Wahrscheinlichkeiten und Verteilungen

Recap zu Wahrscheinlichkeiten

Was sind Wahrscheinlichkeiten?

In diesem Abschnitt wollen wir die Grundbegriffe und Zusammenhänge aus der Statistik noch einmal auffrischen. Sie können sich Wahrscheinlichkeiten folgendermaßen vorstellen:

Recap zu Wahrscheinlichkeiten

Was sind Wahrscheinlichkeiten?

In diesem Abschnitt wollen wir die Grundbegriffe und Zusammenhänge aus der Statistik noch einmal auffrischen. Sie können sich Wahrscheinlichkeiten folgendermaßen vorstellen:

> Wenn Sie ein Experiement unabhängig voneinander und unter den selben Bedingungen sehr oft wiederholen können Sie bestimmen mit welchem Anteil ein bestimmtes Ereignis eintritt. Hier sprechen wir von der **Wahrscheinlichkeit** das dieses Ereignis eintritt.

Diskrete Wahrscheinlichkeiten

Wir wollen uns hier ein paar Beispiele für diskrete Wahrscheinlichkeiten anschauen um unsere Kenntnisse aufzufrischen.

- Diskrete Wahrscheinlichkeiten lassen sich sehr gut mittels Experimente am PC, d.h. Simulationen, nachvollziehen
- Diskrete Wahrscheinlichkeiten begegnen uns z.B. bei Kartenspielen oder beim Ziehen aus einer Urne.

Diskrete Wahrscheinlichkeiten

Wir wollen uns hier ein paar Beispiele für diskrete Wahrscheinlichkeiten anschauen um unsere Kenntnisse aufzufrischen.

- Diskrete Wahrscheinlichkeiten lassen sich sehr gut mittels Experimente am PC, d.h. Simulationen, nachvollziehen
- Diskrete Wahrscheinlichkeiten begegnen uns z.B. bei Kartenspielen oder beim Ziehen aus einer Urne.

In einer Urne befinden sich 2 rote und 3 blauen Kugel, wenn Sie aus dieser Urne zufällig eine Kugel ziehen, wie hoch ist die Wahrscheinlichkeit eine rote Kugel zu ziehen?

Diskrete Wahrscheinlichkeiten

Wir wollen uns hier ein paar Beispiele für diskrete Wahrscheinlichkeiten anschauen um unsere Kenntnisse aufzufrischen.

- Diskrete Wahrscheinlichkeiten lassen sich sehr gut mittels Experimente am PC, d.h. Simulationen, nachvollziehen
- Diskrete Wahrscheinlichkeiten begegnen uns z.B. bei Kartenspielen oder beim Ziehen aus einer Urne.

In einer Urne befinden sich 2 rote und 3 blauen Kugel, wenn Sie aus dieser Urne zufällig eine Kugel ziehen, wie hoch ist die Wahrscheinlichkeit eine rote Kugel zu ziehen?

> 40% -> 2/5 da jede Kugel die gleiche Wahrscheinlichkeit hat gezogen zu werden

Monte Carlo Simulation

> Können Sie am Computer Experimente simulieren um die Wahrscheinlichkeit für ein Ereignis zu berechnen?

Monte Carlo Simulation

> Können Sie am Computer Experimente simulieren um die Wahrscheinlichkeit für ein Ereignis zu berechnen?

Ja, durch Monte-Carlo-Simulationen!

Durch Monte Carlo Simulationen können Sie am Computer bestimmte Aktionen beliebig oft durchspielen

- ✚ Theoretisch optimal wäre es das Experiment unendlich oft zu wiederholen
- ✚ Praktisch sollte es so oft wiederholt werden, dass eine weitere Wiederholung das Ergebnis nur noch unwesentlich verändert $\rightarrow$ Ergebnis konvergiert zum *wahren* Wert
- ✚ Statistisch gesehen nähern Sie sich der tatsächlichen Wahrscheinlichkeit mit steigendem N an

Monte Carlo Simulation

> Können Sie am Computer Experimente simulieren um die Wahrscheinlichkeit für ein Ereignis zu berechnen?

Ja, durch Monte-Carlo-Simulationen!

Durch Monte Carlo Simulationen können Sie am Computer bestimmte Aktionen beliebig oft durchspielen

- Theoretisch optimal wäre es das Experiment unendlich oft zu wiederholen
- Praktisch sollte es so oft wiederholt werden, dass eine weitere Wiederholung das Ergebnis nur noch unwesentlich verändert $\rightarrow$ Ergebnis konvergiert zum *wahren* Wert
- Statistisch gesehen nähern Sie sich der tatsächlichen Wahrscheinlichkeit mit steigendem N an

Bei einer Simulation sollten Sie *immer* einen sogenannten "seed" setzen!

- Generierung von Zufallszahlen
- **Allerdings** werden durch den "seed" immer die gleichen Zufallszahlen generiert

Monte Carlo Simulation

> Wie funktioniert eine Monte-Carlo-Simulation in R?

Monte Carlo Simulation

Wie funktioniert eine Monte-Carlo-Simulation in R?

Durch die `sample` Funktion können wir zufällig eine Kugel aus unserer eben beschriebenen Urne ziehen:

```
urne <- rep( c("rot", "blau"), times = c(2,3))
urne
```

```
[1] "rot"  "rot"  "blau" "blau" "blau"
```

```
sample(urne, 1) # zufälliges Ziehen
```

```
[1] "rot"
```

Mit der Funktion `replicate` können wir diesen Vorgang n-mal wiederholen.

Monte Carlo Simulation

Wir ziehen im folgenden Beispiel $N = 10\,000$ mal zufällig aus der Urne:

```
N <- 10000
ereignis <- replicate(N, sample(urne, 1))
```

Monte Carlo Simulation

Wir ziehen im folgenden Beispiel $N = 10\,000$ mal zufällig aus der Urne:

```
N <- 10000
ereignis <- replicate(N, sample(urne, 1))
```

Betrachten wir nun die Approximation, welche wir durch unsere Monte Carlo Simulation erhalten:

```
tab <- table(ereignis)
prop.table(tab)
```

```
ereignis
  blau    rot
0.6017 0.3983
```

- Statistisch gesehen nähern wir uns den tatsächlichen Wahrscheinlichkeiten mit steigendem N an

Mit und ohne Zurücklegen

Wir können auch ohne die Funktion `replicate` unser Experiment mehrmals wiederholen, wenn wir die Funktion `sample` nicht als ziehen *ohne* Zurücklegen, sondern als ziehen *mit* Zurücklegen spezifizieren. D.h. wir legen die gezogene Kugel in die Urne zurück, nachdem wir sie gezogen haben.

- ✚ Hierzu ergänzen wir das Argument `replace = TRUE`

```
ereignis <- sample(urne, N, replace = TRUE)
prop.table(table(ereignis))
```

```
ereignis
  blau    rot
0.5883 0.4117
```

Unabhängigkeit

Im ersten Beispiel hatten wir uns mit einer Urne und blauen und roten Kugeln beschäftigt.

- Hier ist die Wahrscheinlichkeit eine blaue Kugel aus der Urne zu bekommen recht offensichtlich
- Dies ist jedoch bei etwas komplexeren Problemen oft nicht der Fall
  - Beispielsweise: Wie hoch ist die Wahrscheinlichkeit beim Black Jack zwei Asse auf die Hand zu bekommen?

Unabhängigkeit

Im ersten Beispiel hatten wir uns mit einer Urne und blauen und roten Kugeln beschäftigt.

- Hier ist die Wahrscheinlichkeit eine blaue Kugel aus der Urne zu bekommen recht offensichtlich
- Dies ist jedoch bei etwas komplexeren Problemen oft nicht der Fall
  - Beispielsweise: Wie hoch ist die Wahrscheinlichkeit beim Black Jack zwei Asse auf die Hand zu bekommen?
- In Statistik lernen Sie die Theorie, wie Sie dies berechnen könne

In diesem Kurs wollen wir das Ergebnis von R errechnen lassen.

Unabhängigkeit

Im vorherigen Urnen-Beispiel hatten wir uns unabhängige Ereignisse angeschaut. In diesem Fall war das Ziehen von mehreren Kugeln voneinander *unabhängig*, da der erste Zug den zweiten *nicht beeinflusst* (ziehen mit zurücklegen).

Zwei Ereignisse sind nicht voneinander unabhängig, wenn das erste Ereignis das zweite Ereignis beeinflusst.

- Beispiel: Black Jack oder Poker.
- Die Wahrscheinlichkeit als erste Karte ein Ass zu bekommen liegt bei 4/52, da es 4 Asse bei 52 unterschiedlichen Karten gibt
- Die Wahrscheinlichkeit danach noch einmal ein Ass zu bekommen liegt bei 3/51 und ist somit abhängig von der ersten Karte.
  - Es fehlt eine Karte im Stapel und diese Karte war ein Ass

Bedingte Wahrscheinlichkeiten

Wenn Ereignisse nicht unabhängig voneinander sind können wir sogenannte *bedingte Wahrscheinlichkeiten* berechnen. In der vorherigen Folie haben wir ein Beispiel für eine bedingte Wahrscheinlichkeit gesehen:

$$\Pr(2.\ \text{Karte ist ein Ass} \mid 1.\ \text{Karte ist ein Ass}) = 3/51$$

Der senkrechte Strich $\mid$ wird in der Mathematik oft als Argument 1 "gegeben" Argument 2 verstanden.

Bedingte Wahrscheinlichkeiten

Wenn Ereignisse nicht unabhängig voneinander sind können wir sogenannte *bedingte Wahrscheinlichkeiten* berechnen. In der vorherigen Folie haben wir ein Beispiel für eine bedingte Wahrscheinlichkeit gesehen:

$$\Pr(2.\ \text{Karte ist ein Ass} \mid 1.\ \text{Karte ist ein Ass}) = 3/51$$

Der senkrechte Strich $\mid$ wird in der Mathematik oft als Argument 1 "gegeben" Argument 2 verstanden.

Sind zwei Ereignisse A und B unabhängig, so gilt:

$$\Pr(A \mid B) = \Pr(A)$$

- Dies bedeutet: Auch wenn B passiert ist, so hat dies keine Auswirkung auf die Wahrscheinlichkeit das A eintritt.

