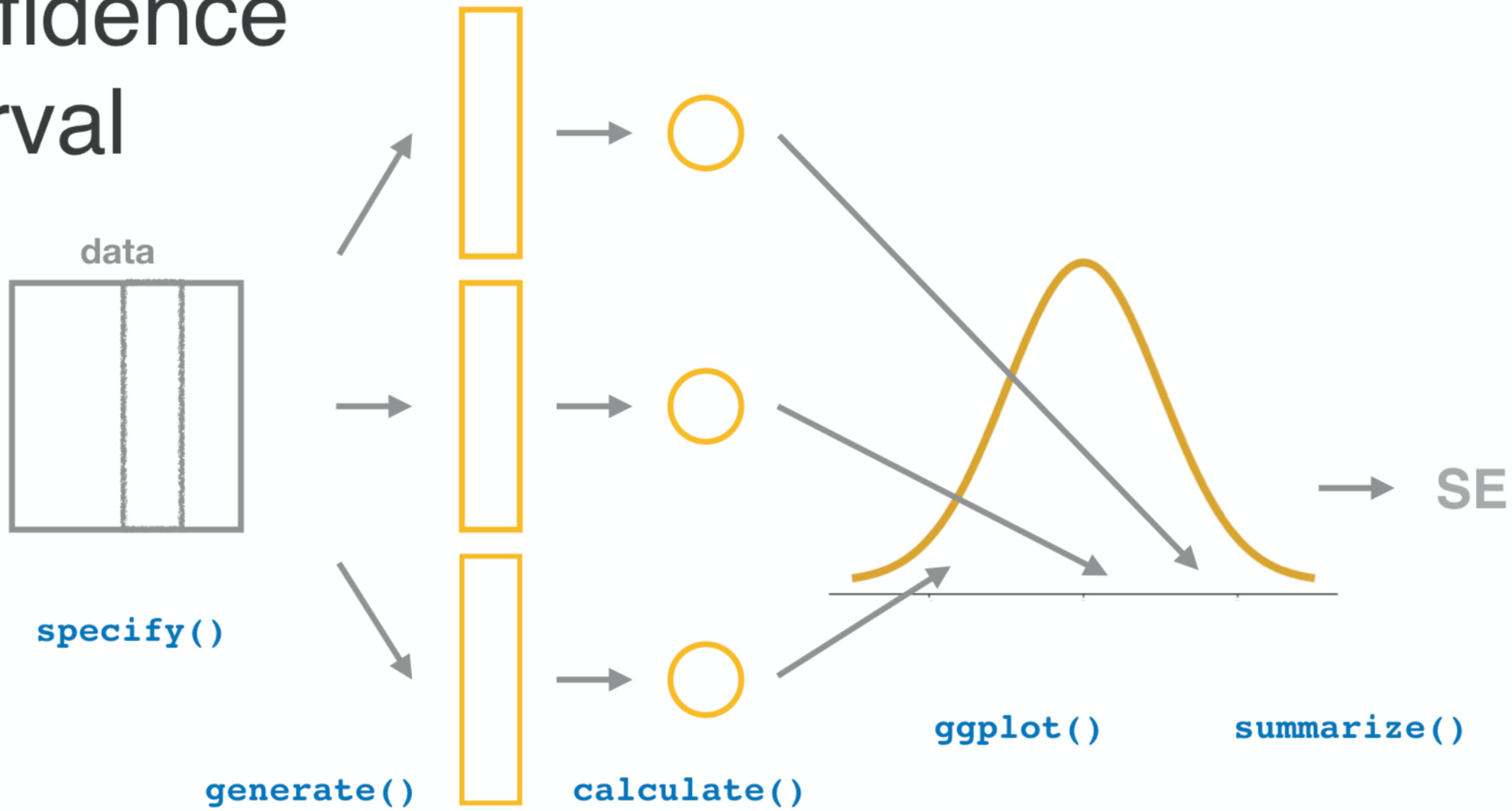
Hypothesis test for a proportion

INFERENCE FOR CATEGORICAL DATA IN R

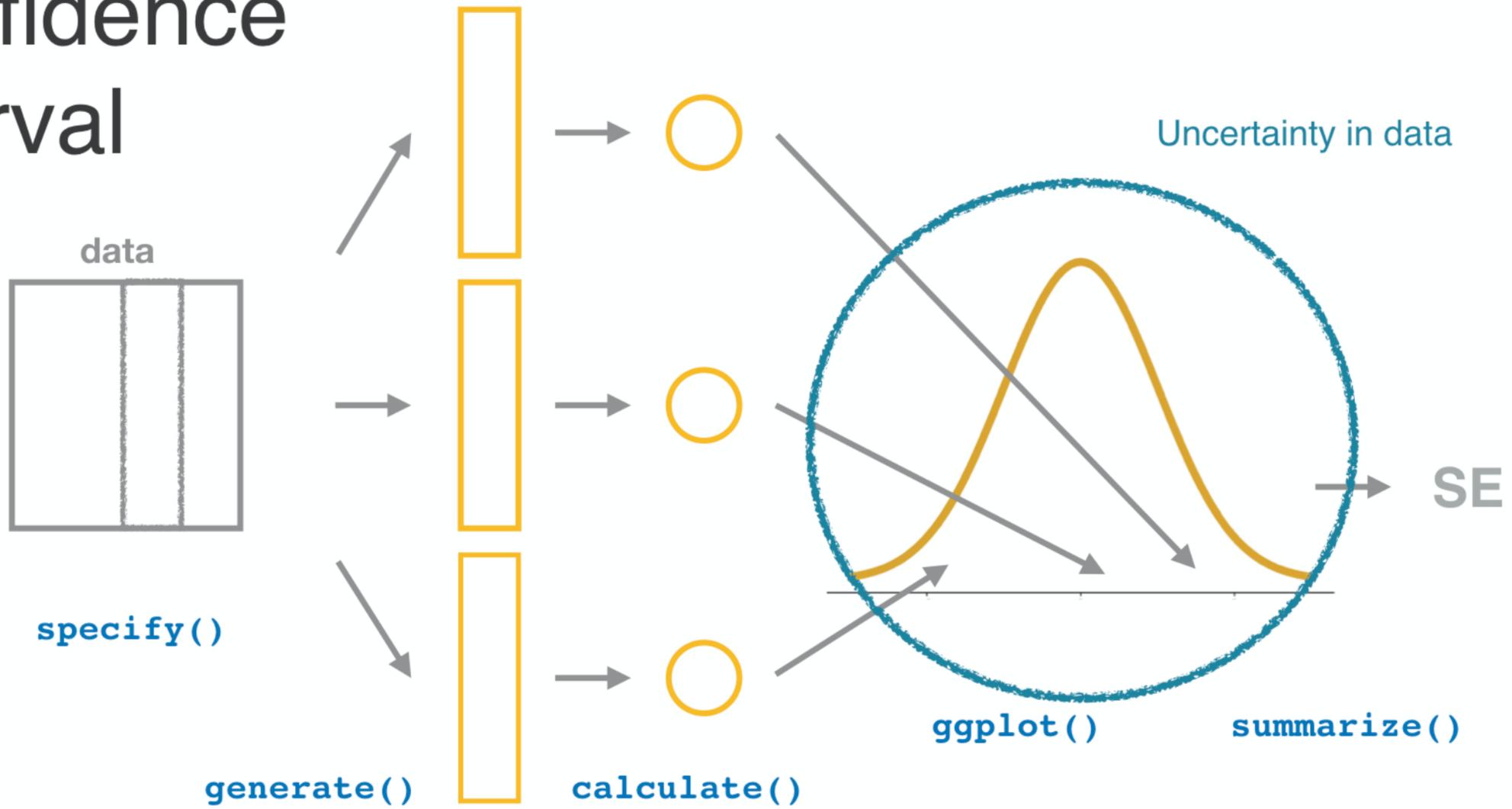
Andrew Bray

Assistant Professor of Statistics at Reed College

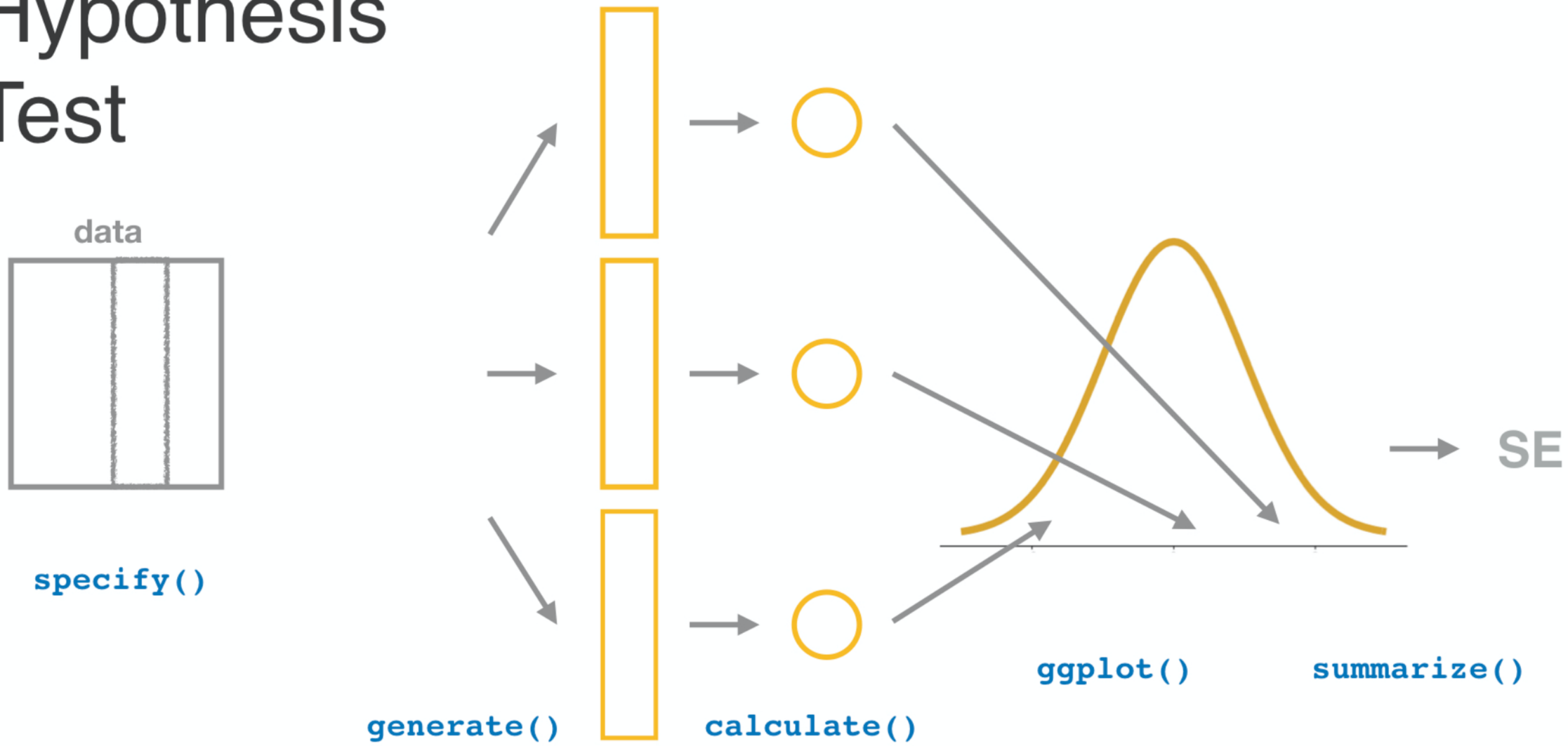
