# Lecture 14: Power and Sample Size Calculations

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Power/Type II Error Rate

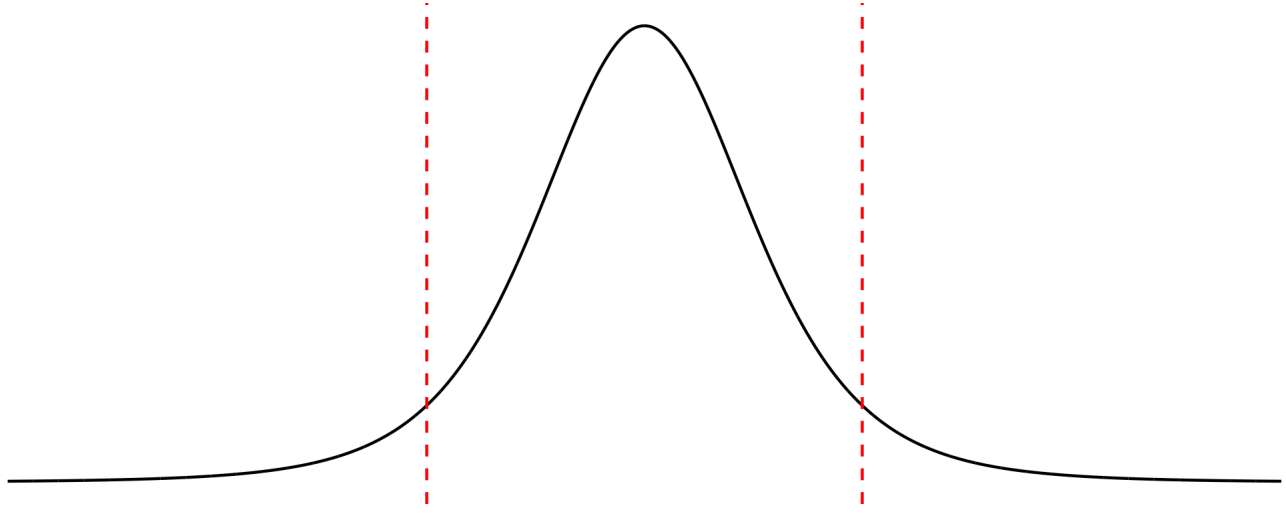
Recall: for every hypothesis test, there is a conclusion. For every conclusion, one of three things will happen:

- we make the right decision
  - i.e. reject when $H_0$ is false, do not reject when $H_0$ is true
- we make a type I error
  - i.e. reject when $H_0$ is actually true
- we make a type II error
  - i.e. fail to reject when $H_0$ is actually false

# Power/Type II Error Rate

We have full control over the type I error rate - it is exactly $\alpha$. To see this, say we are testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$.

$$\begin{aligned} P(\text{reject } H_0 \mid H_0 \text{ true}) &= P(T \text{ very far from } 0 \mid H_0 \text{ true}) \\ &= P(T > t_{n-1,\alpha/2} \mid H_0 \text{ true}) + P(T < -t_{n-1,\alpha/2} \mid H_0 \text{ true}) \\ &= \alpha \end{aligned}$$



Similarly can be done for the two other alternative hypotheses.

# Power/Type II Error Rate

The Type II error rate is a bit more tricky: $P(\text{type II}) = P(\text{fail to reject } H_0 \mid H_0 \text{ false})$.

To get a better sense of how this works, we will consider a simple case that we haven't talked about yet: one sample hypothesis test with known variance $\sigma^2$.

Remember: good old paint thickness data.

```
library(tidyverse); library(distributions3)
paint_thickness <- tibble(
  thickness = c(1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83,
                1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49)
)
```

Let us assume that $\bar{X} \sim N$, and that we know the true variance $\sigma^2 = 0.115$.

