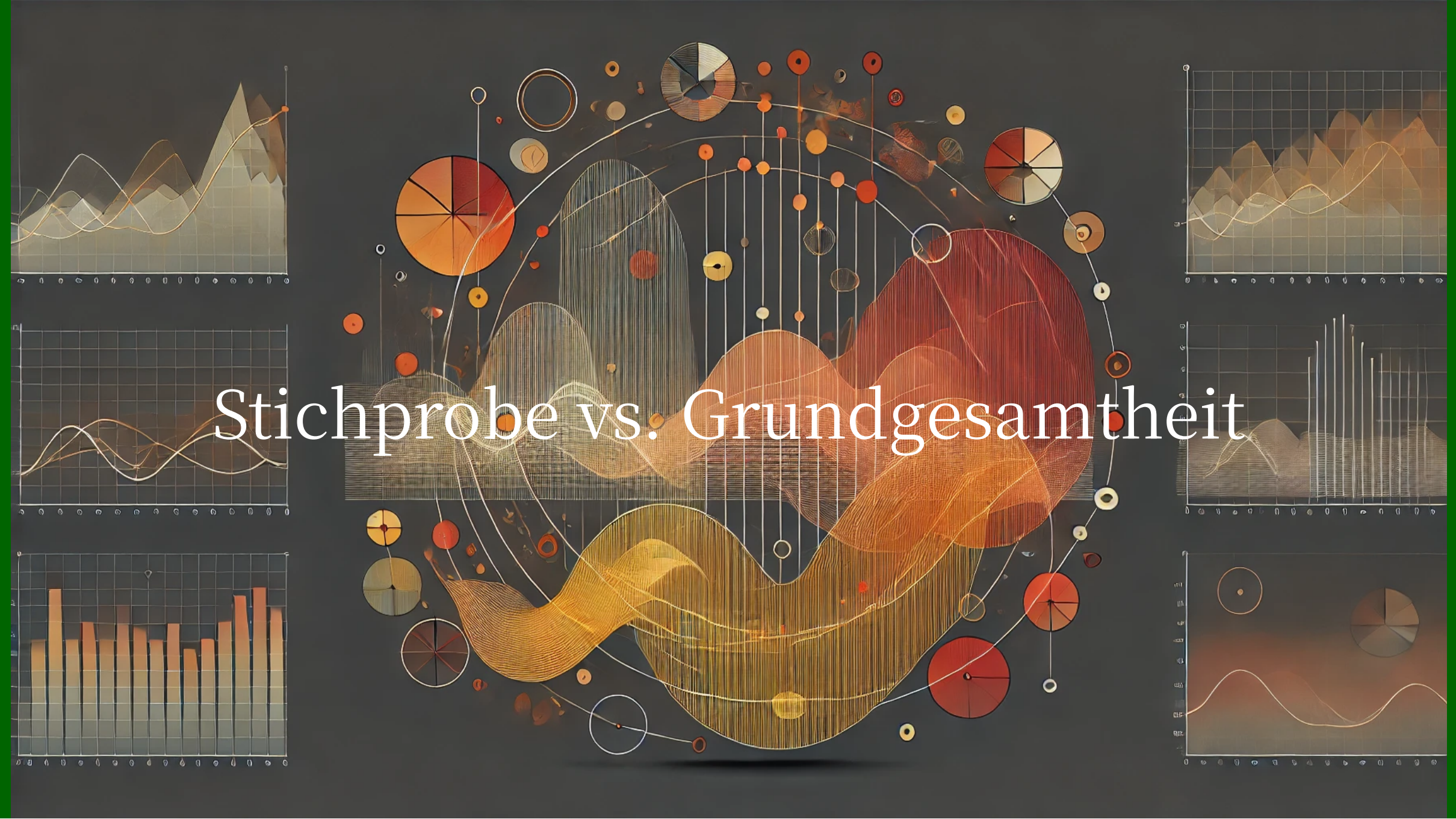