Confidence Interval



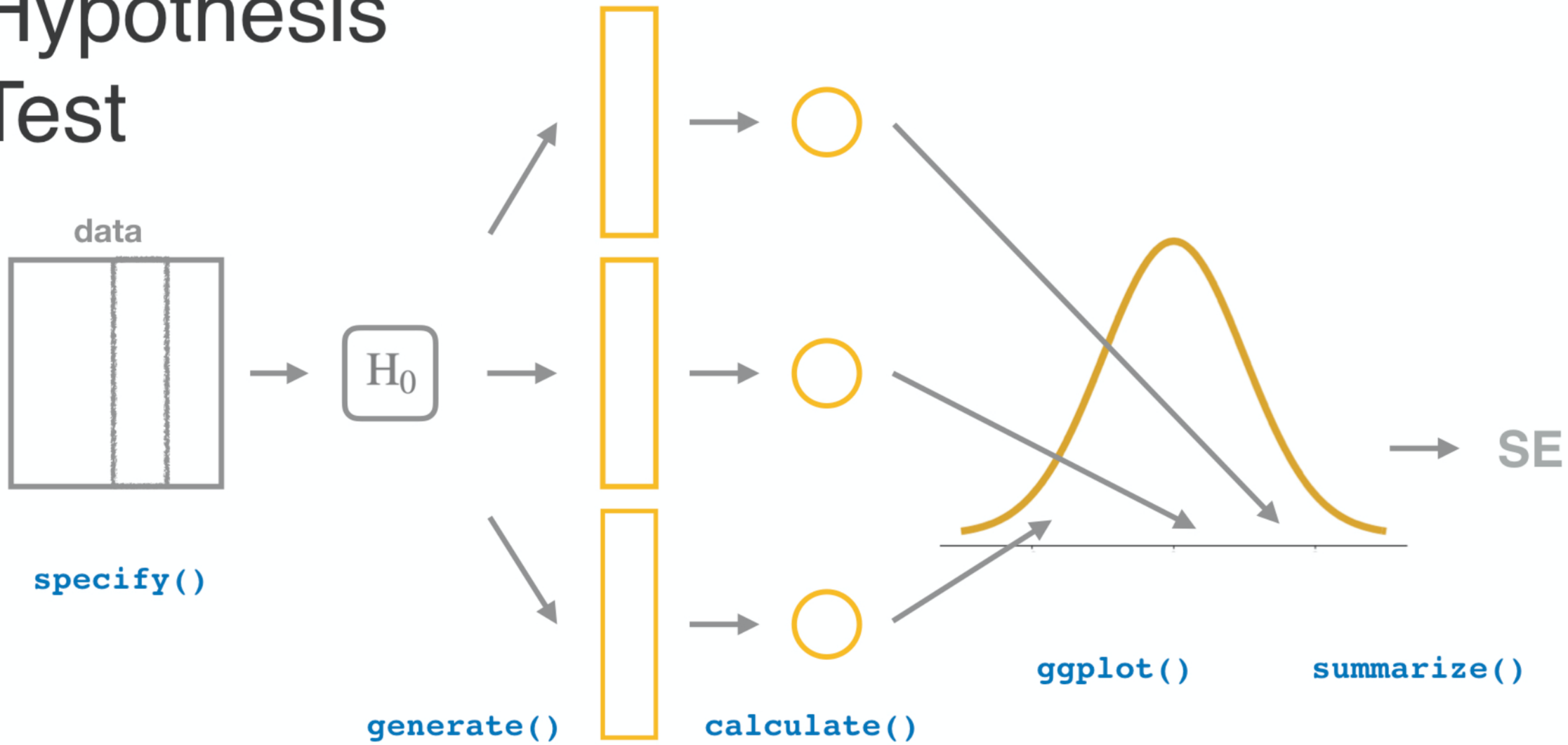
Confidence Interval



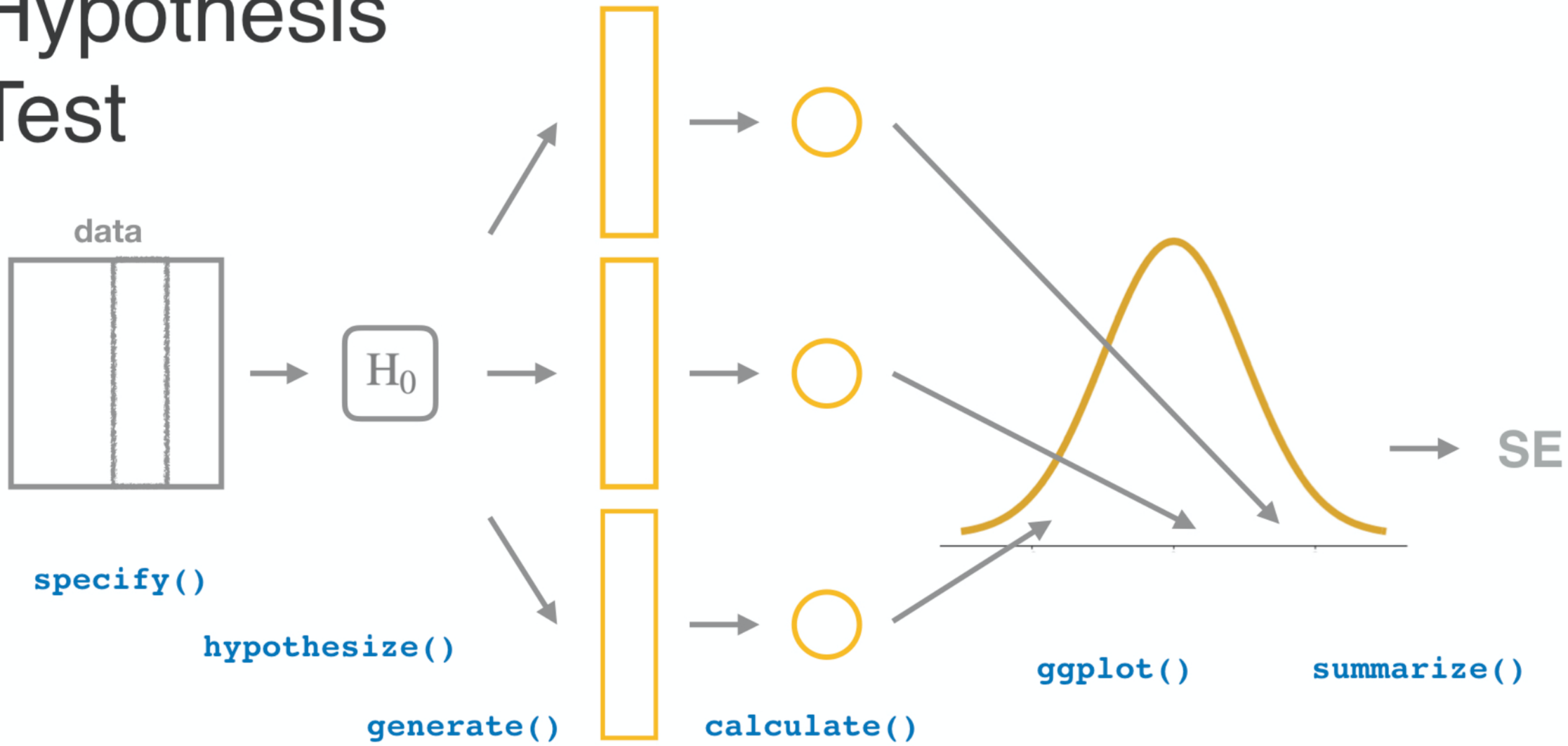
Hypothesis Test



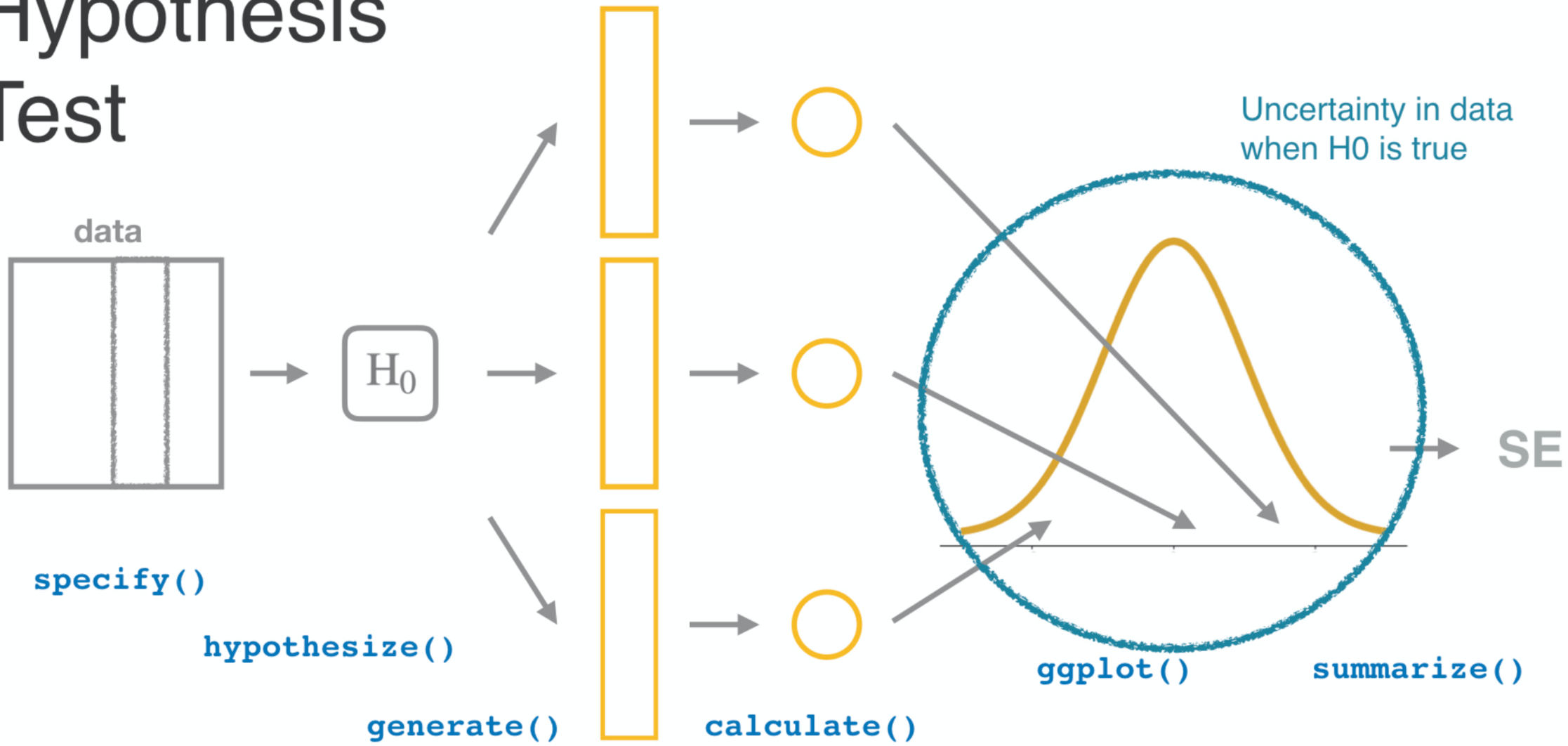
Hypothesis Test



Hypothesis Test



Hypothesis Test