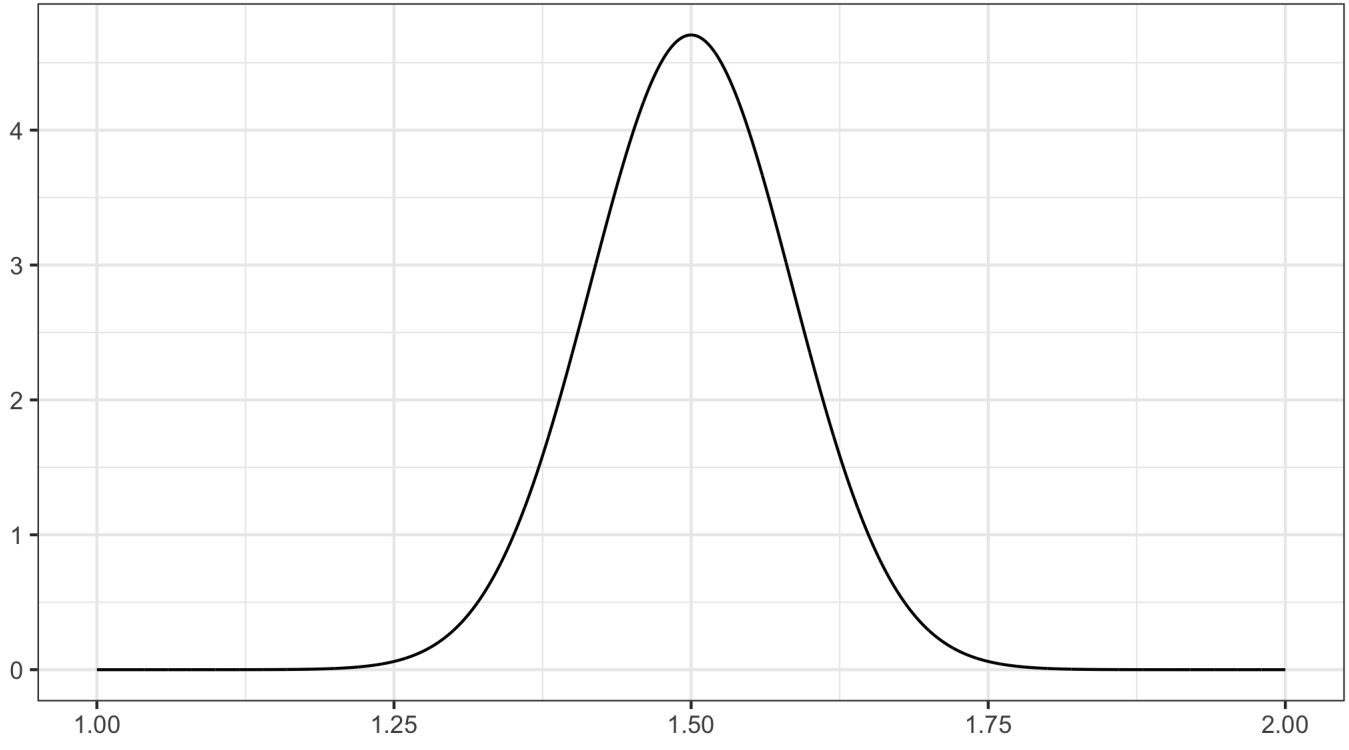
# Power/Type II Error Rate

When testing $H_0 : \mu = 1.5$ against $H_A : \mu \neq 1.5$ using $\alpha = 0.05$, we would reject when $\bar{x}_{\text{obs}}$ is very, very far from $1.5$.

Actually, we reject when $\bar{x}_{\text{obs}}$ is so far from $1.5$ that the probability of $\bar{X}$ being even further from $1.5$ is less than $0.05$: $2 \cdot P(\bar{X} > |\bar{x}_{\text{obs}}| \mid H_0 \text{ true}) < 0.05$.