Multiplikationsregel

Gegeben uns interessiert die Wahrscheinlichkeit das zwei Ereignisse, A und B stattfinden, so berechnet sich dies als:

$$\Pr(A \text{ und } B) = \Pr(A)\ \Pr(B \mid A)$$

Multiplikationsregel

Bleiben wir beim Beispiel Black Jack:

- Wir bekommen zufällig zwei Karten
- Nachdem wir wissen welche Karten wir haben können wir zusätzliche Karten verlangen
- Wahrscheinlichkeit einen Black Jack (Kartenwert 21) durch ein Ass als 1. Karte und eine Bildkarte als 2. Karte zu bekommen liegt bei: $4/52 \times 16/51 \approx 0.024$
  - Erste Karte ist ein Ass
  - Zweite Karte ist eine Bildkarte gegeben, dass die erste Karte ein Ass war

Multiplikationsregel

Die Multiplikationsregel kann auch für mehr als zwei Ereignissen angewendet werden:

$$\Pr(A \text{ und } B \text{ und } C) = \Pr(A)\ \Pr(B \mid A)\ \Pr(C \mid A \text{ und } B)$$

Bei voneinander unabhängigen Ereignissen vereinfacht sich die Berechnung zu:

$$\Pr(A \text{ und } B \text{ und } C) = \Pr(A)\ \Pr(B)\ \Pr(C)$$

Multiplikationsregel

Die Multiplikationsregel kann auch für mehr als zwei Ereignissen angewendet werden:

$$\Pr(A \text{ und } B \text{ und } C) = \Pr(A)\ \Pr(B \mid A)\ \Pr(C \mid A \text{ und } B)$$

Bei voneinander unabhängigen Ereignissen vereinfacht sich die Berechnung zu:

$$\Pr(A \text{ und } B \text{ und } C) = \Pr(A)\ \Pr(B)\ \Pr(C)$$

Die Multiplikationsregel können wir auch nutzen um bedingte Wahrscheinlichkeiten zu berechnen:

$$\Pr(B \mid A) = \frac{\Pr(A \text{ und } B)}{\Pr(A)}$$

Permutationen und Kombinationen

Wir wollen das Ganze in R simulieren.

> Als erstes brauchen wir ein virtuelles Kartendeck.

Dies erstellen wir uns mit der Funktion `paste` und `expand.grid`

- Mit `paste` können wir verschiedene String-Variablen zusammenfügen
- Mit `expand.grid` können wir dann alle möglichen Kombinationen von zwei Vektoren erzeugen lassen

Permutationen und Kombinationen

Wir wollen das Ganze in R simulieren.

> Als erstes brauchen wir ein virtuelles Kartendeck.

Dies erstellen wir uns mit der Funktion `paste` und `expand.grid`

- Mit `paste` können wir verschiedene String-Variablen zusammenfügen
- Mit `expand.grid` können wir dann alle möglichen Kombinationen von zwei Vektoren erzeugen lassen

```
farben <- c("Kreuz", "Pik", "Herz", "Karo")
zahlen <- c("Ass", "König", "Dame", "Bube", "Zehn", "Neun",
            "Acht", "Sieben", "Sechs", "Fünf", "Vier", "Drei", "Zwei")
deck <- expand.grid( farbe=farben, zahl=zahlen)
deck <- paste(deck$farbe, deck$zahl)
length(deck)
```

```
[1] 52
```

Permutationen

Da wir nun ein Deck mit 52 Karten haben können wir unsere vorherige Berechnungen gegenchecken:

```
Ass <- paste( farben, "Ass")
mean(deck %in% Ass) # erste Karte ein Ass
```

```
[1] 0.07692308
```

Permutationen

Da wir nun ein Deck mit 52 Karten haben können wir unsere vorherige Berechnungen gegenchecken:

```
Ass <- paste( farben, "Ass")
mean(deck %in% Ass) # erste Karte ein Ass
```

```
[1] 0.07692308
```

> Das war zu simple?

Permutationen

> Ok, wie steht es dann um die Wahrscheinlichkeit bei der zweiten Karte ein Ass zu bekommen?

Hierfür berechnen wir alle Kombinationen aus dem vorhandenen Deck.

- In R nutzen wir hierfür die Funktion `permutations` aus dem Paket `gtools`
  - Die Reihenfolge der Ziehung wird bei der Funktion `permutations` beachtet

```
library(gtools)
hände <- permutations(52, 2, v = deck)
erste <- hände[,1]
zweite <- hände[,2]
```

Permutationen

Um zu sehen in wie vielen Fällen die erste Karte ein Ass war:

```
Ass <- paste( farben, "Ass")
sum(erste %in% Ass)
```

```
[1] 204
```

Permutationen

Um zu sehen in wie vielen Fällen die erste Karte ein Ass war:

```
Ass <- paste( farben, "Ass")
sum(erste %in% Ass)
```

```
[1] 204
```

Um die bedingte Wahrscheinlichkeit für ein Ass auf der 2. Karte zu berechnen nehmen wir einfach den Anteil der 204 Karten, welche ein Ass als zweite Karte haben:

```
sum(erste %in% Ass & zweite %in% Ass) /
  sum(erste %in% Ass)
```

```
[1] 0.05882353
```

Permutationen

Um zu sehen in wie vielen Fällen die erste Karte ein Ass war:

```
Ass <- paste( farben, "Ass")
sum(erste %in% Ass)
```

```
[1] 204
```

Um die bedingte Wahrscheinlichkeit für ein Ass auf der 2. Karte zu berechnen nehmen wir einfach den Anteil der 204 Karten, welche ein Ass als zweite Karte haben:

```
sum(erste %in% Ass & zweite %in% Ass) /
  sum(erste %in% Ass)
```

```
[1] 0.05882353
```

Dadurch ergibt sich für "Pocket Aces" eine Wahrscheinlichkeit von (4/52) * (3/51) = 0,45%!

Kombinationen

Was jedoch, wenn uns die Reihenfolge egal ist?

In Black Jack wollen wir auf 21 kommen (Summe beider Kartenwerte). Wenn wir nun ein Ass und eine Bildkarte ausgeteilt bekommen, so haben wir gewonnen. Hierbei interessieren uns nicht die Permutationen, sondern die Kombinationen der Karten. D.h. es ist uns egal, ob wir mit der ersten oder der zweiten Karte ein Ass bzw. eine Bildkarte bekommen:

```
Asse <- paste( farben, "Ass")

bildkarte <- c("König", "Dame", "Bube", "Zehn")
bildkarte <- expand.grid( farbe=farben, zahl=bildkarte)
bildkarte <- paste( bildkarte$farbe, bildkarte$zahl )

hände <- combinations(52, 2, v = deck)
mean((hände[,1] %in% Asse & hände[,2] %in% bildkarte) |
       (hände[,2] %in% Asse & hände[,1] %in% bildkarte))
```

```
[1] 0.04826546
```

Monte Carlo Simulation

Neben der Möglichkeit alle Kombinationen auszutesten, um einen Black Jack direkt mit den ersten zwei ausgeteilten Karten zu bekommen, können wir auch eine Monte Carlo Simulation durchführen, um diese Wahrscheinlichkeit zu schätzen.

Hierbei ziehen wir aus unserem Deck immer wieder zwei Karten und notieren uns, wie oft wir dabei direkt eine 21 bekommen. Mit der Funktion `sample` können wir Karten *ohne zurücklegen* ziehen:

```
hand <- sample(deck, 2)
hand
```

```
[1] "Herz Zehn" "Herz Zwei"
```

Monte Carlo Simulation

Neben der Möglichkeit alle Kombinationen auszutesten, um einen Black Jack direkt mit den ersten zwei ausgeteilten Karten zu bekommen, können wir auch eine Monte Carlo Simulation durchführen, um diese Wahrscheinlichkeit zu schätzen.

Hierbei ziehen wir aus unserem Deck immer wieder zwei Karten und notieren uns, wie oft wir dabei direkt eine 21 bekommen. Mit der Funktion `sample` können wir Karten *ohne zurücklegen* ziehen:

```
hand <- sample(deck, 2)
hand
```

```
[1] "Herz Zehn" "Herz Zwei"
```

Anschließend schauen wir uns an, ob eine Karte davon ein Ass bzw. eine Bildkarte ist (wobei 10 zu den Bildkarten zählt)

```
(hand[1] %in% Ass & hand[2] %in% bildkarte) |
  (hand[2] %in% Ass & hand[1] %in% bildkarte)
```

```
[1] FALSE
```

Monte Carlo Simulation

Wir können uns auf Grundlage unserer vorherigen Überlegungen eine Funktion schreiben, welche beide Schritte miteinander verbindet.

Hierbei benötigt die Funktion keine Argumente, da alle Objekte in der globalen Umgebung definiert wurden

- D.h. das Deck aus dem gezogen wird ist fix, die Bildkarten sind definiert und die Asse auch
- Wir müssen explizit beide Möglichkeiten (Ass als erste Karte + Ass als zweite Karte) berechnen

```
black_jack <- function(){
   hand <- sample(deck, 2)
  (hand[1] %in% Ass & hand[2] %in% bildkarte) |
    (hand[2] %in% Ass & hand[1] %in% bildkarte)
}
```

Monte Carlo Simulation

Jetzt spielen wir das Ganze 100 000 mal und schauen wie hoch die Wahrscheinlichkeit ist direkt eine 21 zu erhalten

- Hierbei machen wir uns den Umstand zunutze, dass R intern `TRUE` als 1 und `FALSE` also 0 abspeichert

```
N <- 100000
ergebnis <- replicate(N, black_jack())
mean(ergebnis)
```

```
[1] 0.04798
```

Wir kommen hier auf die gleichen Werte wie in der exakten Berechnung!

Von diskreten zu stetigen Wahrscheinlichkeiten

Stetige Wahrscheinlichkeiten

Diskrete Wahrscheinlichkeiten geben uns einen guten Einblick in die Wahrscheinlichkeitstheorie.

Jedoch: In der empirischen Analyse haben wir es meist nicht mit einer Urne oder Kartenspielen zu tun, sondern wir betrachten z.B. die Körpergröße oder den IQ alle Individuen in Deutschland und möchten nun herausfinden, wie wahrscheinlich es ist, dass eine zufällig ausgewählte Person größer als 2 Meter ist.

Hier befinden wir uns im Bereich der stetigen Wahrscheinlichkeiten.

Stetige Wahrscheinlichkeiten

Diskrete Wahrscheinlichkeiten geben uns einen guten Einblick in die Wahrscheinlichkeitstheorie.

Jedoch: In der empirischen Analyse haben wir es meist nicht mit einer Urne oder Kartenspielen zu tun, sondern wir betrachten z.B. die Körpergröße oder den IQ alle Individuen in Deutschland und möchten nun herausfinden, wie wahrscheinlich es ist, dass eine zufällig ausgewählte Person größer als 2 Meter ist.