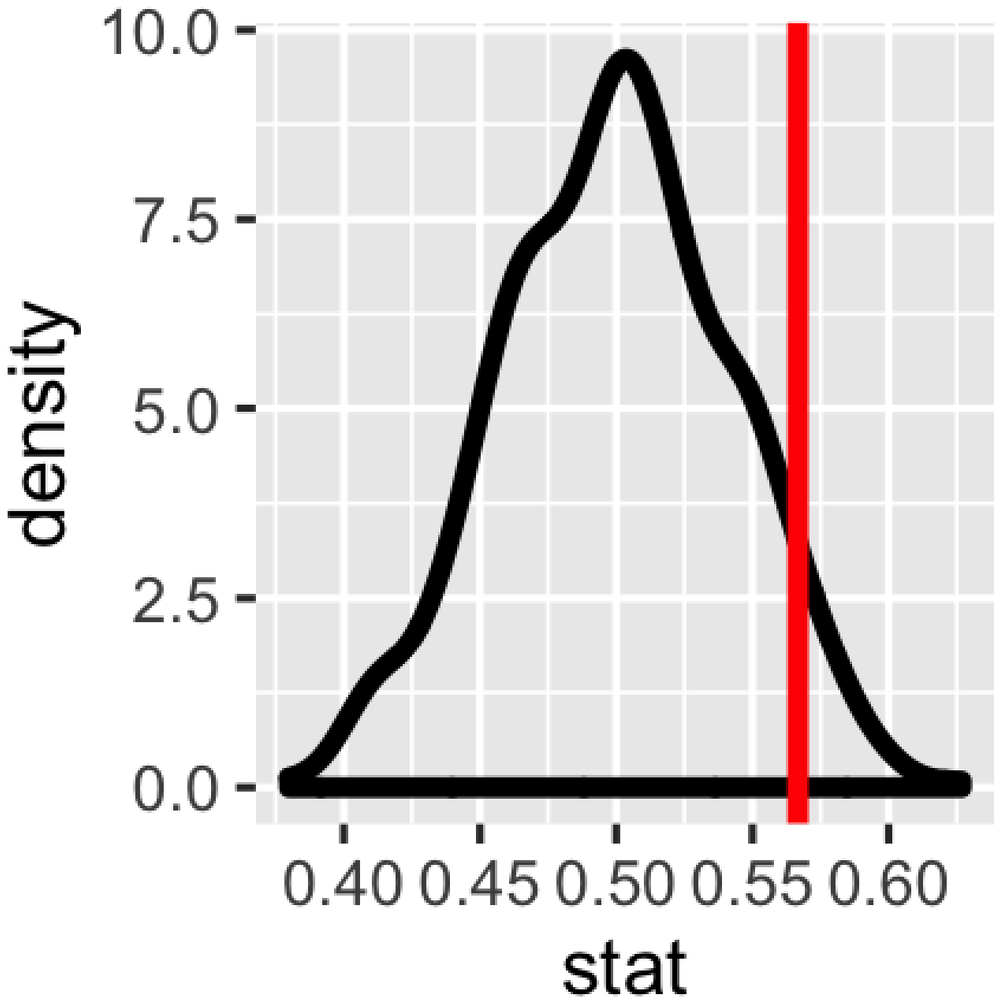


Do half of Americans favor capital punishment?

```
gss2016 %>%
  ggplot(aes(x = cappun)) +
  geom_bar()
```

```
p_hat <- gss2016 %>%
  summarize(mean(cappun == "FAVOR")) %>%
  pull()
p_hat
```

```
0.5666667
```