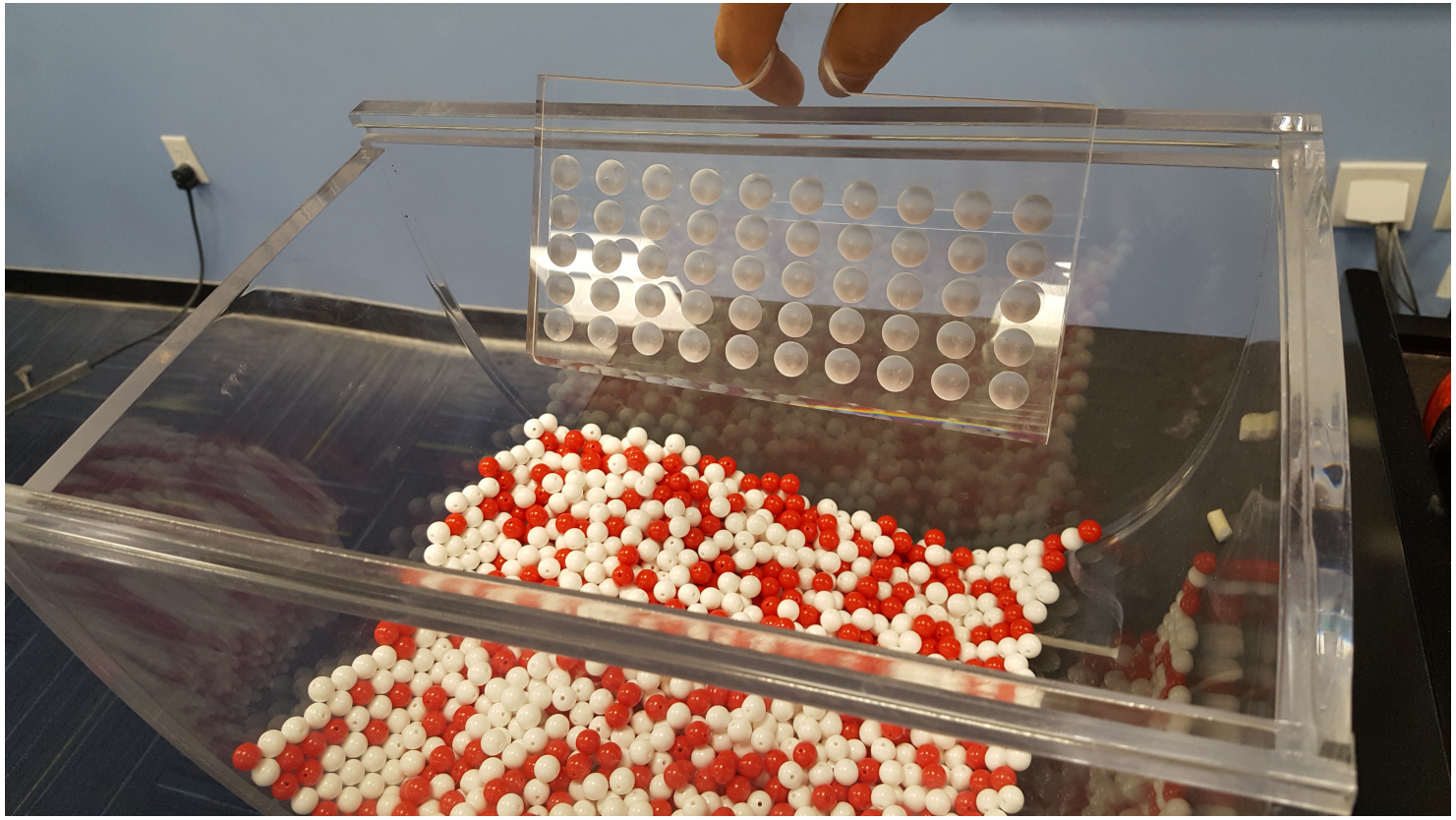