Hier befinden wir uns im Bereich der stetigen Wahrscheinlichkeiten.

Um stetige Wahrscheinlichkeiten greifbarer zu machen wiederholen wir noch einmal das Konzept der Verteilungen!

Verteilungen

Verteilungsfunktionen

> Hier ein einleitendes Beispiel um Verteilungen besser zu verstehen.

- Wir können uns Verteilungen als eine grafische Beschreibung einer Liste mit vielen numerischen Einträgen vorstellen.
  - Nehmen wir zur Verdeutlichung den Datensatz `wage2` aus dem Paket `wooldridge`, welcher Einkommen, Beziehungsstatus, Anzahl der Geschwister, IQ etc. von zufällig ausgewählten Amerikanern enthält
  - Hier haben wir einen kategorialen Vektor (eine Dummyvariable) `married`, welche 1 ist für alle verheirateten und 0 für Singles
  - Die Verteilung ist hier einfach der Anteil an Personen in jeder Kategorie

Verteilungsfunktionen

> Hier ein einleitendes Beispiel um Verteilungen besser zu verstehen.

- Wir können uns Verteilungen als eine grafische Beschreibung einer Liste mit vielen numerischen Einträgen vorstellen.
  - Nehmen wir zur Verdeutlichung den Datensatz `wage2` aus dem Paket `wooldridge`, welcher Einkommen, Beziehungsstatus, Anzahl der Geschwister, IQ etc. von zufällig ausgewählten Amerikanern enthält
  - Hier haben wir einen kategorialen Vektor (eine Dummyvariable) `married`, welche 1 ist für alle verheirateten und 0 für Singles
  - Die Verteilung ist hier einfach der Anteil an Personen in jeder Kategorie

```
prop.table(table(wage2$married))
```

```
        0         1 
0.1069519 0.8930481 
```

Verteilungsfunktionen

Variablen, welche durch eine kleine Gruppe definiert sind fallen unter *kategorische Daten*.

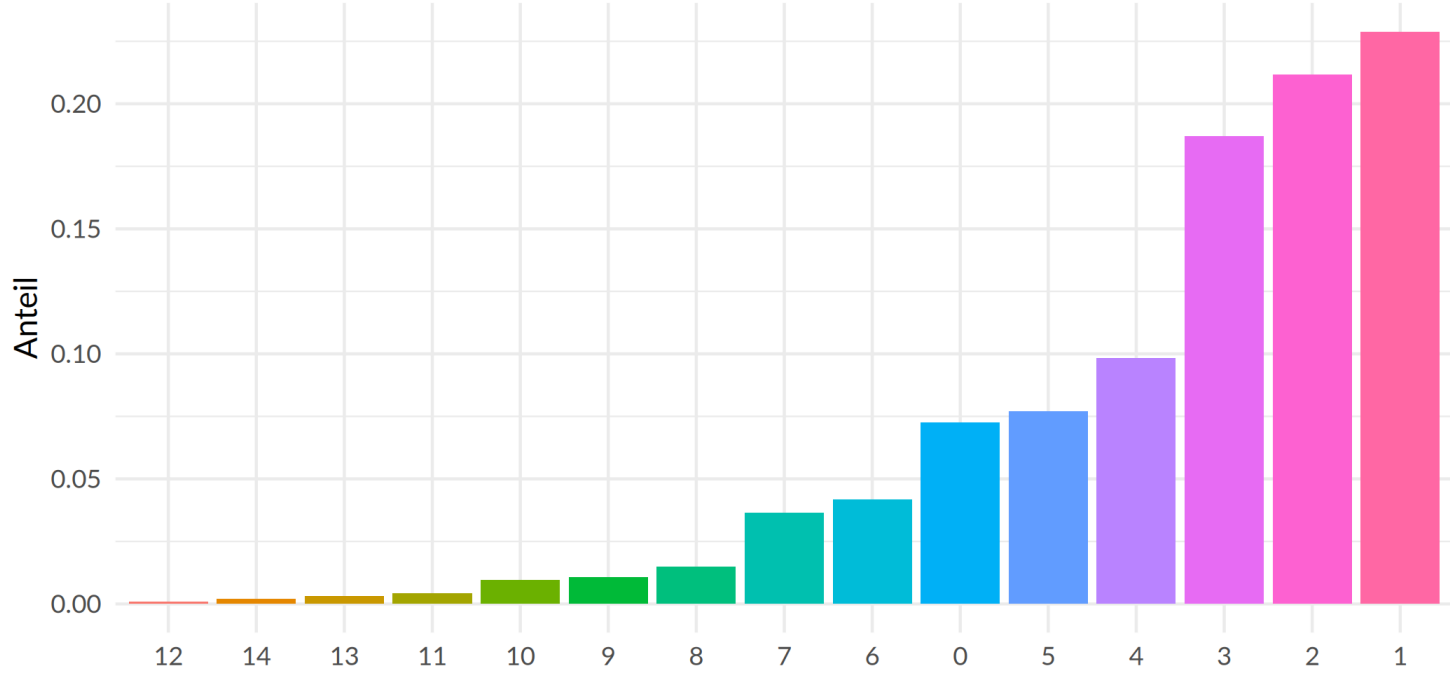
- Beispielsweise Geschlecht (m/w/d) oder Region (Nord, Süd, West, Ost).
- Wenn diese Daten geordnet sind, wie z.B. nach Hitzegrad (kalt, wohltemperiert, heiß), dann sprechen wir von *ordinalen Daten*.

Anders numerische Daten, wie beispielsweise die Bevölkerungsgröße, der IQ oder die Körpergröße. Numerische Variablen können in diskrete oder stetige Daten unterteilt werden.

- Stetige Daten können jeden Wert annehmen, wie z.B. bei der Körpergröße
- Diskrete Daten können nicht jeden beliebigen Zwischenschritt annehmen, sonder müssen zur nächsten (ganzen) Zahl gerundet werden. Z.B. die Bevölkerungsgröße wird auf 83 Mio. gerundet.

Verteilungsfunktionen

Gibt es mehrere Kategorien, wie z.B. bei der Anzahl der Geschwister im `wage2` Datensatz, dann kann ein einfaches Balkendiagramm die Verteilung darstellen:



Anteil mit x Geschwistern im Datensatz

Verteilungsfunktionen

Gegeben wir haben keine Kategorien sondern numerische Daten

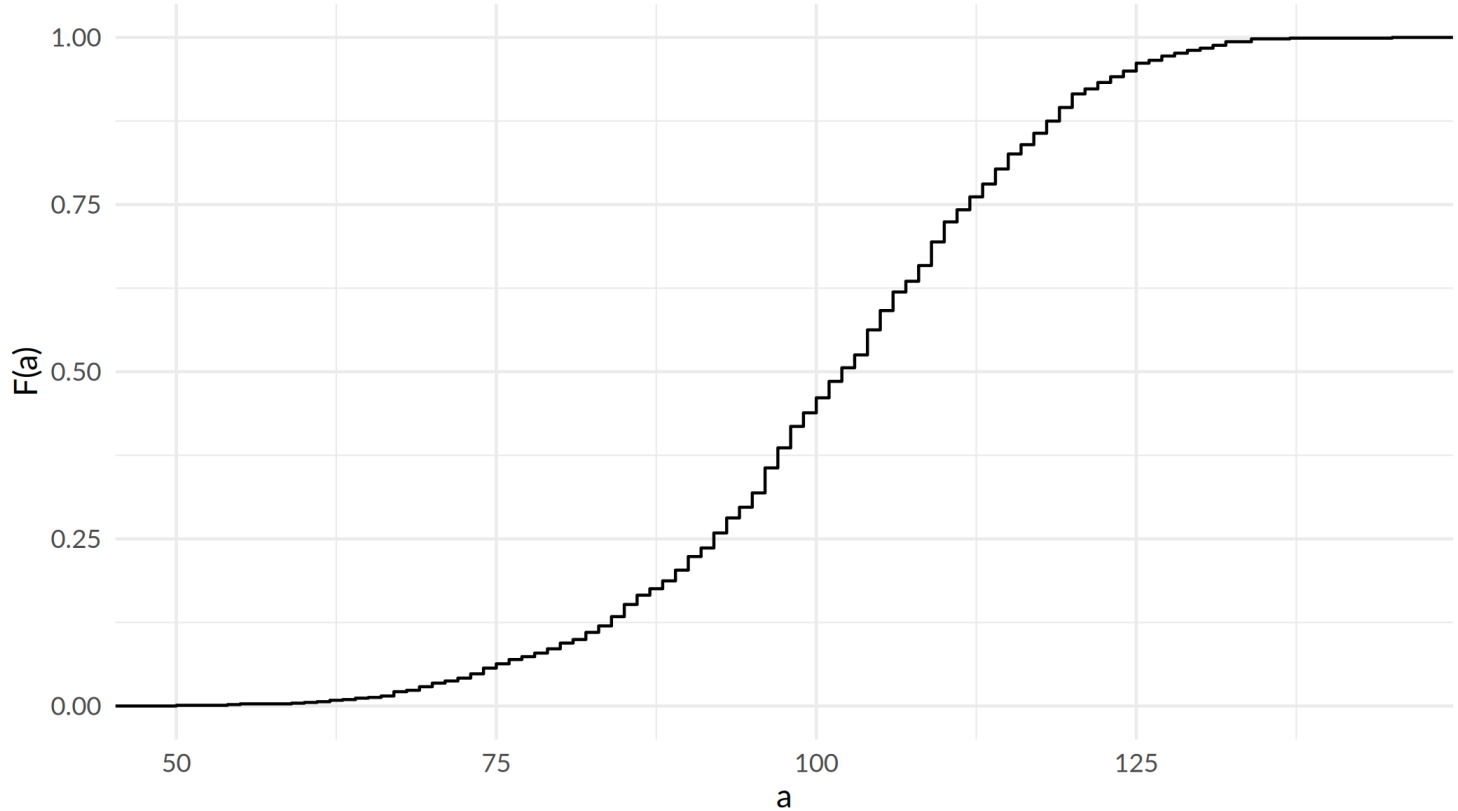
- Häufigkeitstabelle ist hier etwas ungeschickt, da nun der Vektor jeden beliebigen Zwischenschritt annehmen könnte, z.B. in den IQ Daten von `100.345623`
- Folglich gibt es sehr viele verschiedene Einträge
- In der Statistik wird hier die Verteilungsfunktion herangezogen.