Do half of Americans favor capital punishment?

```
null <- gss2016 %>%
  specify(
    response = cappun,
    success = "FAVOR"
  ) %>%
```

```
  hypothesize(
    null = "point",
    p = 0.5
  ) %>%
```

```
  generate(
    reps = 500,
    type = "simulate"
  ) %>%
```

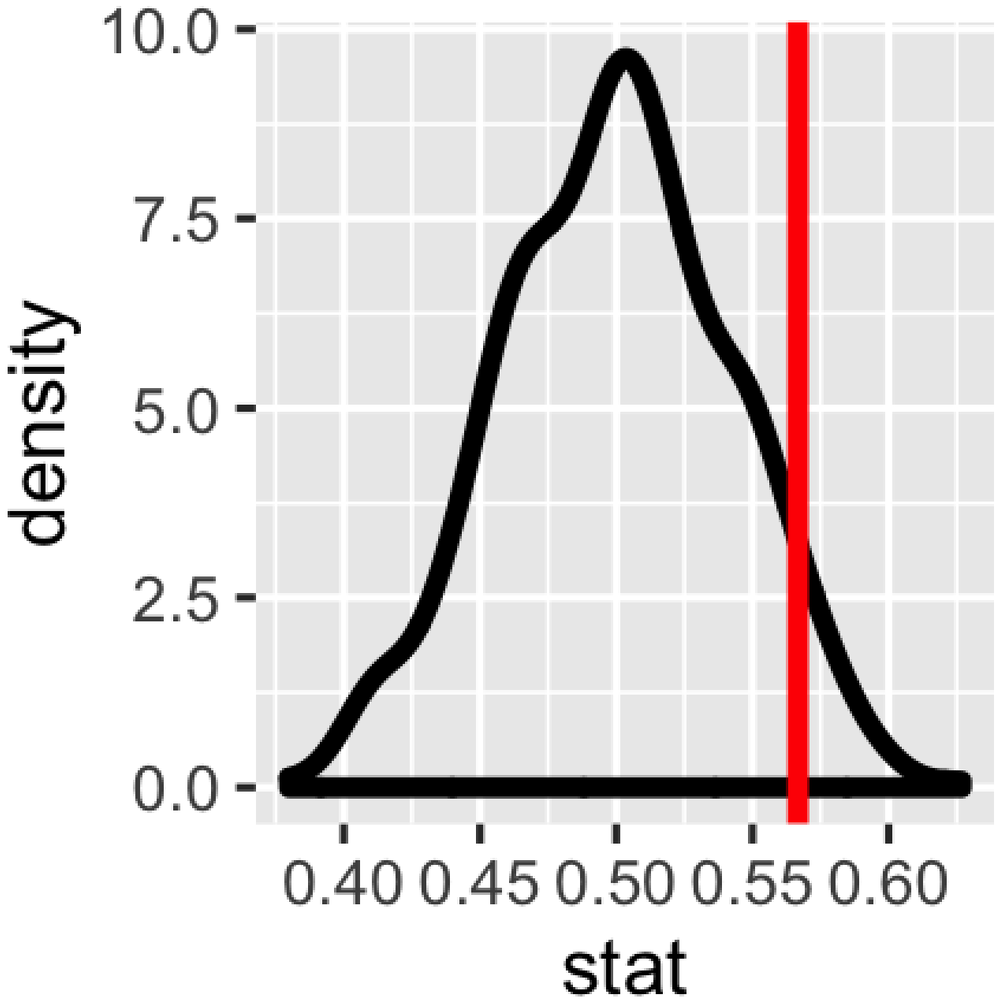
```
  calculate(stat = "prop")
```

```
A tibble: 500 x 2
   replicate  stat
   <fct>     <dbl>
 1 1         0.48
 2 2         0.447
 3 3         0.48
 4 4         0.44
 5 5         0.407
 6 6         0.52
 7 7         0.413
 8 8         0.553
 9 9         0.52
10 10        0.467
# … with 490 more rows
```

Do half of Americans favor capital punishment?

```
ggplot(null, aes(x = stat)) +
  geom_density() +
  geom_vline(
    xintercept = p_hat,
    color = "red"
  )
```

```
null %>%
  summarize(mean(stat > p_hat)) %>%
  pull() * 2
```



Hypothesis test

- Null hypothesis: theory about the state of the world.
- Null distribution: distribution of test statistics assuming null is true.
- p-value: a measure of consistency between null hypothesis and your observations.
  - high p-value: consistent (p-val > alpha)
  - low p-value: inconsistent (p-val < alpha)

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

Intervals for differences

INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

A question in two variables

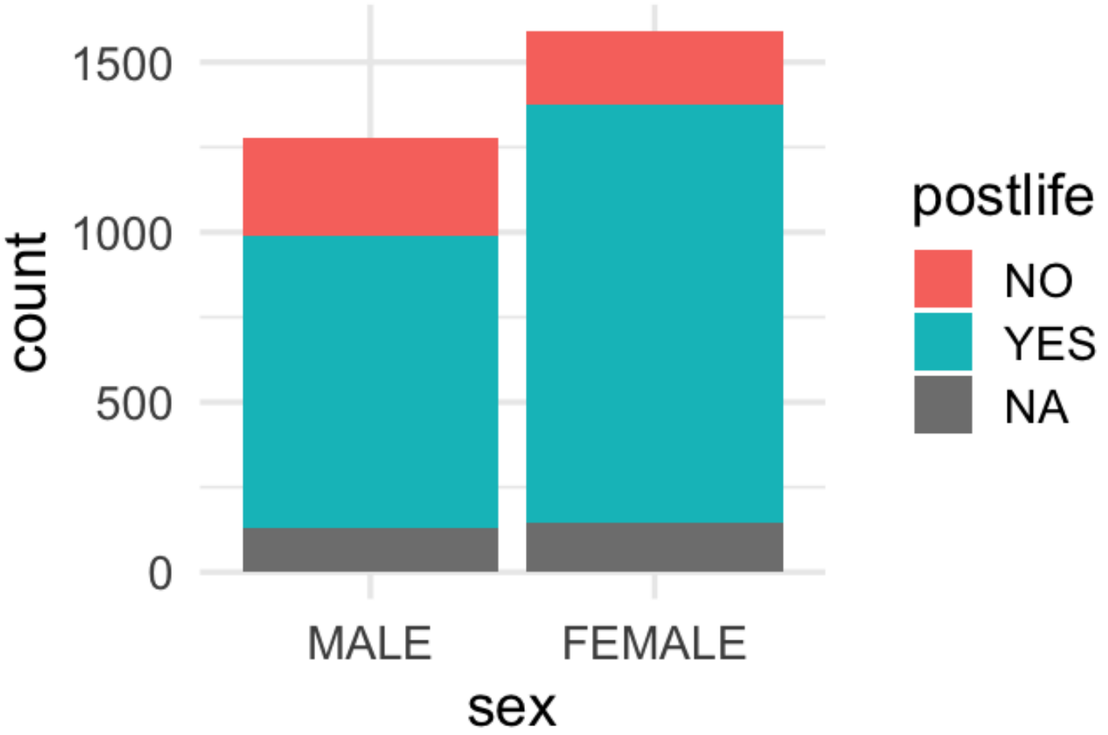
Do women and men believe at different rates?

Let p be the proportion that believe in life after death.