Stichprobe vs. Grundgesamtheit

Die Stichprobe

> **Schätzen Sie:** wie viele rote Kugeln sind in dieser Urne



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

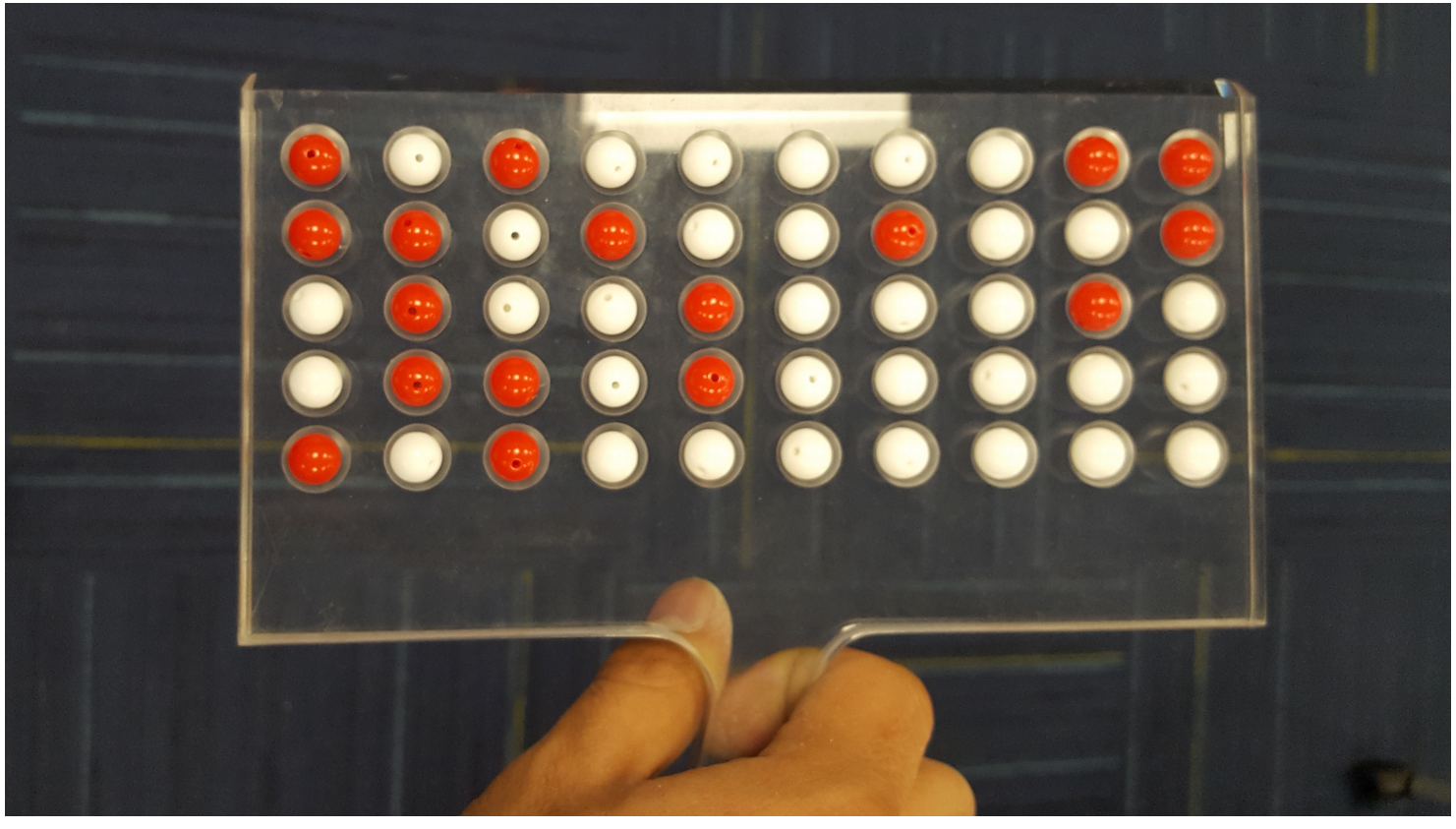
Die Stichprobe

Gibt es eine Möglichkeit auf die Anzahl der Kugeln zu kommen **ohne** alle Kugeln zu zählen?