Die Verteilungsfunktion zeigt den Anteil der Datenpunkte, welche kleiner gleich a sind für alle möglichen Werte von a (reelle Zufallsvariable). Folgende Formel repräsentiert die Verteilungsfunktion:

$$F(a) = \Pr(x \leq a)$$

Verteilungsfunktionen

Wir können die empirische Verteilungsfunktion für unsere IQ-Werte zeichnen:



Empirische Verteilungsfunktion für den IQ.

Verteilungsfunktionen

- Für jeden Wert a kann nun der dazugehörigen Anteil $F(a)$ an Personen, welchen einen IQ von a oder niedriger haben direkt abgelesen werden
  - $F(75) = 0.0631016$ oder $F(100) = 0.4609626$
  - Für jede beliebigen IQ-Werte im Intervall $[a, b]$ können wir durch $F(b) - F(a)$ den Anteil an Personen mit einem IQ zwischen a und b berechnen
  - D.h. der Anteil an Personen mit einem IQ zwischen 75 und 100 beträgt: 0.397861
- Da wir die Verteilungsfunktion F aus uns zur Verfügung stehenden Daten geschätzt haben, sprechen wir hier von der *empirischen* Verteilungsfunktion $\hat{F}_n$

Empirische Verteilungsfunktion

In der Theorie werden meist Verteilungsfunktionen F verwendet und diskutiert, welche wir durch die empirische Verteilungsfunktion $\hat{F}_n$ approximieren können.

Die Dichte f ist die Ableitung der Verteilungsfunktion F.

Empirische Verteilungsfunktion

In der Theorie werden meist Verteilungsfunktionen F verwendet und diskutiert, welche wir durch die empirische Verteilungsfunktion $\hat{F}_n$ approximieren können.

Die Dichte f ist die Ableitung der Verteilungsfunktion F.

Problem: Die empirische Verteilungsfunktion `(\hat{F}_n)` kann nicht abgeleitet werden, da diese nicht differenzierbar (und sogar nicht stetig) ist.

> Durch welches Schaubild wird die empirische Verteilungsfunktion in empirischen Arbeiten oft dargestellt?

Empirische Verteilungsfunktion

Alternative in der Praxis: Histogramm, Kerndichteschätzer

Empirische Verteilungsfunktion

Alternative in der Praxis: Histogramm, Kerndichteschätzer

- In Histogrammen können auf einen Blick verschiedenste Fragen geklärt werden:
  - Ist die Verteilung symmetrisch
  - Ist die Verteilung zentriert
  - Welche Werte liegen im 95% Konfidenzintervall

Nachteil:

- Wahl der Bandbreite willkürlich
  - Bandbreite zu groß oder zu klein führt zu einer schlechten Approximation der Dichte
- Histogramm ist lokal konstant und nicht stetig
  - Dichte ist meist weder lokal konstant noch stetig

Kerndichteschätzer

Kerndichteschätzer sind eine Generalisierung von Histogrammen und eine etwas bessere Methode zur Schätzung der Dichte. Ein *Kern* ist eine messbare Funktion $K : \mathbb{R} \to [0, \infty)$ deren Definiton die Gleiche ist, wie die Definition der Dichte:

- $K(x) \geq 0$ für alle $x \in \mathbb{R}$
- $\int_{\mathbb{R}} K(x)dx = 1$

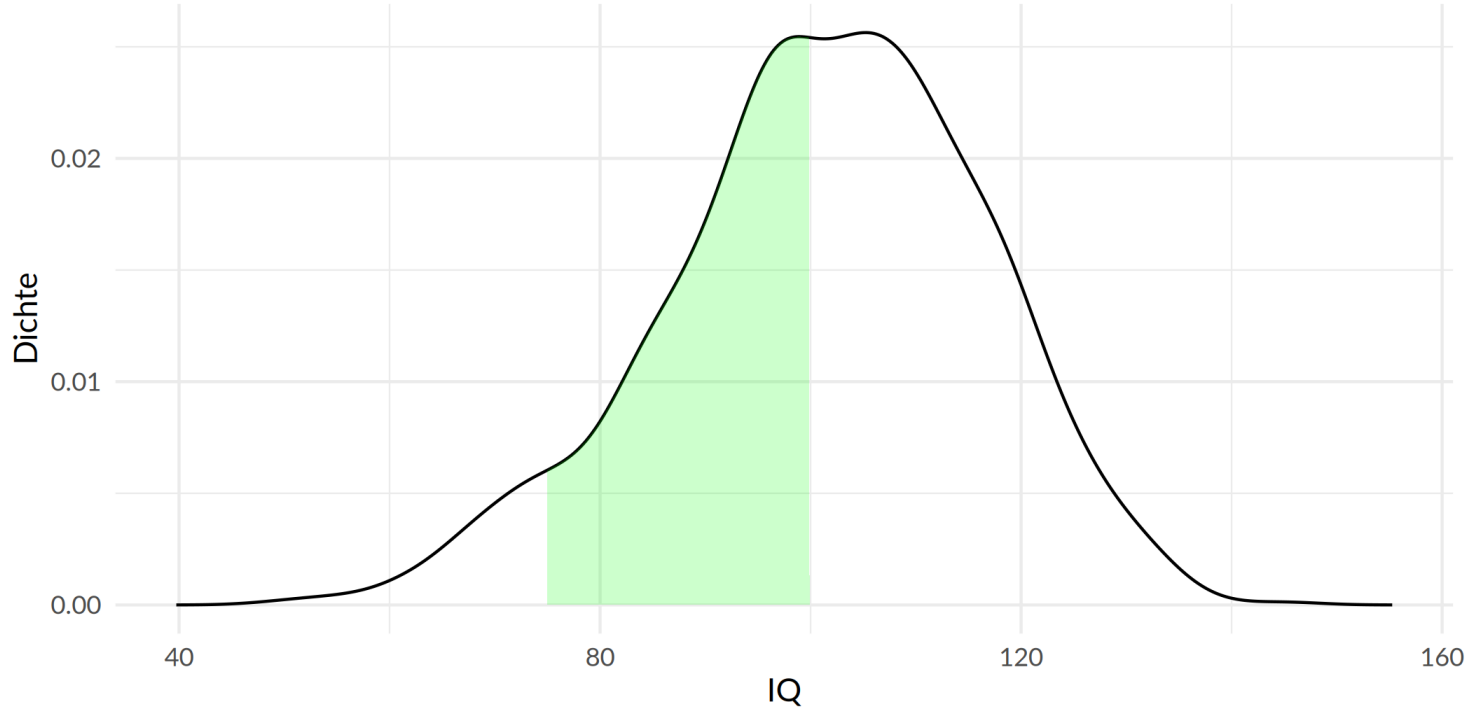
Interpretation der y-Achse

Ist nicht trivial bei einer Kerndichteschätzung

- Skaliert, so dass der Bereich unter der Kurve 1 ergibt
- Wir können den Anteil der Datenpunkte im Intervall $[a, b]$ berechnen indem wir die Fläche unter dem Intervall bestimmen

Interpretation der y-Achse

Beispiel: Anteil der Datenpunkte zwischen IQ 75 und 100.



Der Anteil dieser Fläche ist 0.4, d.h. 40% aller Personen in unserem Datensatz hat einen IQ zwischen 75 und 100.

Dies entspricht unserem Ergebnis mit der empirischen Verteilungsfunktion.

Die Normalverteilung

Die Normalverteilung

Histogramm und Kerndichte sind eine sehr schöne Möglichkeit empirische Verteilungen zu verdeutlichen.

> Nun gehen wir einen Schritt weiter und beschäftigen uns mit theoretischen Verteilungen, genauer: der **Normalverteilung**

Die Normalverteilung

Histogramm und Kerndichte sind eine sehr schöne Möglichkeit empirische Verteilungen zu verdeutlichen.

> Nun gehen wir einen Schritt weiter und beschäftigen uns mit theoretischen Verteilungen, genauer: der **Normalverteilung**

- Die Normalverteilung ist ein sehr wichtiges Konzept in der Mathematik
- Viele Verteilungen sind approximativ normal verteilt
  - Hierunter zählen: Körpergröße, Gewicht, Blutdruck, IQ-Werte, ...

In dieser Veranstaltung konzentrieren wir uns nicht darauf, warum dies so ist, sondern wie uns die Normalverteilung weiterhelfen bei der statistischen Analyse

Die Normalverteilung

Die mathematische Definition der Normalverteilung besagt, dass der Anteil im Interval (a, b) durch folgende Formel berechnet werden kann:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\} dx$$

Damit ist die Normalverteilung durch zwei Parameter definiert (μ und σ). Hier ist μ der Mittelwert und σ die Standardabweichung der Verteilung.

Die Normalverteilung ist

- symmetrisch
- zentriert um den Mittelwert
- 95% aller Werte liegen innerhalb von 2 Standardabweichungen vom Mittelwert

Die Normalverteilung

Wenn unser Datensatz nun durch die Normalverteilung approximiert werden kann, so bedeutet dies, dass wir auch unseren Datensatz mit Hilfe von Mittelwert und Standardabweichung darstellen können.

Wir können den Mittelwert μ unserer IQ-Verteilung einfach berechnen:

```
x <- wage2$IQ
mu <- sum(x) / length(x)
```

und die Standardabweichung ist definiert als

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

```
SD.t <- sqrt( sum( (x-mu)^2) / length(x) )
```

- ✚ Interpretation der Standardabweichung: Durchschnittliche Abweichung zwischen den Werten der Verteilung und deren Mittelwert

Die Normalverteilung

Mittelwert und Standardabweichung für die Verteilung der IQ Daten

- R berechnet die Standardabweichung für eine Stichprobe, d.h. SD.e wird durch n-1 geteilt.