- $H_0 : p_{female} - p_{male} = 0$
- $H_A : p_{female} - p_{male} \neq 0$

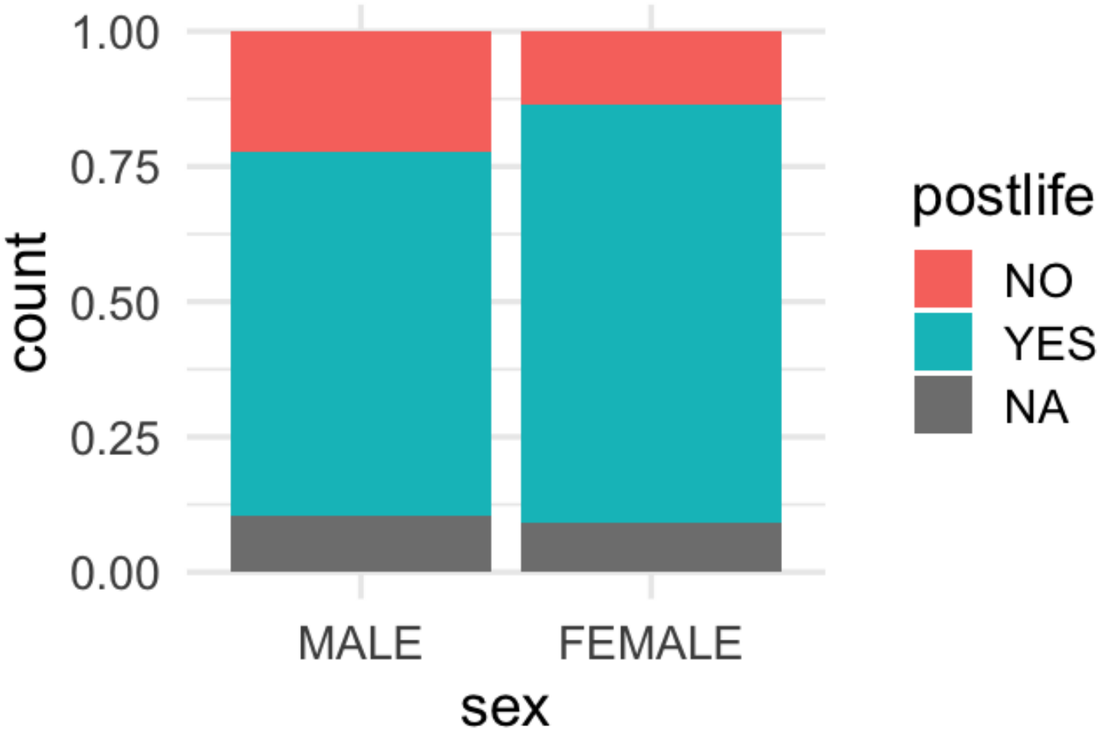
Do women and men have different opinions on life after death?

```
ggplot(gss2016, aes(x = sex, fill = postlife)) +
  geom_bar()
```



Do women and men have different opinions on life after death?

```
ggplot(gss2016, aes(x = sex, fill = postlife)) +
  geom_bar(position = "fill")
```



Do women and men have different opinions on life after death?

```
p_hats <- gss2016 %>%
  group_by(sex) %>%
  summarize(mean(postlife == "YES", na.rm = TRUE)) %>%
  pull()
```

```
d_hat <- diff(p_hats)
d_hat
```

```
0.1472851
```

Generating data from H0

- $H_0 : p_{female} - p_{male} = 0$
- There is no association between belief in the afterlife and the sex of a subject.
- The variable `postlife` is independent from the variable `sex`.

⇒ **Generate data by permutation**

Do women and men have different opinions on life after death?

```
gss2016 %>%
  specify(
    response = postlife,
    explanatory = sex,
    success = "YES"
  ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1, type = "permute")
```

Do women and men have different opinions on life after death?

```
gss2016 %>%
  specify(
    postlife ~ sex,  # this line is new
    success = "YES"
  ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1, type = "permute")
```

```
Response: postlife (factor)
Explanatory: sex (factor)
Null Hypothesis:  independence
# A tibble: 137 x 3
# Groups:   replicate [1]
   postlife sex    replicate
   <fct>    <fct>      <int>
 1 YES      FEMALE         1
 2 YES      MALE           1
 3 YES      FEMALE         1
 4 YES      MALE           1
 5 YES      MALE           1
 6 YES      FEMALE         1
 7 NO       FEMALE         1
```

Do women and men have different opinions on life after death?

```
gss2016 %>%
  specify(
    postlife ~ sex,
    success = "YES"
  ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1, type = "permute")
```

```
Response: postlife (factor)
Explanatory: sex (factor)
Null Hypothesis:  independence
# A tibble: 137 x 3
# Groups:   replicate [1]
   postlife sex    replicate
   <fct>    <fct>      <int>
 1 YES      FEMALE         1
 2 NO       MALE           1
 3 NO       FEMALE         1
 4 YES      MALE           1
 5 YES      MALE           1
 6 YES      FEMALE         1
 7 YES      FEMALE         1
```

Do women and men have different opinions on life after death?

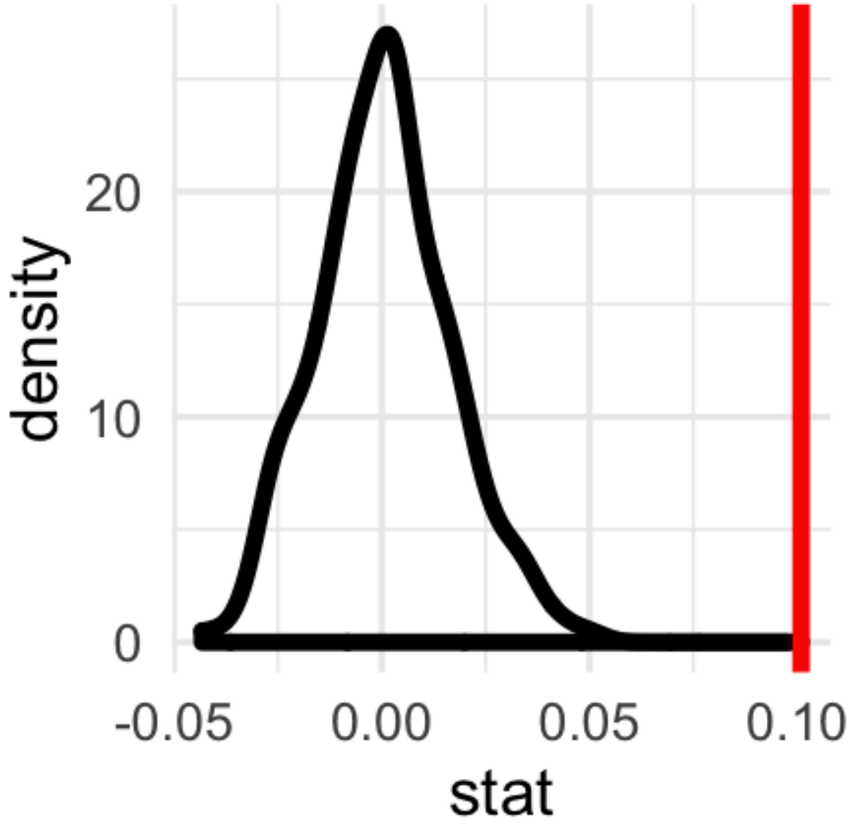
```
gss2016 %>%
  specify(postlife ~ sex, success = "YES") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500, type = "permute") %>%
  calculate(stat = "diff in props", order = c("FEMALE", "MALE"))
```

```
Warning message:
Removed 13 rows containing missing values.
```

Do women and men have different opinions on life after death?

```
ggplot(null, aes(x = stat)) +
  geom_density() +
  geom_vline(xintercept = d_hat, color = "red")
```

These data suggest that there is a difference between sexes in the belief of life after death.



Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

Statistical errors

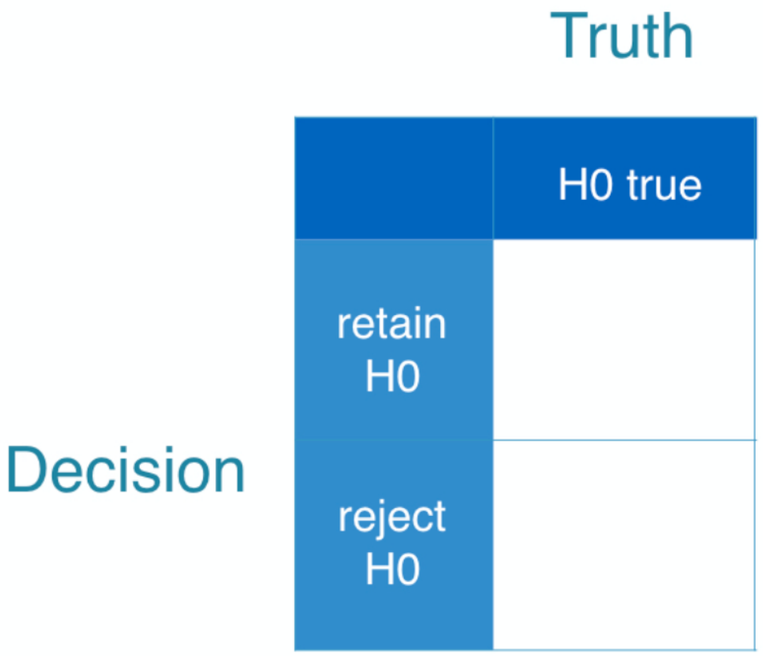
INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

Statistical Errors

Statistical Errors



Statistical Errors

Truth

Decision	H0 true
retain H0	OK
reject H0	

Statistical Errors

Truth

Decision	H0 true
retain H0	OK
reject H0	Type I

Statistical Errors

Truth

Decision	H0 true
retain H0	OK
reject H0	Type I

What is the probability that you will reject a true null hypothesis?

Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true
retain H0	OK
reject H0	Type I

What is the probability that you will reject a true null hypothesis?