Die Stichprobe

Gibt es eine Möglichkeit auf die Anzahl der Kugeln zu kommen **ohne** alle Kugeln zu zählen?

> Ja! Nehmen Sie eine Stichprobe aus der Urne (hier 50 Kugeln).



Die Stichprobe

Lassen Sie uns hier die Urne virtuell nachbauen:

- **Design:** 38% rote, 62% weiße Kugeln

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

Die Stichprobe

Lassen Sie uns hier die Urne virtuell nachbauen:

- **Design:** 38% rote, 62% weiße Kugeln

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

Nun können Sie eine Stichprobe von 50 Bällen entnehmen:

```
set.seed(1234)
probe1 <- urne |>
  sample_n(size = 50)
```

Wie viele sind rot?

```
probe1 |>
  summarize(anteil_rot = mean(farbe=="rot")) |>
  pull()
```

```
[1] 0.32
```

Wie viele sind rot?

```
probe1 |>
  summarize(anteil_rot = mean(farbe=="rot")) |>
  pull()
```

```
[1] 0.32
```

> Ist dies der tatsächliche Anteil an roten Kugeln in der Urne?

Stichprobenvarianz

Wie sieht es aus, wenn Sie mehrere Stichproben aus der Urne entnehmen, sagen wir **50**?

Dies können wir mit Hilfe des `infer`-Pakets und der Funktion `rep_sample_n` simulieren:

```
library(infer)

N <- 50
proben50 <- urne |>
  rep_sample_n(size=50, reps = N)
```

Mit welcher Wahrscheinlichkeit erhalten Sie nun eine rote Kugel?

Stichprobenvarianz

Wie sieht es aus, wenn Sie mehrere Stichproben aus der Urne entnehmen, sagen wir **50**?

Dies können wir mit Hilfe des `infer`-Pakets und der Funktion `rep_sample_n` simulieren:

```
library(infer)

N <- 50
proben50 <- urne |>
  rep_sample_n(size=50, reps = N)
```

Mit welcher Wahrscheinlichkeit erhalten Sie nun eine rote Kugel?

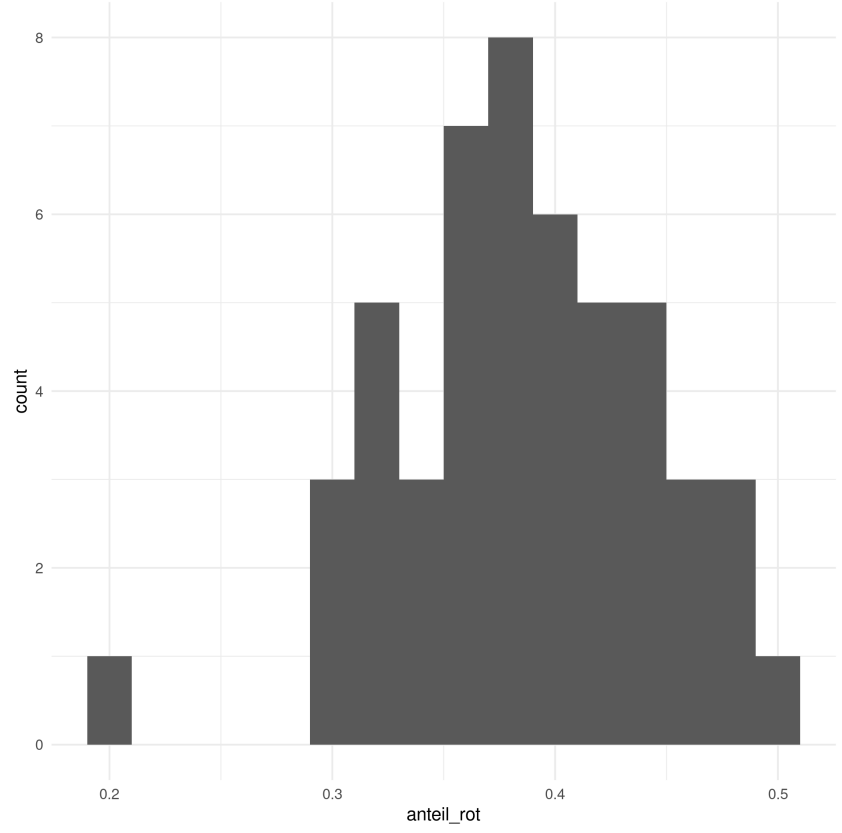
```
urne_anteil50 <- proben50 |>
  group_by(replicate) |>
  summarize(anteil_rot = mean(farbe == "rot"))

mean(urne_anteil50$anteil_rot)
```

```
[1] 0.386
```

