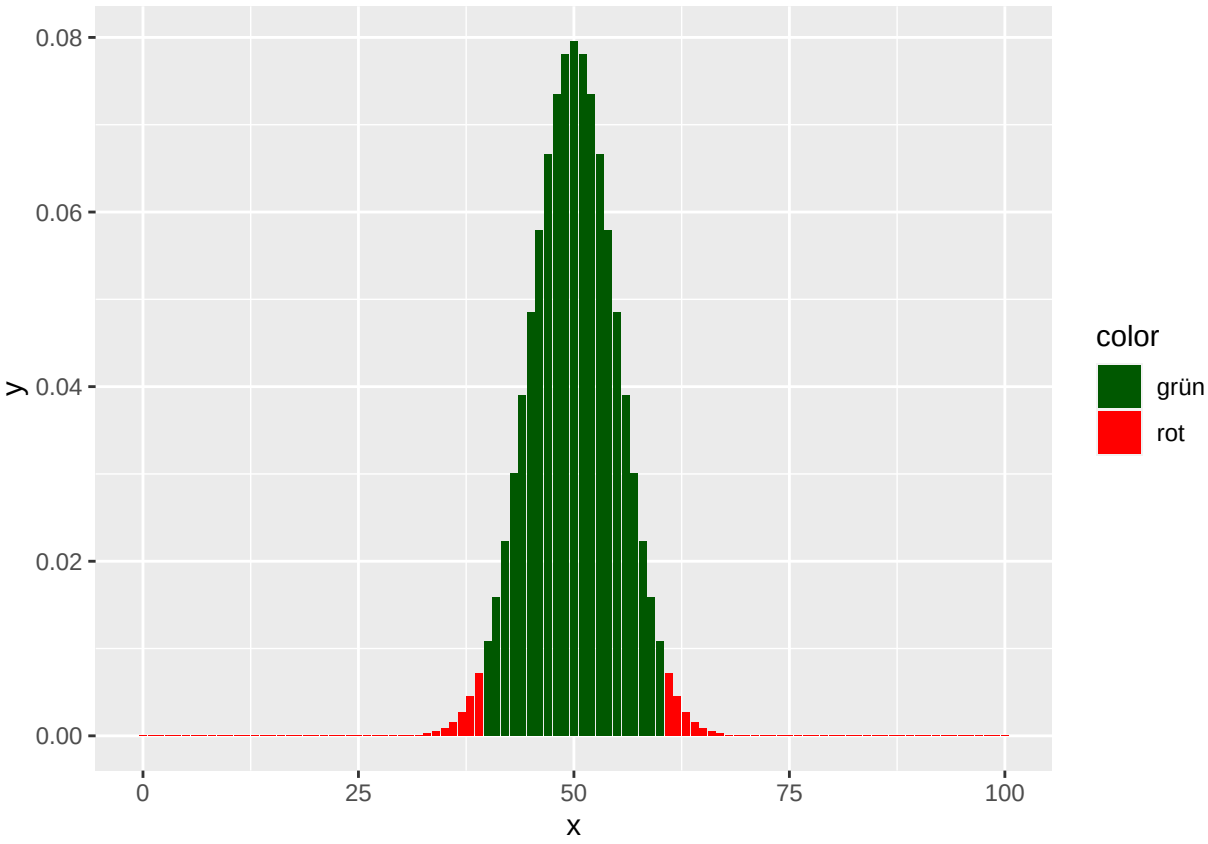
Im Falle der Binomialverteilung geht das so (der Annahmebereich ist grün, der Ablehnungsbereich rot):

```
p=1/2
n=100
alpha <- 0.05
lower_limit <- qbinom(alpha/2,n,1/2)
upper_limit <- n-lower_limit

bin_df <- function(p){
  tibble(
  x=0:n,
  y=dbinom(x,n,p),
  color= ifelse((x >= lower_limit) & (x <= upper_limit), "grün", "rot")
)}

bar_colors <- c(rot="red",grün=muted("green"))


bin_df(1/2) %>%
  ggplot(aes(x=x,y=y,fill=color)) +
  geom_col() +
  scale_fill_manual(values=bar_colors)
```