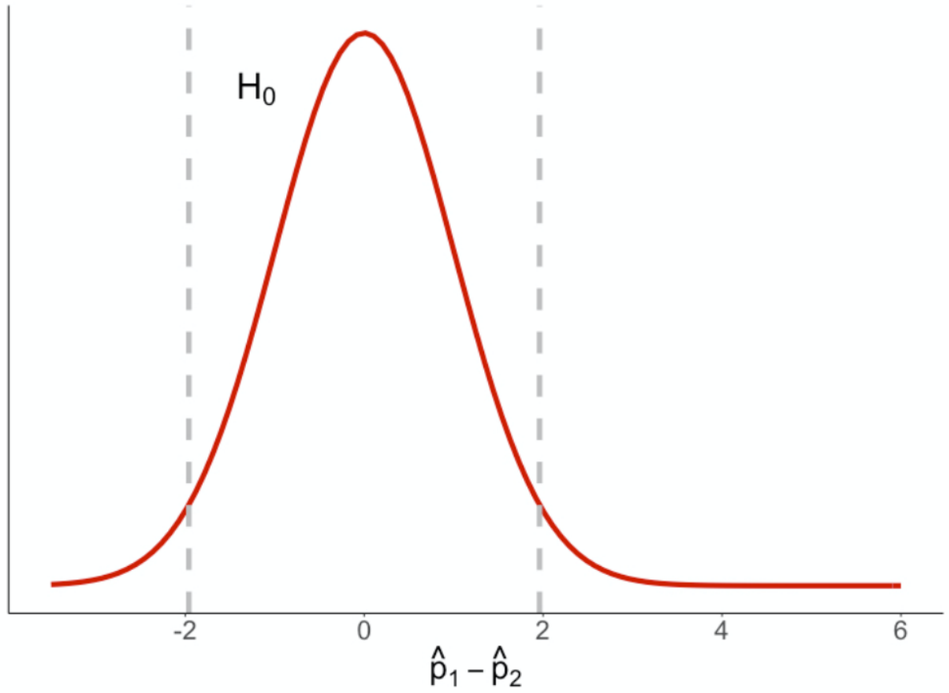
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true
retain H0	OK
reject H0	Type I



What is the probability that you will reject a true null hypothesis?

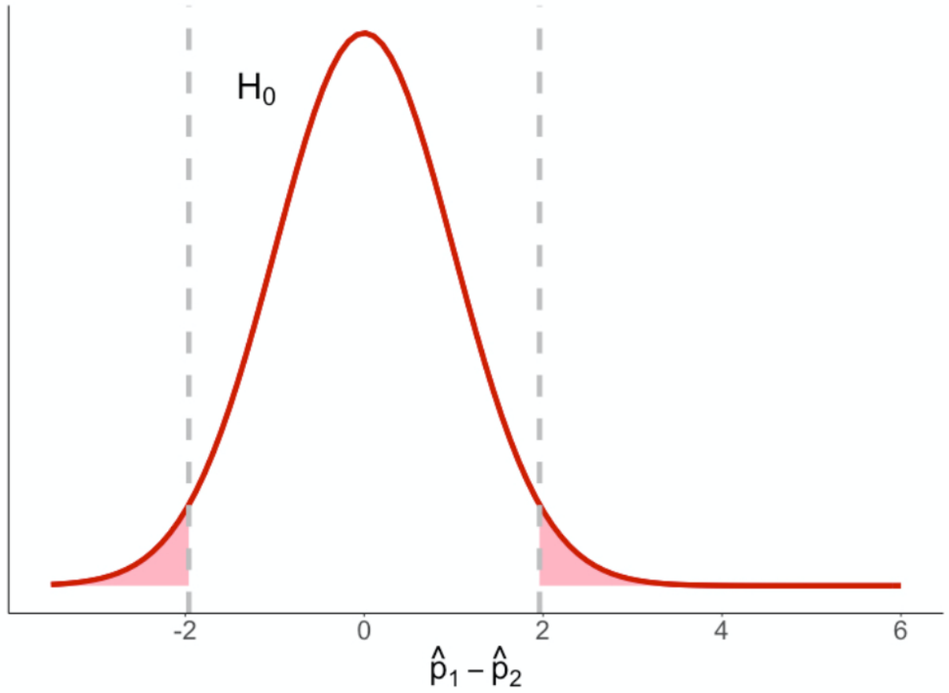
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true
retain H0	OK
reject H0	Type I



What is the probability that you will reject a true null hypothesis?

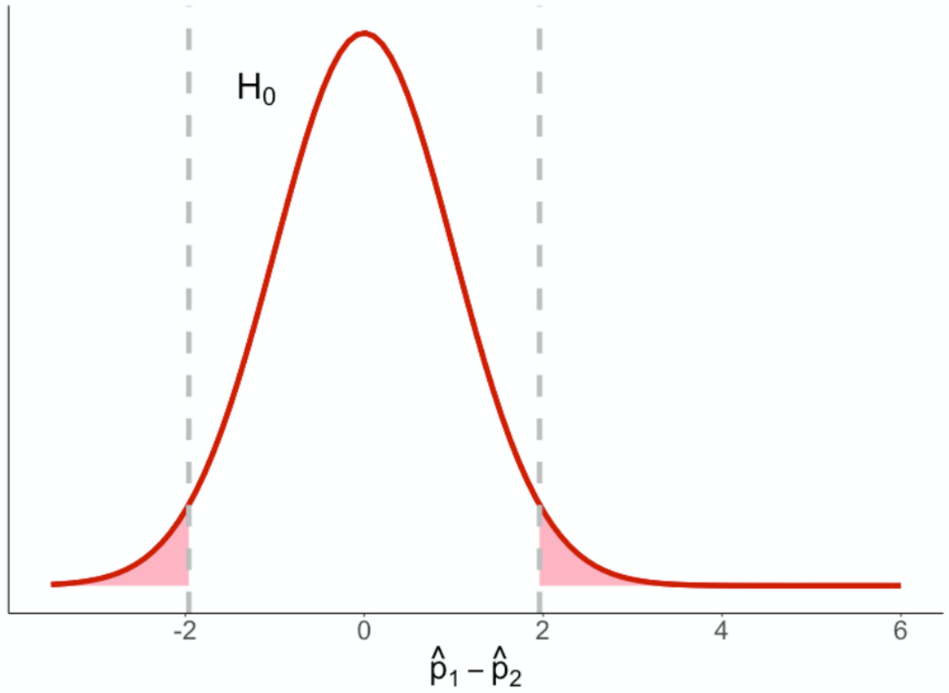
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Decision \ Truth	H0 true
retain H0	OK
reject H0	Type I α

What is the probability that you will reject a true null hypothesis?

α



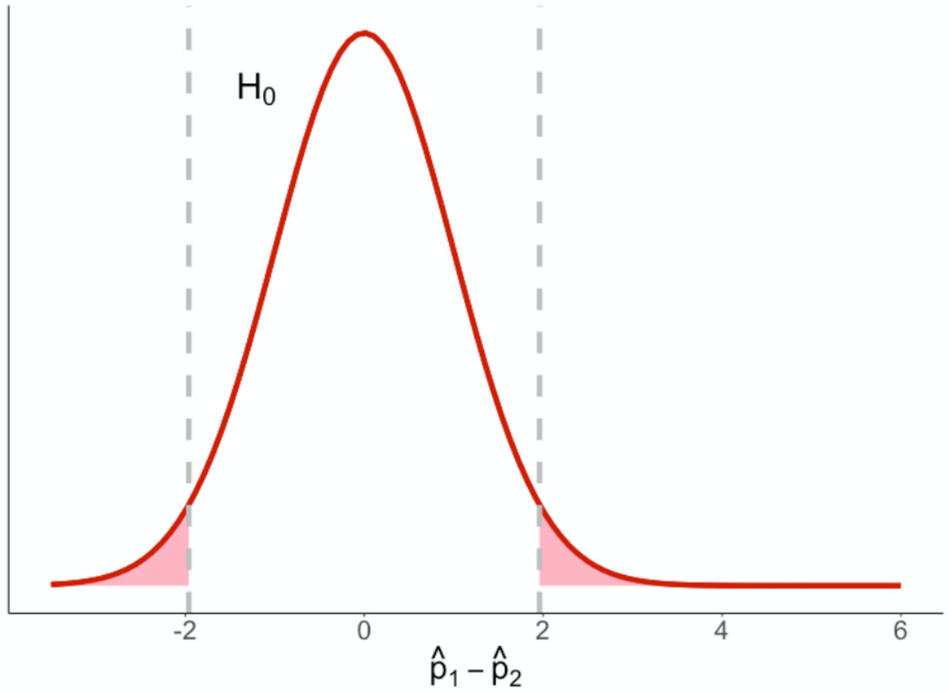
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true
retain H0	OK
reject H0	Type I α