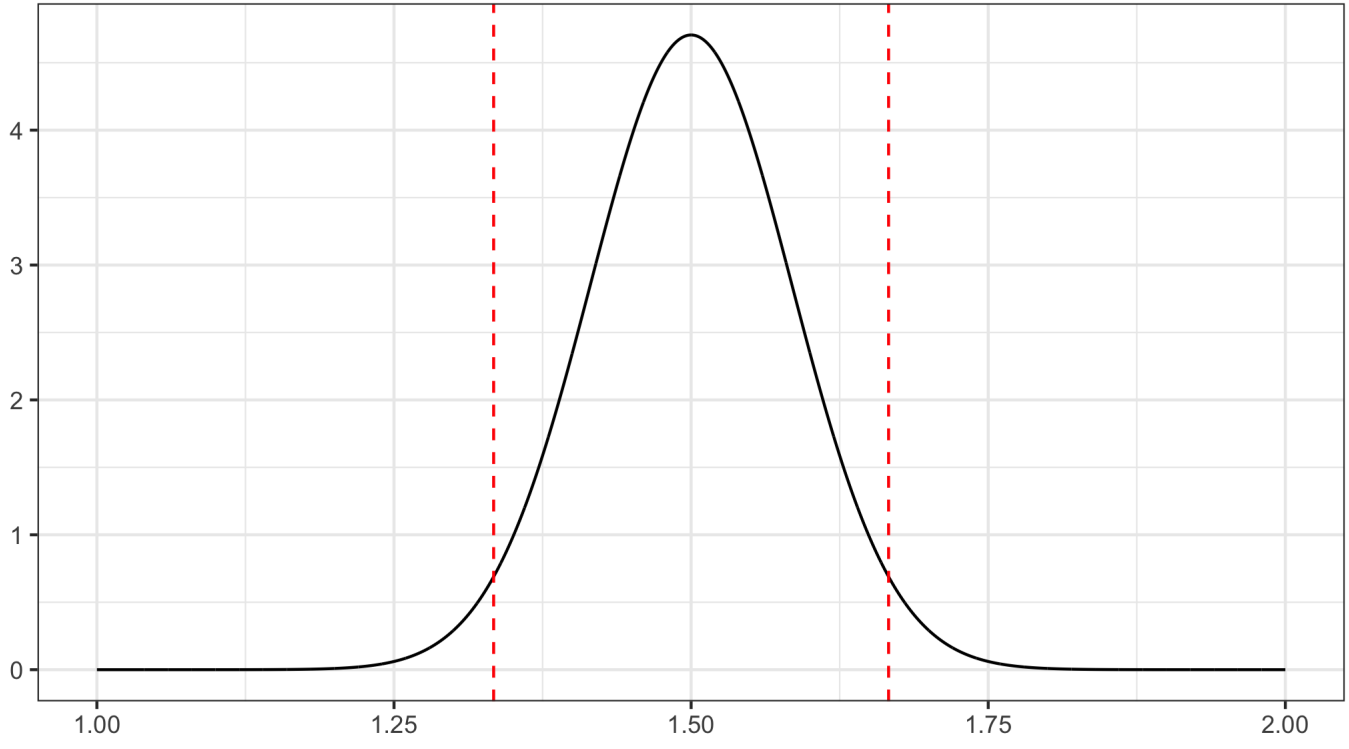
# Power/Type II Error Rate

When $H_0$ is true, $\bar{X} \sim N(1.5, 0.115/16)$. So, when $H_0$ is true, $\bar{X}$ follows this distribution:



# Power/Type II Error Rate

Will reject when $\bar{X}$ outside red dotted lines.




Type II error rate has to do with what happens when the *alternative* is true:
$P(\text{fail to reject } H_0 \mid H_0 \text{ not true}) = P(\text{fail to reject } H_0 \mid H_A \text{ true})$.

Power = 1 - Type II error rate.

# Power/Type II Error Rate

Assuming we know the true $\sigma^2 = 0.115$, and that the null hypothesis is true, then $\bar{X} \sim N(1.5, 0.115/16)$.

So, we reject when $\bar{X}$ outside red dotted lines. I.e. if $0.025 > P\left(\bar{X} > |\bar{x}_{\text{obs}}|\right)$.

```
X_H0 <- Normal(mu = 1.5, sigma = sqrt(0.115/16))

quantile(X_H0, c(0.025, 0.975))
```