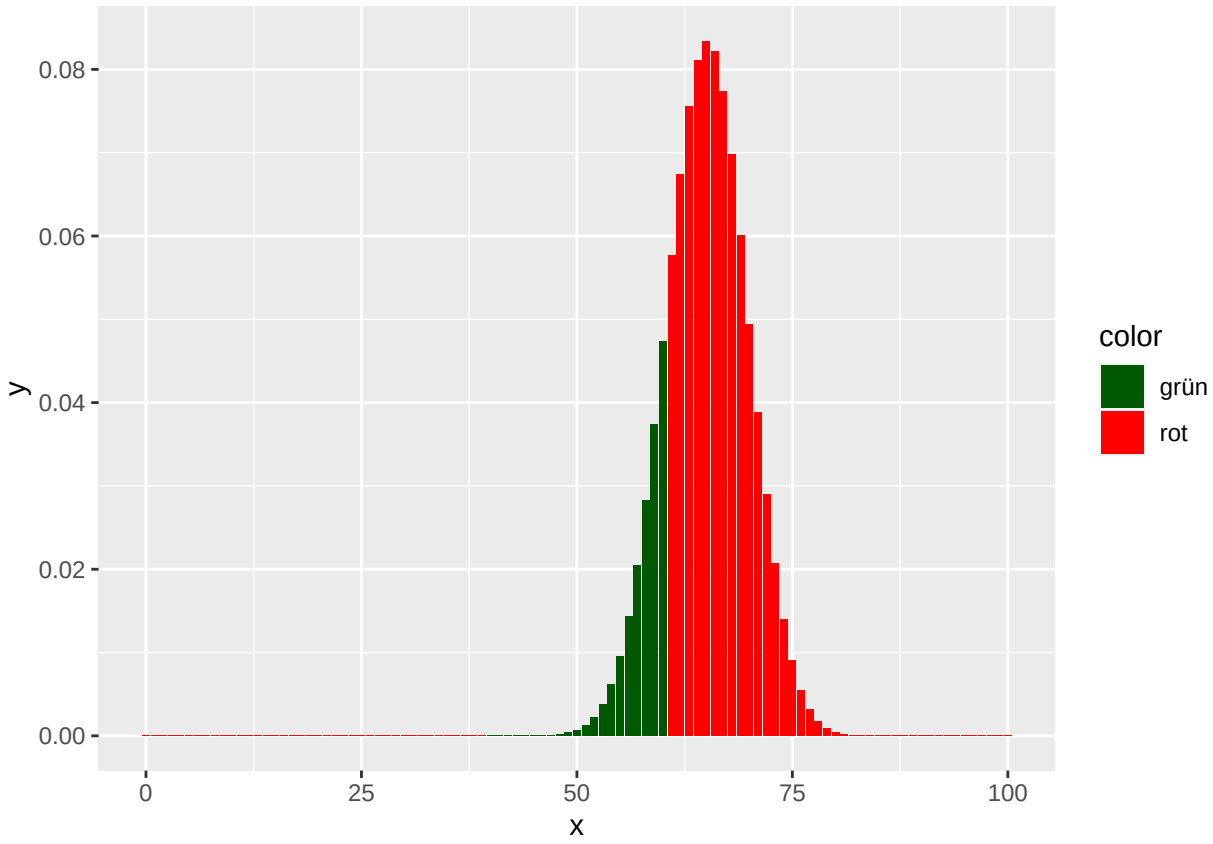


Wenn der beobachtete Wert im grünen Bereich liegt, dann wird die Nullhypothese angenommen, sonst wird sie abgelehnt. Da in unserem konkreten Fall die Grenzen `rlower_limit` und 60 sind, wird etwa bei einem Versuchsergebnis vom 54 die Nullhypothese, dass das Versuchsergebnis aus einer Binomialverteilung mit $p = \frac{1}{2}$ stammt, angenommen. Wenn das Versuchergebnis 33 ist, dann wird die Nullhypothese abgelehnt.

Wenn wir wissen wollen, mit welcher Wahrscheinlichkeit eine Abweichung von der Nullhypothese (also $p \neq \frac{1}{2}$) ist, dann berechnen wir die Wahrscheinlichkeit, dass eine Zufallsvariable mit dieser Verteilung in den mit Hilfe der $p = \frac{1}{2}$ verteilten Nullhypothesenverteilung definierten roten Bereich fällt.