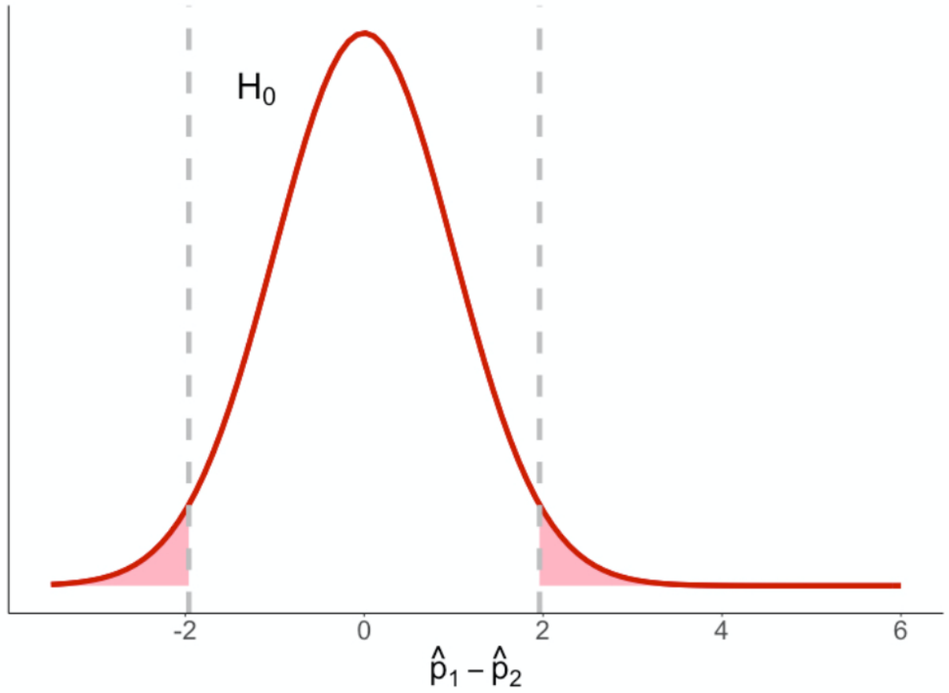
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	
reject H0	Type I α	



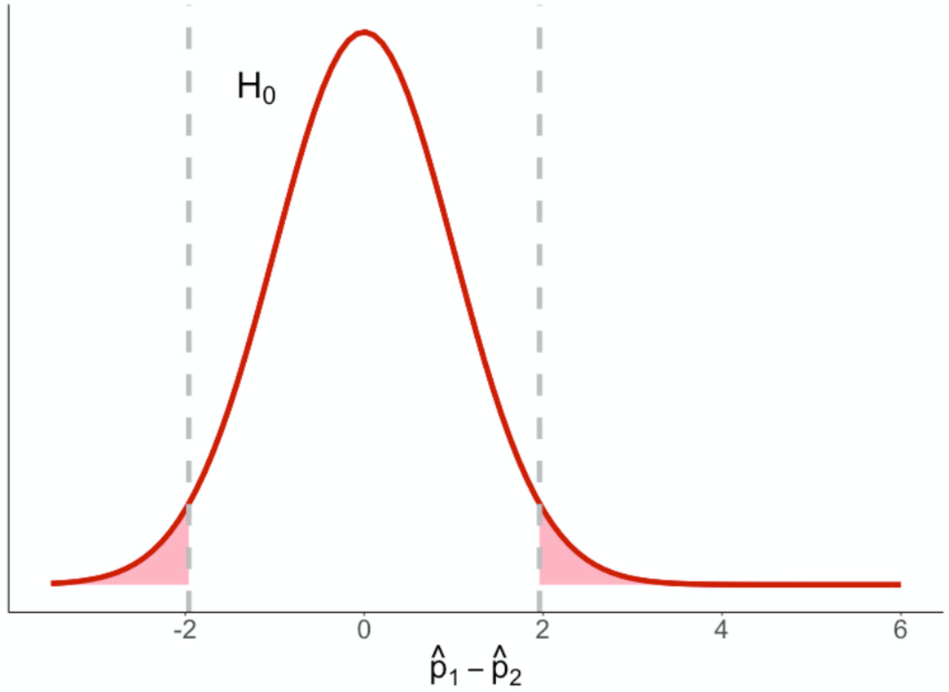
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	
reject H0	Type I α	OK



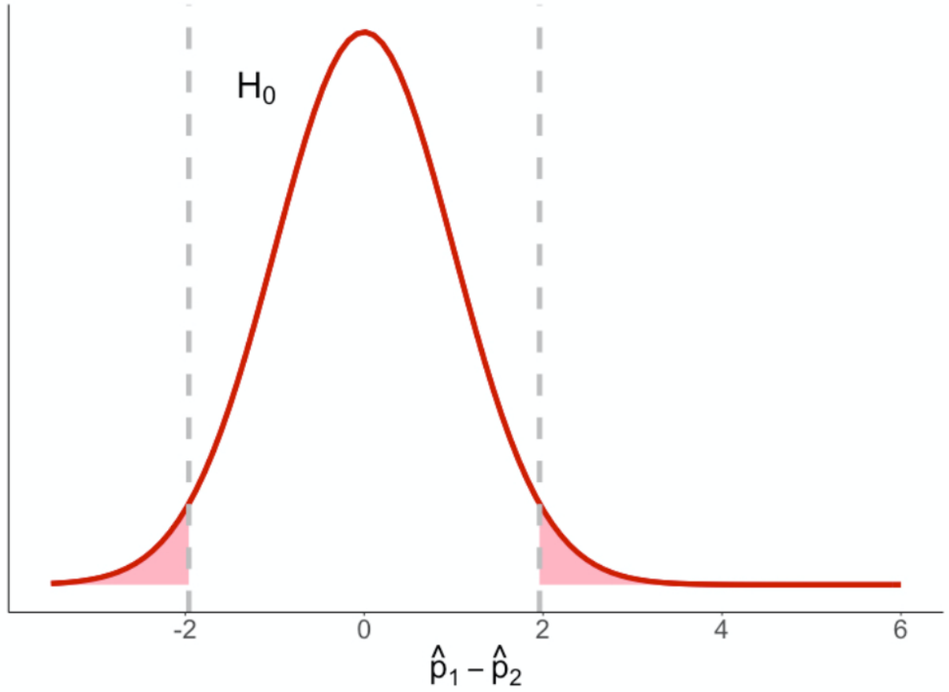
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II
reject H0	Type I α	OK



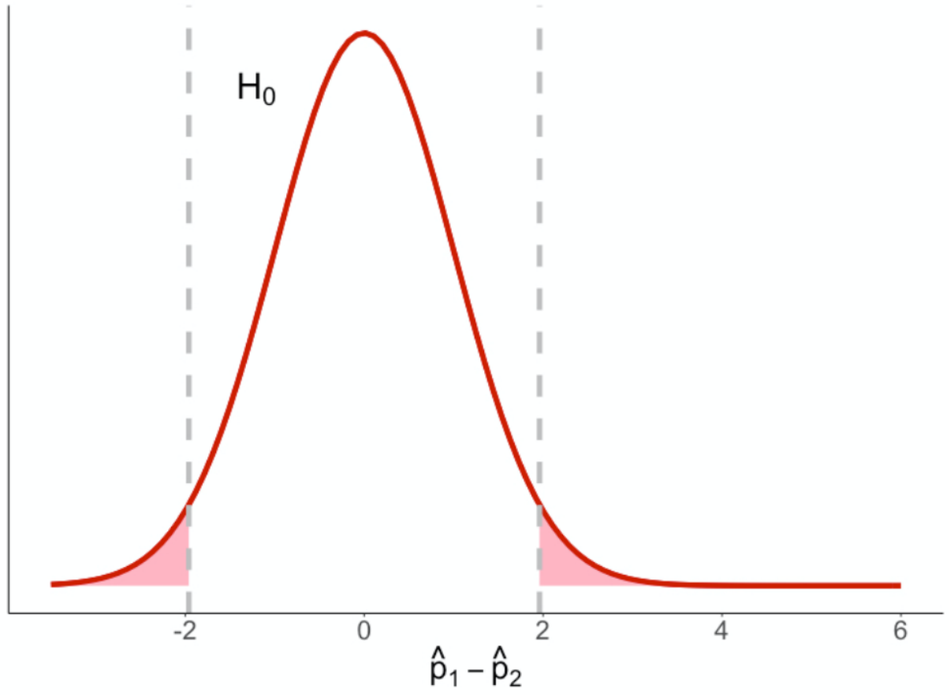
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II
reject H0	Type I α	OK



What is the probability that you will fail to reject a false null hypothesis?