Plotten der Stichprobenverteilung

```
urne_anteil50 |>
  ggplot(aes(x = anteil_rot)) +
  geom_histogram(binwidth = 0.02, farbe = "rot")
```



Stichprobenvarianz

- Manchmal ist der Anteil an roten Kugeln kleiner als 30%
- Manchmal ist der Anteil an roten Kugeln größer als 45%

Jedoch: Der häufigste Anteil liegt jedoch zwischen 35% und 45%.

Diese Unterschiede in der Anzahl an roten Kugeln erhalten wir auf Grund der *Stichprobenvarianz*

Stichprobenvarianz

Was passiert, wenn Sie die Anzahl an Stichproben erhöhen?

> Ziehen Sie 1000, 5000 und 10000 mal aus der Urne

Stichprobenvarianz

Was passiert, wenn Sie die Anzahl an Stichproben erhöhen?

> Ziehen Sie 1000, 5000 und 10000 mal aus der Urne

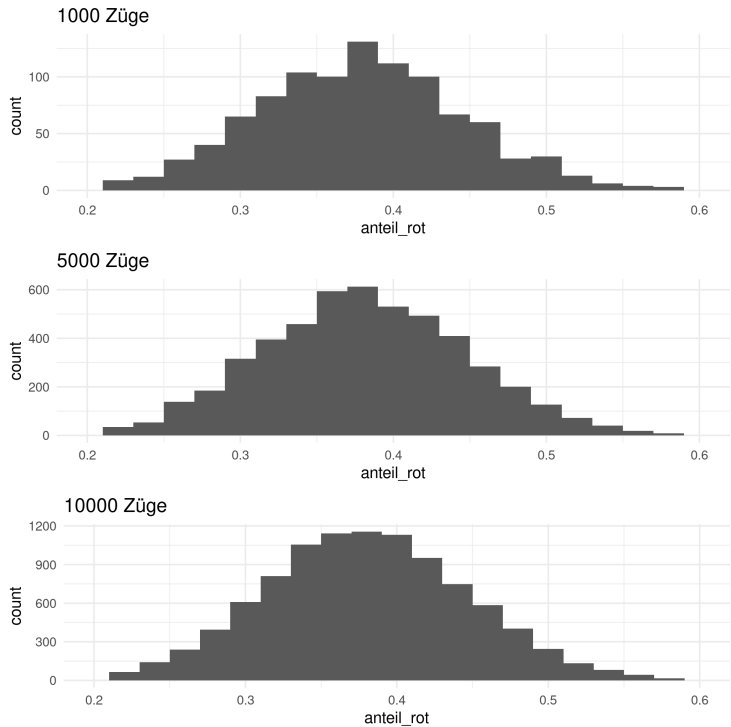
Wie groß ist hier der Anteil an roten Kugeln?