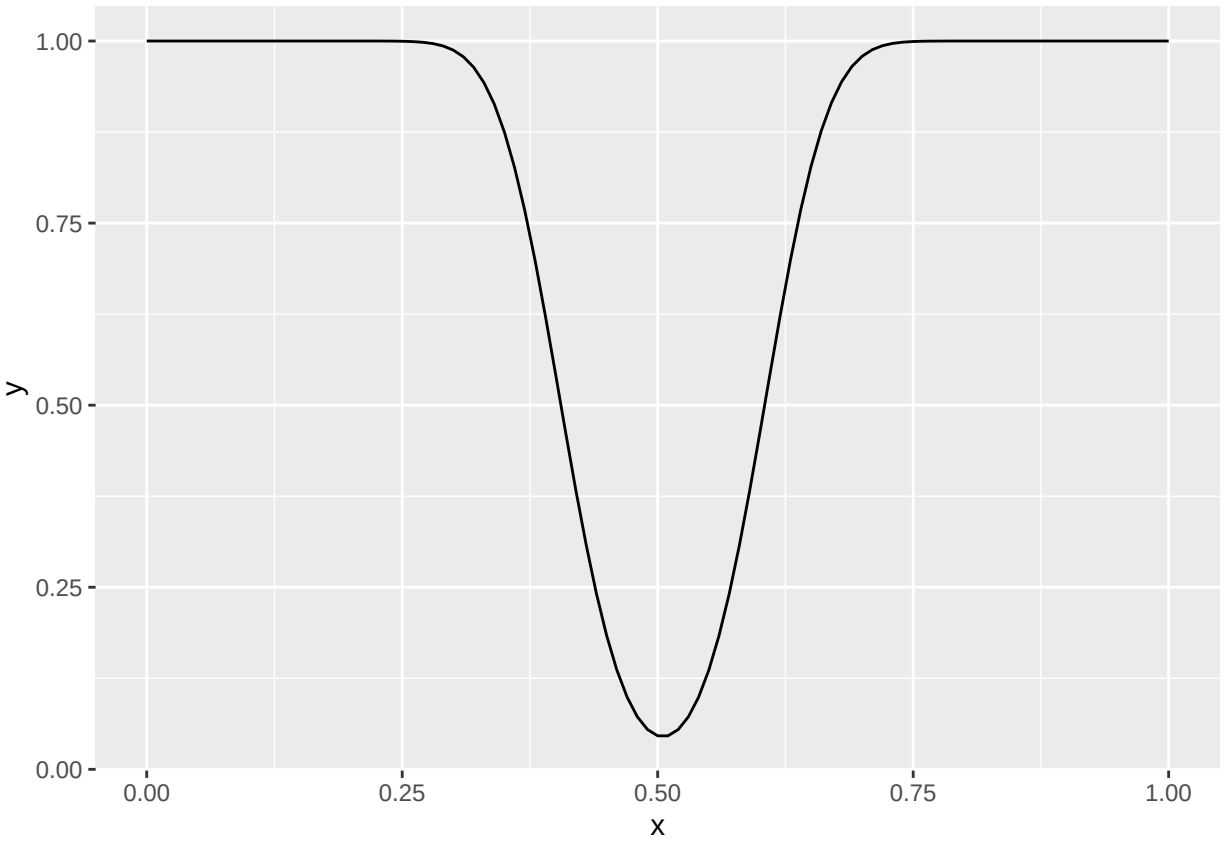
Hier das Beispiel mit einer Binomialverteilung mit $p = 0.65$

```
bin_df(0.65) %>%
  ggplot(aes(x=x,y=y,fill=color)) +
  geom_col() +
  scale_fill_manual(values=bar_colors)
```



Das als Funktion von p berechnet sieht so aus:

```
tibble(x=c(0,1),y=c(0,1)) %>%
  ggplot(aes(x=x,y=y)) +
  stat_function(fun=function(p)pbinom(lower_limit,n,p)+1-pbinom(upper_limit,n,p))
```



Binomialtest mit Normalverteilungsapproximation

X Binomialverteiltes Experiment

$\hat{p}$ Anteil in der Stichprobe

p_0 Nullhypothese

$\frac{\hat{p}-p_0}{\sqrt{n\hat{p}(1-\hat{p})}}$ ist normalverteilt.