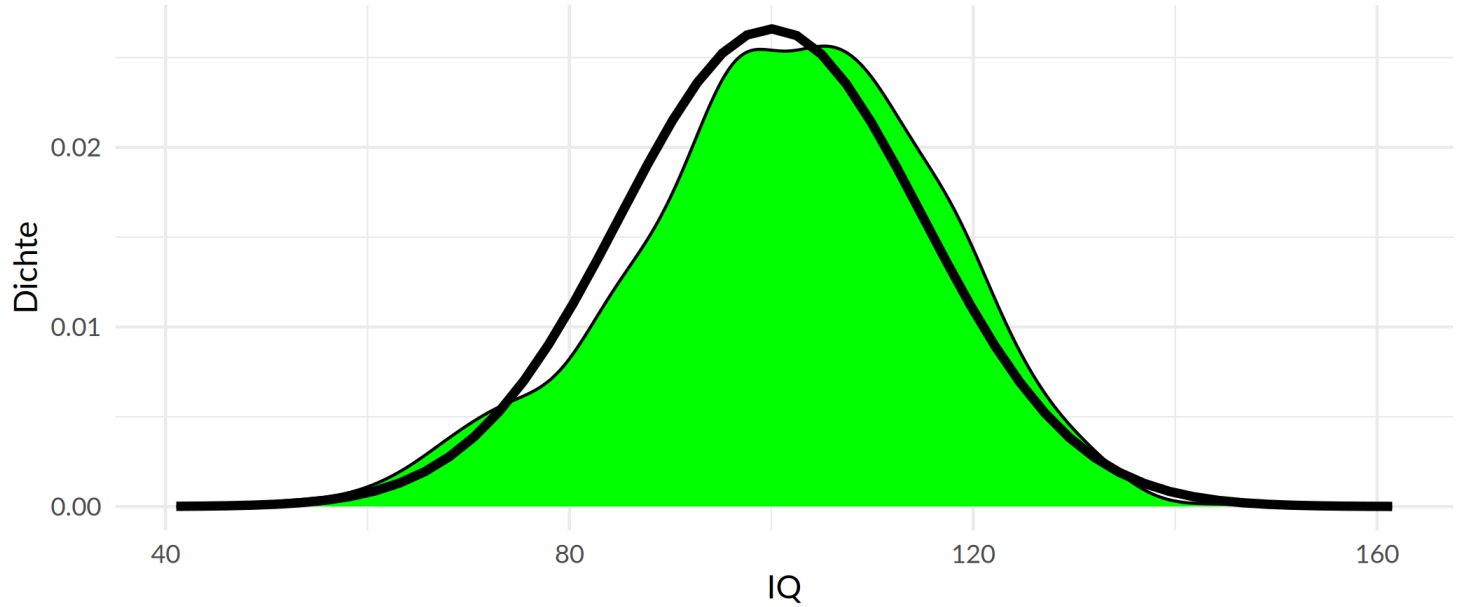
```
average <- mean(x, na.rm = T)
SD.e <- sd(x, na.rm = T)
c(average = average, SD = SD.e)
```

```
  average         SD
101.28235  15.05264
```

APPROXIMATION DURCH NORMALVERTEILUNG

Wir können nun den Kerndichteschätzer der empirischen IQ-Verteilung gegenüber einer Normalverteilung mit Mittelwert 100 und Standardabweichung 15 anschauen

- Verteilung leicht nach rechts versetzt
- Interpretation: Personen im Datensatz sind etwas intelligenter als der Durchschnitt



Wir können die empirischen IQ-Verteilung durch die Normalverteilung sehr gut approximieren!

Normalverteilung und stetige Wahrscheinlichkeiten

Bisher haben wir noch keine Wahrscheinlichkeiten im Kontext von Verteilungen eingeführt!

Hierzu können wir uns folgende Frage stellen:

> Wenn Sie eine Person zufällig aus ihrem Datensatz `wage2` ziehen, wie hoch ist die Chance, dass diese Person einen IQ größer 125 hat?

Normalverteilung und stetige Wahrscheinlichkeiten

Bisher haben wir noch keine Wahrscheinlichkeiten im Kontext von Verteilungen eingeführt!

Hierzu können wir uns folgende Frage stellen:

> Wenn Sie eine Person zufällig aus ihrem Datensatz `wage2` ziehen, wie hoch ist die Chance, dass diese Person einen IQ größer 125 hat?

Da jede Person die gleiche Wahrscheinlichkeit hat gezogen zu werden können wir diese Frage umschreiben zu:

> Wie hoch ist der Anteil an Personen mit einem IQ größer als 125?

Normalverteilung und stetige Wahrscheinlichkeiten

Bisher haben wir noch keine Wahrscheinlichkeiten im Kontext von Verteilungen eingeführt!

Hierzu können wir uns folgende Frage stellen:

> Wenn Sie eine Person zufällig aus ihrem Datensatz `wage2` ziehen, wie hoch ist die Chance, dass diese Person einen IQ größer 125 hat?

Da jede Person die gleiche Wahrscheinlichkeit hat gezogen zu werden können wir diese Frage umschreiben zu:

> Wie hoch ist der Anteil an Personen mit einem IQ größer als 125?

```
F <- function(a) mean(x<=a)
1 - F(125)
```

```
[1] 0.03850267
```

Stetige Wahrscheinlichkeiten

Wenn wir gewillt sind die Normalverteilungsannahme für unsere Daten, sagen wir für den IQ, zu akzeptieren, dann benötigen wir nicht unseren kompletten Datensatz um die Frage von gerade zu beantworten:

> Was ist die Wahrscheinlichkeit das eine zufällig gezogene Person einen IQ von mehr als 145 hat?

Stetige Wahrscheinlichkeiten

Wenn wir gewillt sind die Normalverteilungsannahme für unsere Daten, sagen wir für den IQ, zu akzeptieren, dann benötigen wir nicht unseren kompletten Datensatz um die Frage von gerade zu beantworten:

> Was ist die Wahrscheinlichkeit das eine zufällig gezogene Person einen IQ von mehr als 145 hat?

Nun benötigen wir nur noch den Mittelwert (`(\mu)`) und die Standardabweichung (`(\sigma)`) des IQ im Datensatz!

D.h. wir berechnen die Wahrscheinlichkeit, das ein bestimmtes Ereignis eintritt als:

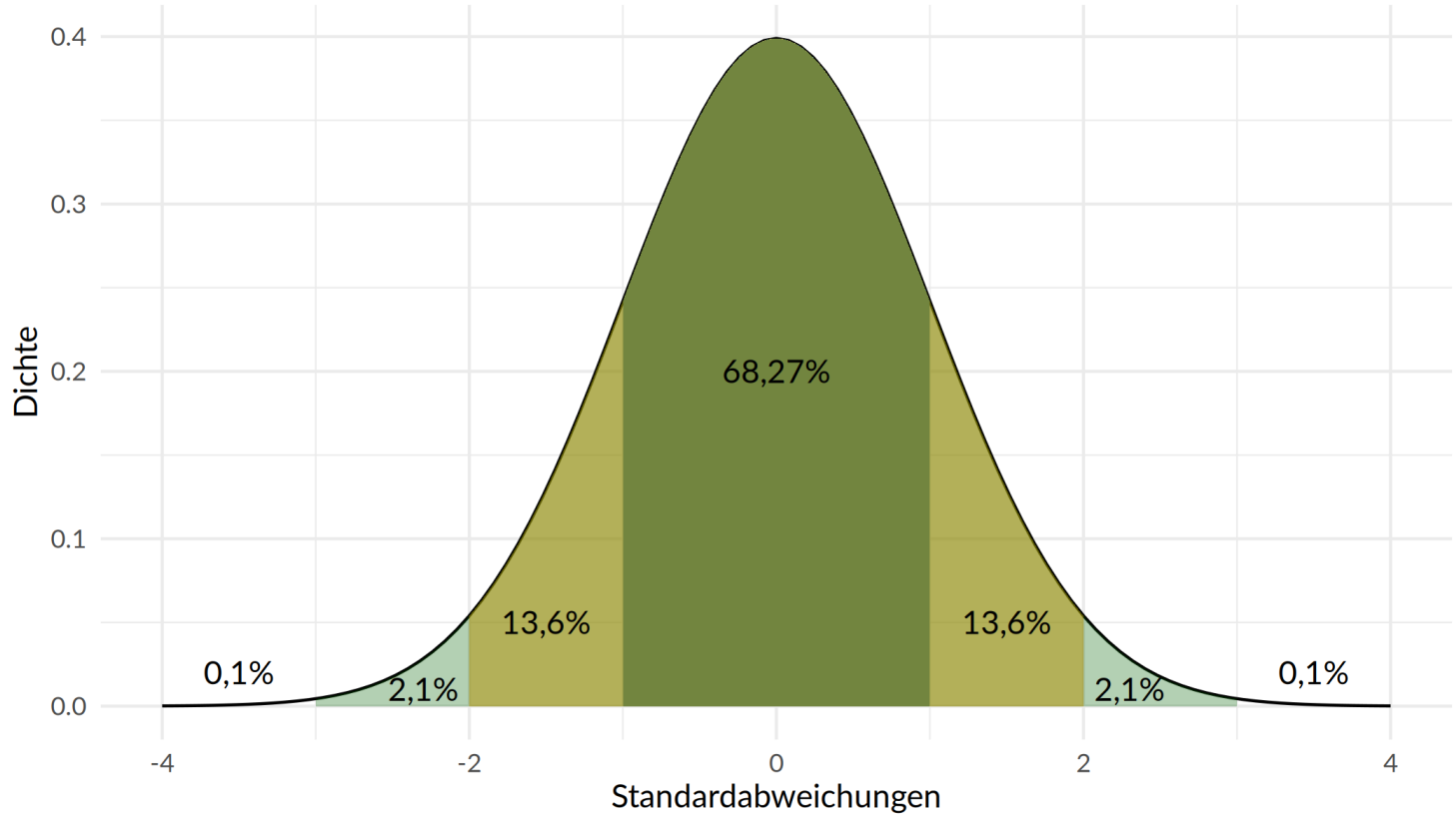
```
m <- mean(x) #Erinnern Sie sich noch was x ist?
s <- sd(x)
1 - pnorm(145, m, s)
```

```
[1] 0.001840269
```

```
#> [1] 0.001840269
```

> Die Normalverteilungsannahme ist eine vereinfachende Annahme, die uns in der empirischen Analyse viel erleichtert.

Dichte der Normalverteilung



Die Stichprobe und Stichprobenvarianz

Zufallsvariablen

In der empirischen Analyse arbeiten wir fast immer mit Daten die in irgendeiner Art durch den Zufall beeinflusst wurden:

- Eine zufällige Stichprobe
- Zufällige Messungenauigkeiten der einzelnen Variablen
- Daten kommen von einem zufälligen Event selbst
- ...

> Wir wollen diese Zufälligkeit quantifizieren und ihr Rechnung tragen.