Stichprobenvarianz

Was passiert, wenn Sie die Anzahl an Stichproben erhöhen?

> Ziehen Sie 1000, 5000 und 10000 mal aus der Urne

Wie groß ist hier der Anteil an roten Kugeln?



Unterschiedliche Stichprobengrößen

Sie können auch die Stichprobengröße variieren.

Unterschiedliche Stichprobengrößen

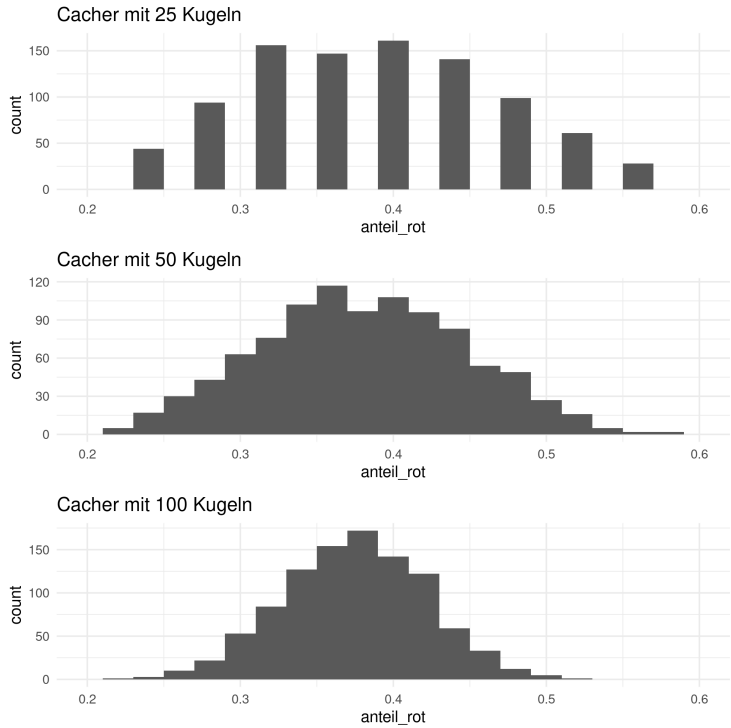
Sie können auch die Stichprobengröße variieren.

- Nutzen Sie einen Cacher mit 25, 50 und 100 Kugeln
- Anzahl der Entnahmen konstant bei 1000

Unterschiedliche Stichprobengrößen

Sie können auch die Stichprobengröße variieren.

- Nutzen Sie einen Cacher mit 25, 50 und 100 Kugeln
- Anzahl der Entnahmen konstant bei 1000



Stichprobe

- Je größer die Stichprobe, desto geringer die Varianz zwischen den einzelnen Stichproben
- Verteilung ist zentrierter um den gleichen Wert
- Alle Verteilungen zentriert um etwa 40%

Sie können die die Variation in ihren Daten mittels der Standardabweichung quantifizieren:

Stichprobe

- Je größer die Stichprobe, desto geringer die Varianz zwischen den einzelnen Stichproben
- Verteilung ist zentrierter um den gleichen Wert
- Alle Verteilungen zentriert um etwa 40%

Sie können die die Variation in ihren Daten mittels der Standardabweichung quantifizieren:

Stichprobengröße	
Anzahl der Kugeln	Standardabweichung
25	0.0958206
50	0.0698487
100	0.0466991

Stichprobe

- Stichproben als Grundlage für Schätzungen hilfreich
- Stichproben werden häufig genutzt da Grundgesamtheit nicht verfügbar oder zu umfangreich

Sie sollten bzgl. Stichproben zwei Grundkonzepte verinnerlichen:

- Welchen Effekt hat die Stichprobenvariation auf ihre Schätzer
- Welchen Effekt hat die Stichprobengröße auf ihre Stichprobenvariation