Wir rechnen mit Sicherheit $\alpha = 0.95$

$$-1.96 \leq \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq 1.96$$

Ungeformt

$$\hat{p} - 1.96\sqrt{n\hat{p}(1-\hat{p})} \leq p \leq \hat{p} + 1.96\sqrt{n\hat{p}(1-\hat{p})}$$

Das ist ein Konfidenzintervall

Wichtige statistische Tests

t-Test

Wenn wir 2 Messreihen $x_1, x_2, \ldots x_n$ und $y_1, y_2 \ldots y_m$ haben, dann sind nach dem zentralen Grenzwertsatz die Mittelwerte $\bar{x}$ und $\bar{y}$ annähernd normalverteilt. Die Stichprobenvarianzen sind dann s_x^2 und s_y^2. Beide Datenreihen entstammen jeweils einer Normalverteilung mit gegebenem Mittelwert, die Mittelwerte der beiden Verteilungen können gleich oder verschieden sein.

Die Nullhypothese, dass die Mittelwerte der beiden Verteilungen gleich sind, testet man mit der Modellannahme, dass die Varianzen der beiden Normalverteilungen gleich sind, so:

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

Ist in diesem Fall $t-$verteilt mit $n-m-2$ Freiheitsgraden.

Ein entsprechender Test geht so:

```
studdat %>%
  select(Size,Gender) %>%
  drop_na() %$%
  t.test(Size ~ Gender,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Size by Gender
## t = -31.066, df = 1102, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.90592 -12.25367
## sample estimates:
## mean in group female   mean in group male
##             167.1845             180.2643
```

Wenn man nicht voraussetzen kann, dass die Varianzen in den beiden Normalverteilungen gleich sind, dann lautet der t-Test

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_1}+\frac{s_y^2}{m_1}}}$$

Diese Testgröße ist t-verteilt mit $\frac{(\frac{s_x^2}{n_1}+\frac{s_y^2}{m_1})^2}{\frac{(\frac{s_x^2}{n})^2}{n-1}+\frac{(\frac{s_y^2}{m})^2}{m-1}}$ Freiheitsgraden.

Ein entsprechender Test geht so:

```
studdat %>%
  select(Size,Gender) %>%
  drop_na() %$%
  t.test(Size ~ Gender,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Size by Gender
## t = -31.814, df = 675.76, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.88706 -12.27254
## sample estimates:
## mean in group female   mean in group male
##             167.1845             180.2643
```

χ^2-Test

Ein weiterer wichtiger Test ist der χ^2-Test.

Dazu gibt es 2 Varianten.

Die erste Variante testet, ob eine gegebene Häufigkeitsverteilung mit einer theoretischen Verteilung „zusammenpasst“.

Wenn es n verschiedene mögliche Werte gibt und o_i die beobachteten absoluten Häufigkeiten und e_i die erwarteten absoluten Häufigkeiten sind, dann ist die Testgröße

$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$

nach einer χ^2 mit $n-1$ Freiheitsgraden verteilt.

Wenn also bei 100x würfeln die Augenzahlen mit folgender Häufigkeit auftreten:

```
wuerfel <- tribble(
  ~Augen, ~Anzahl,
 1, 13,
2,  18,
3,  14,
4,  17,
5,  16,
6,  22
)
```

dann können wir so testen, ob alle 6 Augenzahlen gleich wahrscheinlich sind:

```
wuerfel %$% Anzahl %>% chisq.test()
```

```
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 3.08, df = 5, p-value = 0.6877
```

Der χ^2-Test wird auch angewendet, wenn es darum geht, ob 2 Merkmale A und B voneinander unabhängig sind, oder, anders gesagt, die Häufigkeitsverteilung für Merkmal B in allen durch Merkmal A erzeugten Untergruppen gleich ist.