Grundbegriffe

- **Grundgesamtheit**: Alle Individuen oder Beobachtungen die für uns interessant sind. Die Grundgesamtheit wird oft mit *N* abgekürzt.
- **Parameter in der Grundgesamtheit**: Parameter welchen wir gerne kennen würden, jedoch nicht kennen. Bspw. den *Mittelwert der Grundgesamtheit*, welchen wir mathematisch als μ deklarieren. Oder den Anteil roter Kugeln (*Anteil der Grundgesamtheit*), welches mathematisch p wäre.
- **Zensus**: Eine Zählung aller Individuen in unserer Grundgesamtheit
- **Stichprobe**: Untersuchung nur einer bestimmten Anzahl *n* von Individuen der Grundgesamtheit.
- **Punktschätzer**: Geschätzter Parameter auf Basis einer Stichprobe *n*. Mit dem Punktschätzer soll der unbekannte Parameter der Grundgesamtheit geschätzt werden. In unserem Beispiel der Anteil an roten Kugeln. Den Punktschätzer den wir hier erhalten würde mathematisch als $\hat{p}$ definiert um anzuzeigen, dass er auf Basis einer Stichprobe geschätzt wurde.
- **Repräsentative Stichprobe**: Eine Stichprobe ist repräsentativ, wenn diese der Grundgesamtheit sehr ähnlich sieht, d.h. wenn deren Charakteristika derer der Grundgesamtheit entsprechen

Grundbegriffe

- **Verallgemeinerbar**: Eine Stichprobe ist verallgemeinerbar, wenn Resultate aus der Stichprobe auch auf die Grundgesamtheit zutreffen. Ist $\hat{p}$ eine gute Abschätzung für p?
- **Stichprobenverzerrung**: Entsteht, wenn einige Individuen oder Beobachtungen in der Gesamtpopulation eine höhere Wahrscheinlichkeit haben in der Stichprobe repräsentiert zu sein. Eine Stichprobe ist *unverzerrt* wenn alle Individuen die gleiche Chance haben in die Stichprobe aufgenommen zu werden
- **Zufällige Stichprobe**: Wenn zufällig und nicht verzerrt aus der Grundgesamtheit gezogen wird

Grundbegriffe

Wenn ihre Stichprobe mit der Größe *n* zufällig gezogen wird, dann ist ihre Stichprobe

- *unverzerrt* und *repräsentativ* für ihre Grundgesamtheit *N*
- alle Resultate aus der Stichprobe sind *verallgemeinerbar* für die Grundgesamtheit
- die *Punktschätzer* sind eine gute Abschätzung des Parameters der Population

Somit müssen Sie keinen Zensus durchführen um Aussagen über die Grundgesamtheit machen zu können.

Eine Stichprobe

> Gegeben unsere Verteilung ist eine zufällige Stichprobe aus der Grundgesamtheit

Hier wollen wir wieder unser Urnenbeispiel heranziehen:

```
set.seed(123)
urne <- as.tibble(rep( c("rot", "weiß"), times = c(760,1240)))
urne <- urne |> mutate(id = rownames(urne))
colnames(urne) <- c("farbe", "id")

stichprobe <- sample_n(urne, 50)

stichprobe |> count(farbe)
```

```
# A tibble: 2 × 2
  farbe     n
  <chr> <int>
1 rot      17
2 weiß     33
```

Stichprobenvarianz

Wir betrachten hier Zufallsvariablen!

> Können wir mittels dieser einen Stichprobe etwas über die Stichprobenvarianz (Verteilung mehrerer Stichproben) aussagen?

Stichprobenvarianz

Wir betrachten hier Zufallsvariablen!

> Können wir mittels dieser einen Stichprobe etwas über die Stichprobenvarianz (Verteilung mehrerer Stichproben) aussagen?

```
library(infer)
stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot")
```

```
Response: farbe (factor)
# A tibble: 50 × 1
   farbe
   <fct>
 1 rot
 2 rot
 3 rot
 4 rot
 5 rot
 6 weiß
 7 weiß
 8 weiß
 9 weiß
10 weiß
# i 40 more rows
```

Stichprobenvarianz

```
bootstrap <- stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot") |>
  generate(reps = 48, type = "bootstrap")
```

Stichprobenvarianz

```
bootstrap <- stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot") |>
  generate(reps = 48, type = "bootstrap")
```

- **farbe** = welche Farbe der Ball hat
- **replicate** = aus welchem Zug der Ball stammt (insgesamt 48 Züge)

Hier ziehen wir mit zurücklegen 48 mal aus unserer Stichprobe und erhalten so eine Verteilung möglicher Stichproben!

Stichprobenvarianz

```
bootstrap
```

```
Response: farbe (factor)
# A tibble: 2,400 × 2
# Groups:   replicate [48]
   replicate farbe
       <int> <fct>
 1         1 weiß
 2         1 rot
 3         1 rot
 4         1 rot
 5         1 weiß
 6         1 weiß
 7         1 weiß
 8         1 rot
 9         1 weiß
10         1 weiß
# i 2,390 more rows
```

Bootstrap

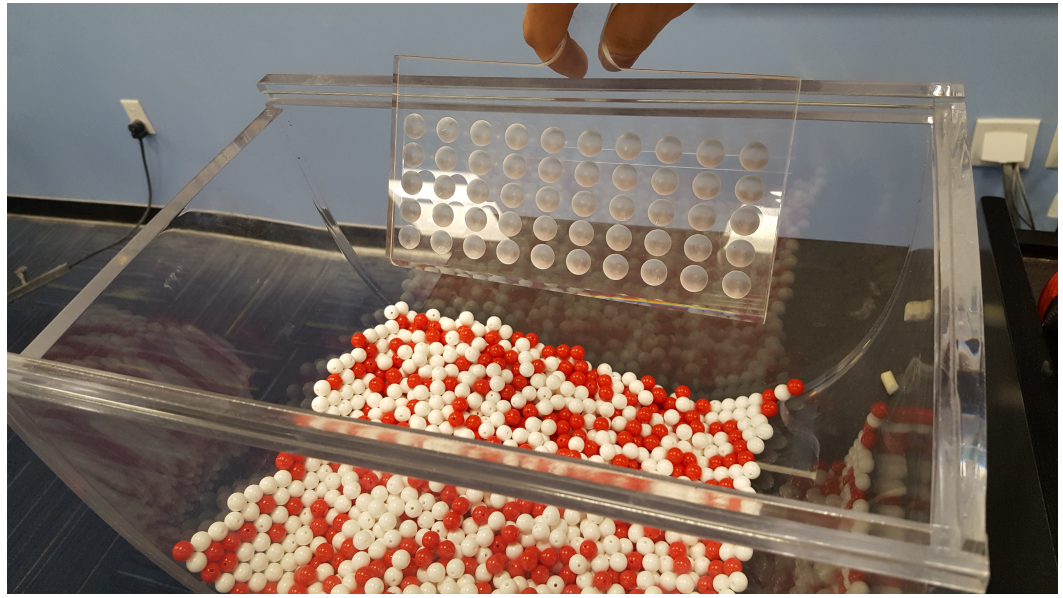
Ausgangslage:

- Eine Stichprobe von 17 roten und 33 weiße Kugeln
- Was Sie gerne hätten wäre die Grundgesamtheit:

Bootstrap

Ausgangslage:

- Eine Stichprobe von 17 roten und 33 weiße Kugeln
- Was Sie gerne hätten wäre die Grundgesamtheit:



Quelle: https://moderndive.com/7-sampling.html

Bootstrap

Können Sie diese Grundgesamtheit durch häufiges Ziehen mit Zurücklegen aus ihrer Stichprobe replizieren? Dieses ziehen mit Zurücklegen wird hier *bootstrap* genannt.

Bootstrap

Können Sie diese Grundgesamtheit durch häufiges Ziehen mit Zurücklegen aus ihrer Stichprobe replizieren? Dieses ziehen mit Zurücklegen wird hier *bootstrap* genannt.

Bootstrap bedeutet frei übersetzt "sich selbst an seinem Schopf aus dem Sumpf ziehen", was soviel heißt wie "auf Grund seiner eigenen Fähigkeiten Erfolg haben".

- Statistisch gesehen meint dies die Effekte der Stichprobenvarianz nur auf der Basis einer einzelnen Stichprobe herauszufinden
- Besser: Sie können mit dem Bootstrap approximativ eine Stichprobenverteilung konstruieren, nur auf Basis einer einzelnen Stichprobe

Bootstrap Verteilung

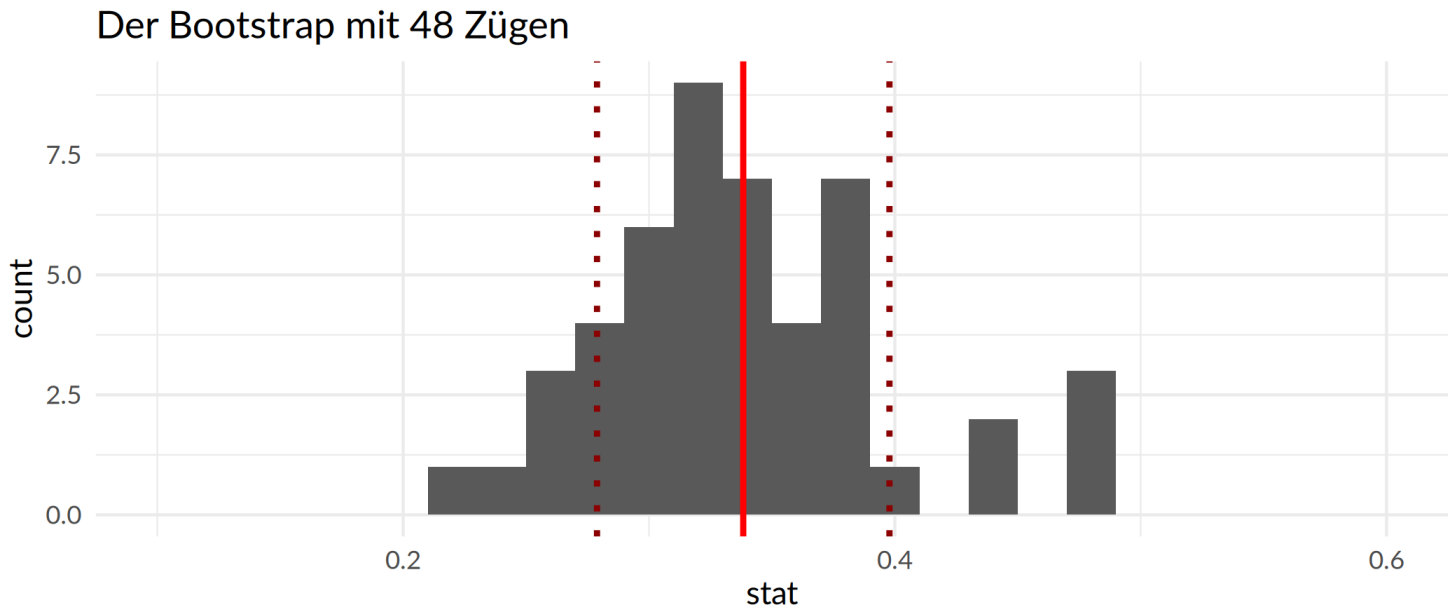
```
bootstrap1 <- stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot") |>
  generate(reps = 48, type = "bootstrap") |>
  calculate(stat = "prop")
```

> Können Sie aus ihren Daten ein Gefühl für die Stichprobenvarianz in der Gesamtpopulation erhalten?