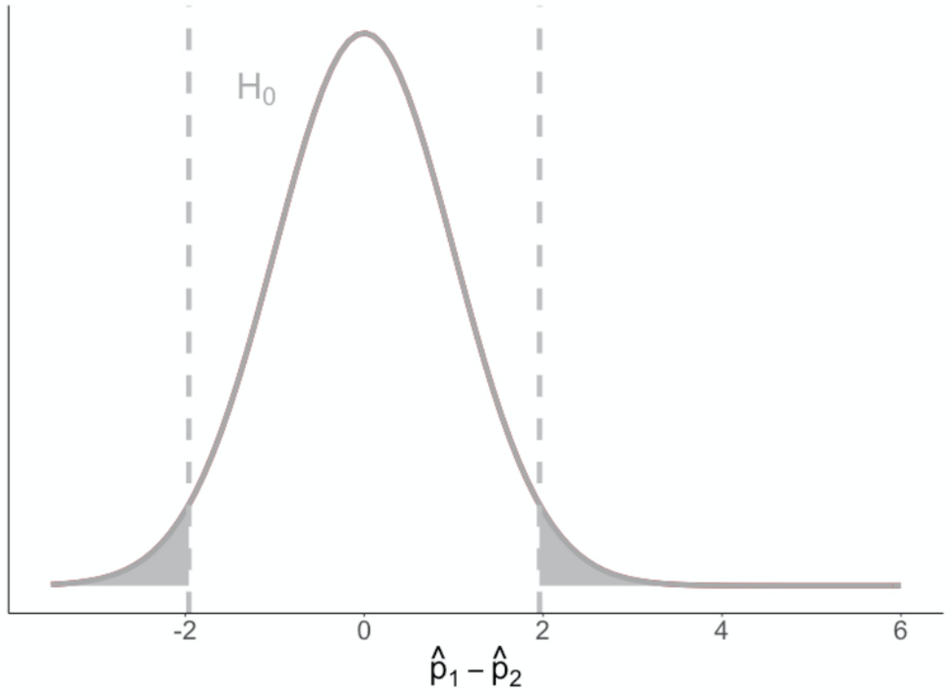
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II
reject H0	Type I α	OK



What is the probability that you will fail to reject a false null hypothesis?

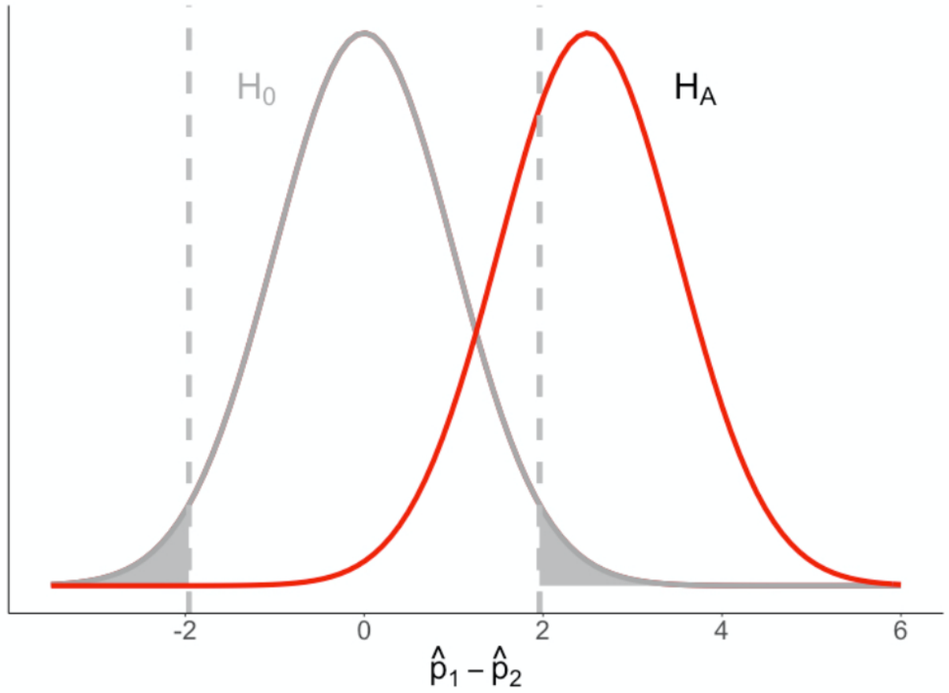
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II
reject H0	Type I α	OK



What is the probability that you will fail to reject a false null hypothesis?

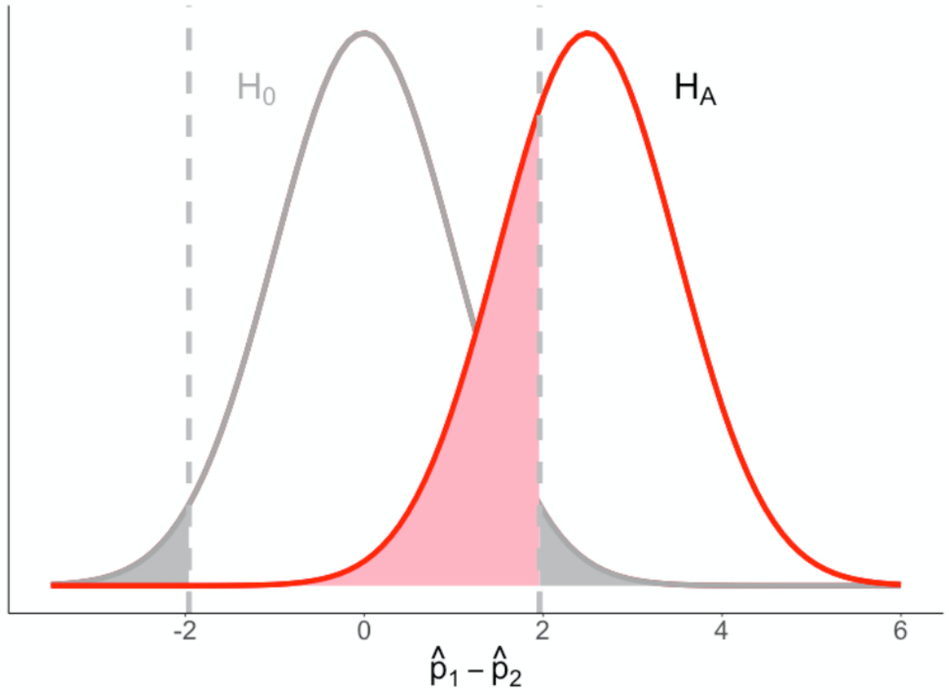
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II
reject H0	Type I α	OK



What is the probability that you will fail to reject a false null hypothesis?

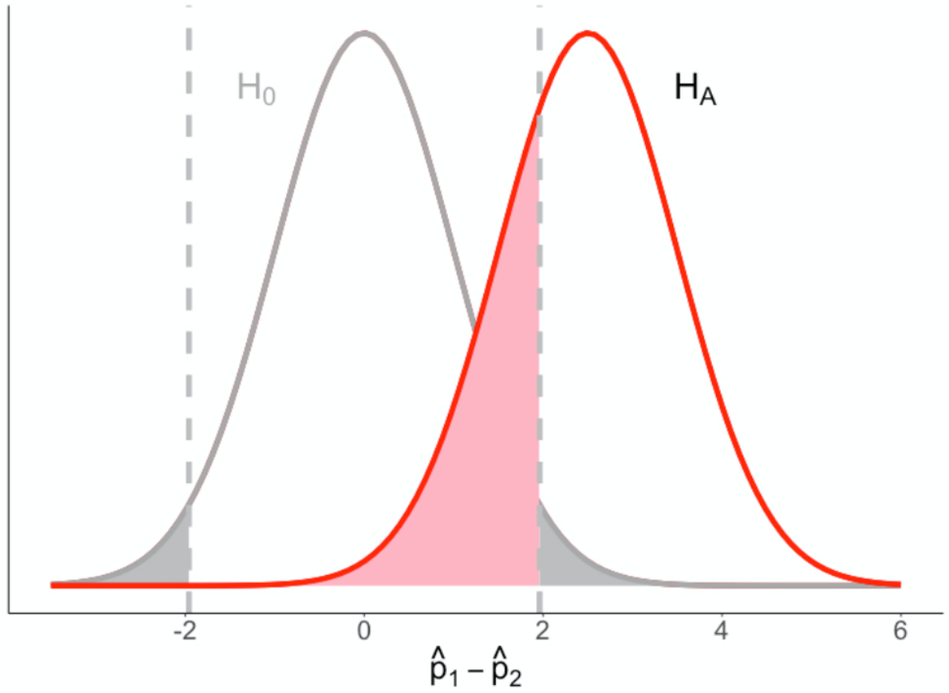
Statistical Errors

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 \neq 0$$

Truth

Decision

	H0 true	H0 false
retain H0	OK	Type II β
reject H0	Type I α	OK



What is the probability that you will fail to reject a false null hypothesis?

β

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R