Wenn wir mit den Studierendendaten testen wollen, ob die Häufigkeitsverteilung der Mathematiknoten bei Frauen und Männern gleich ist, dann geht das so:

```
studdat %>%
  select(Gender,Mathgrade) %>%
  drop_na() %>% table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 2.8012, df = 3, p-value = 0.4233
```

Varianzalyse

```
studdat %>%
  select(SmokeMother,Size) %>%
  drop_na() %>%
  aov(Size ~ SmokeMother,data=.) %>%
  summary()
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## SmokeMother     1    111  110.70   1.421  0.233
## Residuals    1083  84352   77.89
```

```
studdat %>%
  select(SmokeMother,Size) %>%
  drop_na() %>%
  lm(Size ~ SmokeMother,data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = Size ~ SmokeMother, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4025  -6.4025   0.5975   6.5975  28.5975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    176.4025     0.2947 598.642   <2e-16 ***
## SmokeMotheryes  -0.8439     0.7079  -1.192    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.825 on 1083 degrees of freedom
## Multiple R-squared:  0.001311,   Adjusted R-squared:  0.0003885
## F-statistic: 1.421 on 1 and 1083 DF,  p-value: 0.2335
```

Korrelations- und Regressionsanalyse

```
studdat %>%
  filter(Gender=="male") %>%
  select(Size,Weight) %>%
  drop_na() %>%
  lm(Weight ~ Size, data=.)
```

```
##
## Call:
## lm(formula = Weight ~ Size, data = .)
##
## Coefficients:
## (Intercept)         Size
##    -72.2267       0.8159
```

```
studdat %>%
  filter(Gender=="male") %>%
  select(Size,Weight) %>%
  drop_na() %>%
  lm(Weight ~ Size, data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = Weight ~ Size, data = .)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.800  -6.641  -1.641   4.807  42.200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -72.22675    9.96601  -7.247 1.05e-12 ***
## Size          0.81593    0.05526  14.765  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.967 on 759 degrees of freedom
## Multiple R-squared:  0.2231, Adjusted R-squared:  0.2221
## F-statistic:   218 on 1 and 759 DF,  p-value: < 2.2e-16
```

```
studdat %>%
  filter(Gender=="female") %>%
  select(Size,Weight) %>%
  drop_na() %>%
  lm(Weight ~ Size, data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = Weight ~ Size, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.262  -4.783  -1.311   3.527  35.936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.85540   11.71000  -3.831 0.000154 ***
## Size          0.61975    0.06996   8.859  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.713 on 318 degrees of freedom
## Multiple R-squared:  0.1979, Adjusted R-squared:  0.1954
## F-statistic: 78.48 on 1 and 318 DF,  p-value: < 2.2e-16
```

```
studdat %>%
  filter(Gender=="male") %>%
  select(Size,SizeFather,SizeMother) %>%
  drop_na() %>%
  lm(Size ~ SizeFather + SizeMother, data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = Size ~ SizeFather + SizeMother, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -20.7547  -3.2280  -0.1132   3.4133  17.7615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.63540    7.49329   7.959 8.28e-15 ***
## SizeFather   0.33697    0.03452   9.761  < 2e-16 ***
## SizeMother   0.36837    0.03917   9.404  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.464 on 622 degrees of freedom
## Multiple R-squared:  0.2949, Adjusted R-squared:  0.2926
## F-statistic:   130 on 2 and 622 DF,  p-value: < 2.2e-16
```

```
studdat %>%
  filter(Gender=="male") %>%
  drop_na() %>%
  select(Size,SizeFather,SizeMother,SmokeMother) %>%
  drop_na() %>%
  lm(Size ~ SizeFather + SizeMother + SmokeMother,
     data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = Size ~ SizeFather + SizeMother + SmokeMother, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4629  -3.0638  -0.0056   3.2228  17.5505
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    60.98240    9.41052   6.480 2.62e-10 ***
## SizeFather      0.33770    0.04330   7.798 5.22e-14 ***
## SizeMother      0.36009    0.05011   7.186 3.17e-12 ***
## SmokeMotheryes -0.96819    0.70799  -1.368    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.491 on 411 degrees of freedom
## Multiple R-squared:  0.2827, Adjusted R-squared:  0.2775
## F-statistic: 53.99 on 3 and 411 DF,  p-value: < 2.2e-16
```