Bootstrap Verteilung

```
bootstrap1 <- stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot") |>
  generate(reps = 48, type = "bootstrap") |>
  calculate(stat = "prop")
```

> Können Sie aus ihren Daten ein Gefühl für die Stichprobenvarianz in der Gesamtpopulation erhalten?



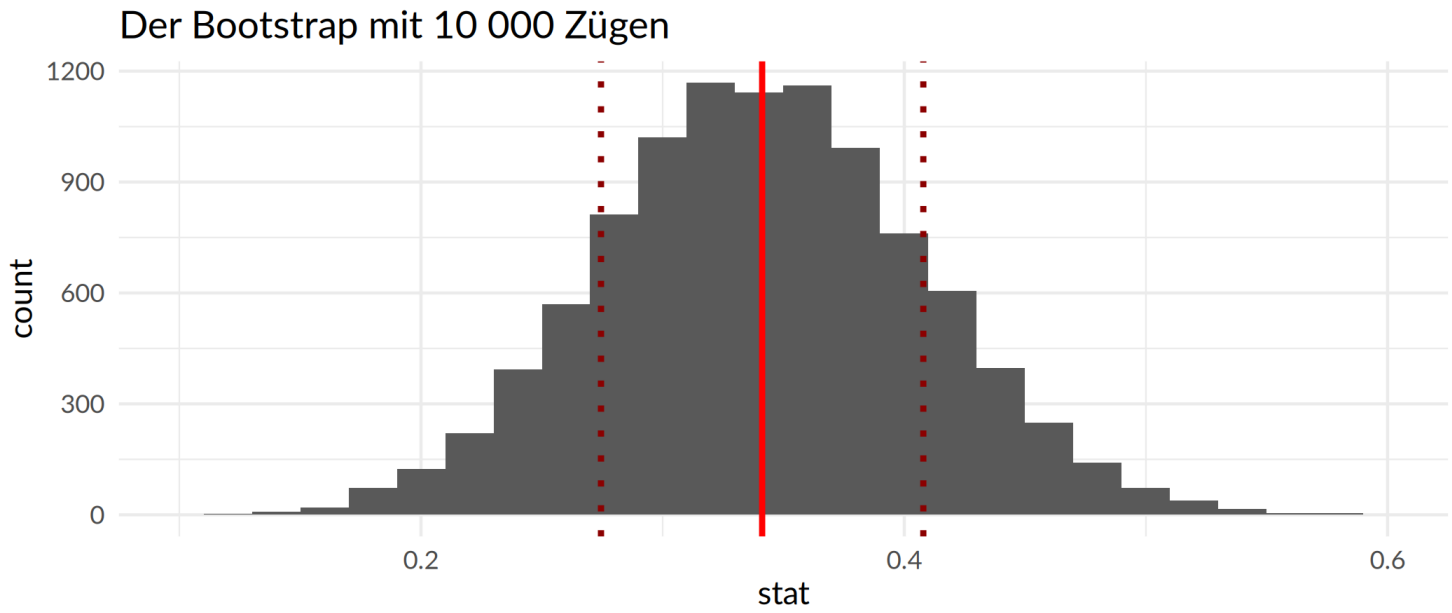
Bootstrap Verteilung (mit 10 000 Ziehungen)

> Wie sieht es aus, wenn Sie die Anzahl der Ziehungen erhöhen?

Bootstrap Verteilung (mit 10 000 Ziehungen)

> Wie sieht es aus, wenn Sie die Anzahl der Ziehungen erhöhen?

```
bootstrap2 <- stichprobe |>
  specify(formula = farbe ~ NULL, success = "rot") |>
  generate(reps = 10000, type = "bootstrap") |>
  calculate(stat = "prop")
```

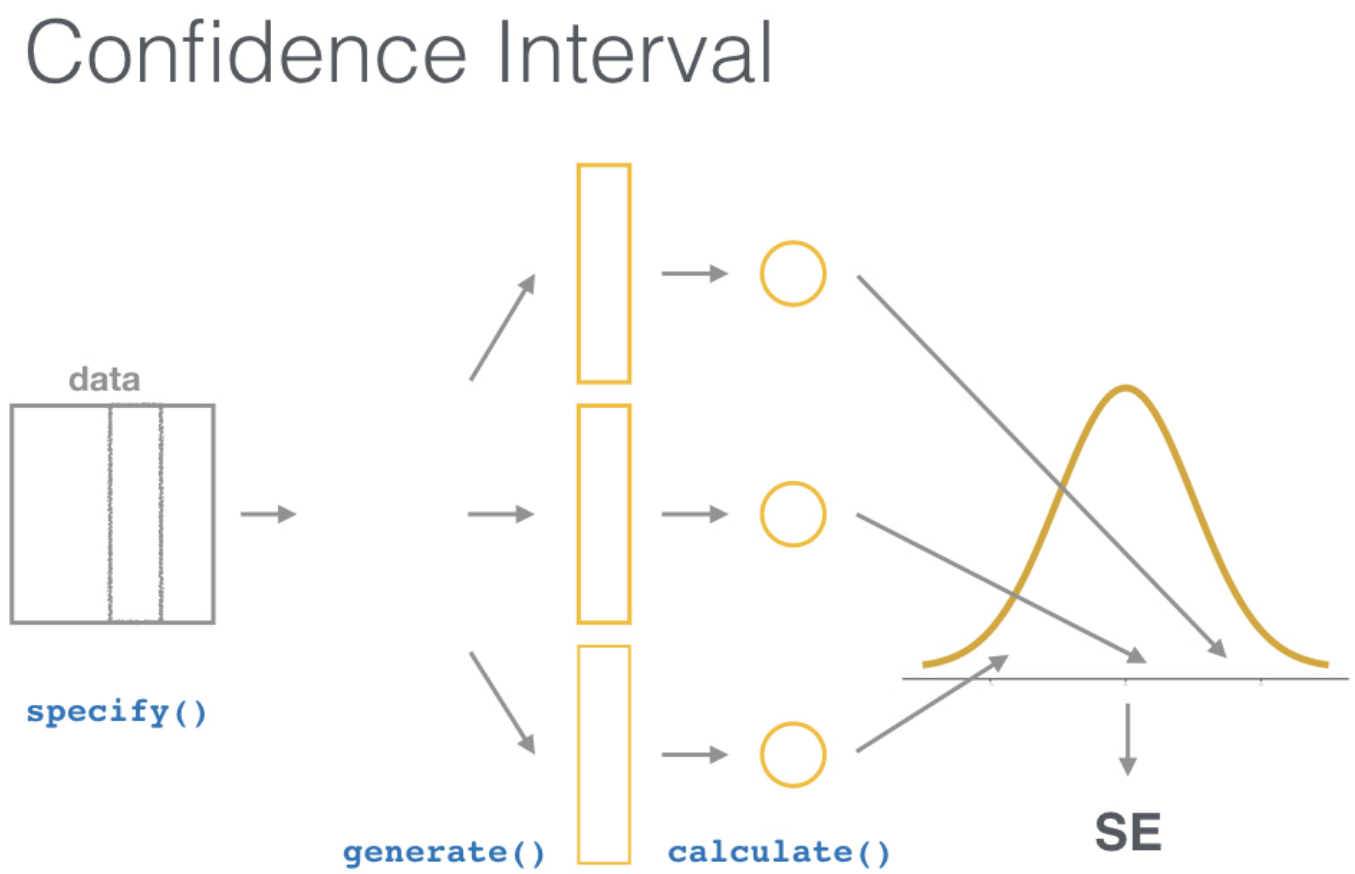


Verteilungen

Bootstrap Verteilung:

- Die Bootstrap Verteilung hat eine sehr ähnliche Form und Schwankungsbreite wie die Stichprobenverteilung. D.h. Bootstrapping kann dazu genutzt werden eine sehr genaue Abschätzung des Standardfehlers bzw. Konfidenzintervalls zu erhalten
- **Annahme: Unabhängigkeit der Ziehungen!**
  - Im vorherigen Urnen-Beispiel hatten wir unabhängige Ereignisse beschrieben. In diesem Fall war das Ziehen von mehreren Kugeln voneinander *unabhängig*, da der erste Zug den zweiten *nicht beeinflusst* (ziehen mit zurücklegen).

Konfidenzintervalle mit dem `infer` Paket



Konfidenzintervalle

Aus dem vorherigen Beispiel sehen Sie

- **Punktschätzer**: Genauer Wert aus einer Schätzung auf Basis der Stichprobe
- **Konfidenzintervall**: Bandbreite plausibler Werte auf Basis der Stichprobe
  - Das Konfidenzintervall ist eng verwandt mit der **Standardabweichung**

Konfidenzintervalle werden in der empirischen Forschung sehr häufig verwendent um die **Schätzunsicherheit** anzugeben. Dies gilt nicht nur für die Wirtschaftswissenschaften, sondern alle Bereiche der empirischen Forschung.

Konfidenzintervalle auf Basis der Normalverteilung

Eine Möglichkeit Konfidenzintervalle zu berechnen ist auf Basis einer Verteilungsannahme.

In der Regel wird die Normalverteilungsannahme getroffen.

- Die Normalverteilung ist ein sehr wichtiges Konzept in der Mathematik
- Viele Verteilungen sind approximativ normal verteilt
  - Hierunter zählen: Körpergröße, Gewicht, Blutdruck, IQ-Werte, ...

In dieser Veranstaltung konzentrieren wir uns nicht darauf, warum dies so ist, sondern wie Sie die Normalverteilung nützen können

Konfidenzintervalle auf Basis der Normalverteilung

Wenn unser Datensatz nun durch die Normalverteilung approximiert werden kann, so bedeutet dies, dass wir auch unseren Datensatz mit Hilfe von Mittelwert und Standardabweichung darstellen können.

Wie wir bereits wissen ist der Mittelwert definiert als:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

und die Standardabweichung kann definiert werden als

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

- Interpretation der Standardabweichung: Durchschnittliche Abweichung zwischen den Werten der Verteilung und deren Mittelwert

Konfidenzintervalle auf Basis der Normalverteilung

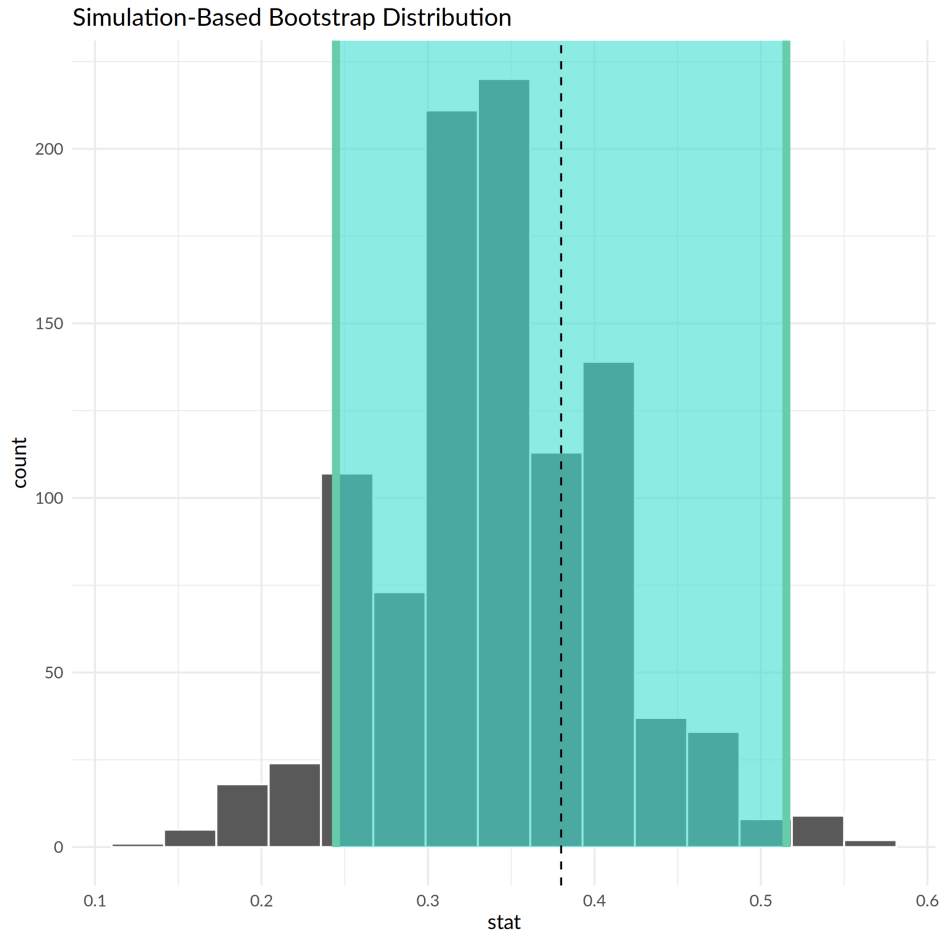
Wenn wir dieses Wissen auf unsere Stichprobe anwenden, dann können wir die Konfidenzintervalle entsprechend einzeichen (μ = 0.38):

```
conf <- stichprobe |>
  specify(response = farbe, success = "rot") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop")

standard_error_ci <- conf |>
  get_confidence_interval(type = "se", point_estimate = 0.38)

conf |> visualize() +
  shade_confidence_interval(endpoints = standard_error_ci) +
  geom_vline(xintercept = 0.38, linetype = "dashed")
```

KI auf Basis der Normalverteilung



Konfidenzintervalle

Bei einem 95% Konfidenzintervall:

- Zu 95% liegt der wahre Wert von x in unserem Intervall
- Das 95% Konfidenzintervall beinhaltet alle Werte, welche bis zu 2 Standardfehler von unserem geschätzten Mittelwert abweichen $[(\mu - 2\sigma), (\mu + 2\sigma)]$
- Die Intervallgrenzen sind Zufallsvariablen!

> Wird das Konfidenzintervall größer oder kleiner bei einem Konfidenzniveau von 99%?

Konfidenzintervalle

Um die Wahrscheinlichkeit zu bestimmen, dass x im Intervall liegt berechnen wir folgendes:

$$\Pr\left(\mu - 2\sigma \leq x \leq \mu + 2\sigma)\right)$$

Dies lässt sich zu folgender Formel umschreiben:

$$\Pr\left(-2 \leq \frac{x - \mu}{\sigma} \leq 2\right)$$

Konfidenzintervalle

Der mittlere Term ist hierbei approximativ normalverteilt mit einem Erwartungswert von 0 und einem Standardfehler von 1. Nennen wir diese Zufallsvariable Z:

$$\Pr\left(-2 \leq Z \leq 2\right)$$

Das heißt, wenn unser Konfidenzintervall 2 Standardfehler vom geschätzten Mittelwert umfasst, so beinhaltet unser Konfidenzintervall mit 95%-iger Wahrscheinlichkeit den wahren Wert x.

```
pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

Konfidenzintervalle

Der mittlere Term ist hierbei approximativ normalverteilt mit einem Erwartungswert von 0 und einem Standardfehler von 1. Nennen wir diese Zufallsvariable Z:

$$\Pr\left(-2 \leq Z \leq 2\right)$$

Das heißt, wenn unser Konfidenzintervall 2 Standardfehler vom geschätzten Mittelwert umfasst, so beinhaltet unser Konfidenzintervall mit 95%-iger Wahrscheinlichkeit den wahren Wert x.

```
pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

Für ein Konfidenzintervall von genau 95% reicht es einen etwas kleineren Bereich als 2σ anzuschauen:

```
qnorm(0.975)
```

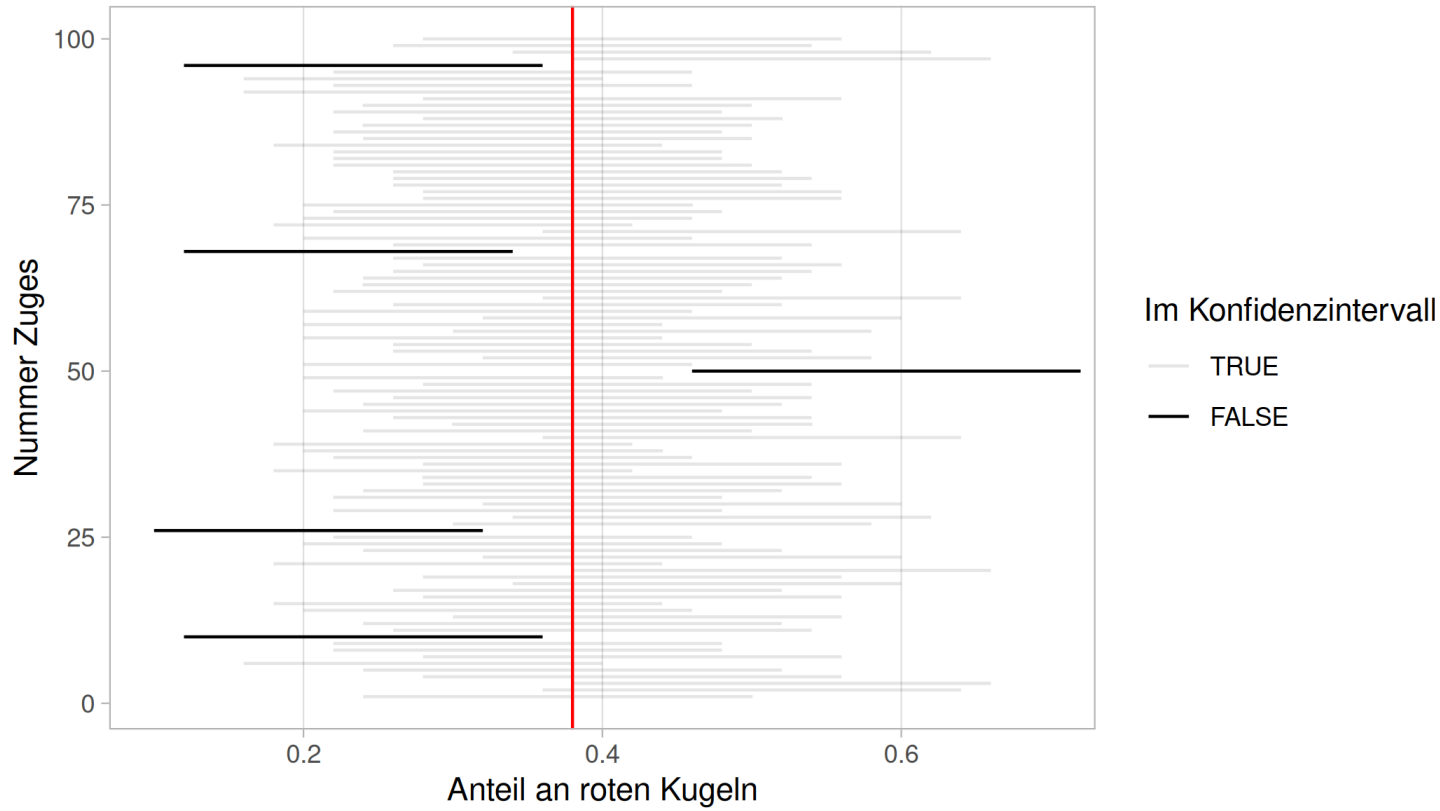
```
[1] 1.959964
```

Die korrekte Beschreibung

Bitte beachten Sie folgendes:

- Das Intervall welches Sie sich oben anschauen unterliegt zufälligen Schwankungen, nicht x.
- Es ist falsch zu sagen, dass x eine 95%-ige Wahrscheinlichkeit hat innerhalb des Intervalls zu liegen
- Die 95% beziehen sich auf die Wahrscheinlichkeit, dass dieses zufällige Konfidenzintervall auf x fällt, d.h. x beinhaltet
- Das Intervall schwankt um x nicht umgekehrt

Das 95% Konfidenzintervall



Grafik in Anlehnung an die Visualisierung aus Kapitel 8.5: Ismay, C., & Kim, A. Y. (2019). Statistical Inference via Data Science: A ModernDive into R and the Tidyverse.

Power

Wenn ein Konfidenzintervall die Null beinhaltet, so können wir die Nullhypothese das kein Effekt vorhanden ist *nicht* ablehnen.

Dies kann jedoch auf verschiedenen Eigenschaften der Stichprobe zurückzuführen sein:

- große Standardabweichung
- großes Konfidenzintervall, d.h. es wurde ein zu hohes Signifikanzniveau ausgewählt
- zu kleine Stichprobe. Durch eine größere Stichprobe wird der Standardfehler des Koeffizienten kleiner: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$