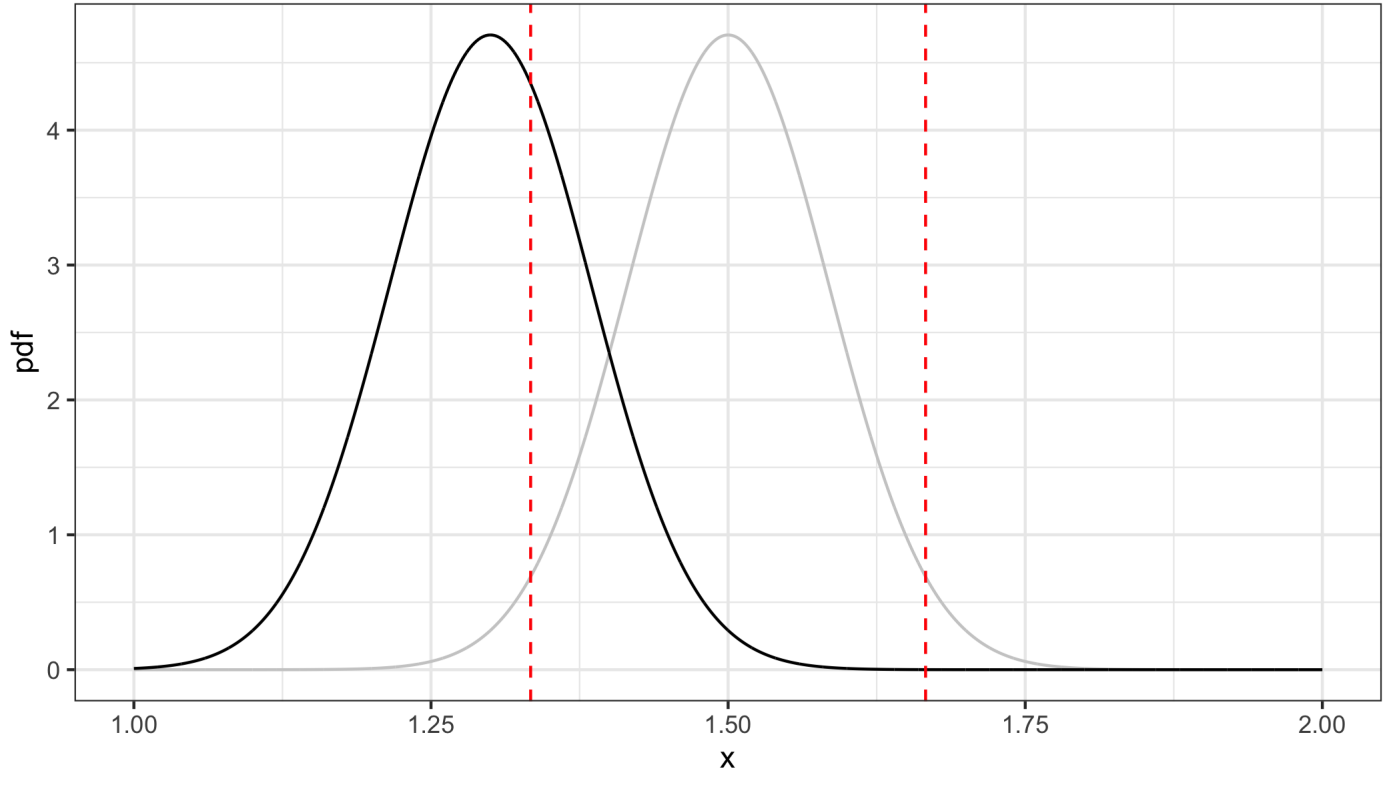
```
## [1] 1.333836 1.666164
```

So, we reject when $\bar{x} < 1.33$ or $\bar{x} > 1.66$.

# Power/Type II Error Rate

But what if the true mean is NOT $1.5$? What if, instead, it is $1.3$? What is $P(\text{type II error})$?

To find the probability of rejecting, we need to look at a different curve, because if $\mu = 1.3$, then $\bar{X} \sim N(1.3, 0.115/16)$.



# Power/Type II Error Rate

But what if the true mean is NOT $1.5$? What if, instead, it is $1.3$? What is $P(\text{type II error})$?

To find the probability of rejecting, we need to look at a different curve, because if $\mu = 1.3$, then $\bar{X} \sim N(1.3, 0.115/16)$.

$$\begin{aligned} P(\text{type II error}) &= P(\text{fail to reject} \mid \mu = 1.3) \\ &= P(1.33 < \bar{X} < 1.66 \mid \mu = 1.3) \\ &= P(\bar{X} < 1.66 \mid \mu = 1.3) - P(\bar{X} < 1.33 \mid \mu = 1.3) \end{aligned}$$

```
X_HA <- Normal(1.3, sqrt(0.115/16))

cdf(X_HA, 1.66) - cdf(X_HA, 1.33)
```

```
## [1] 0.3617108
```

That is, **IF** the true mean is $1.3$, we would fail to reject the idea that $\mu = 1.5$ about $37\%$ of the time.

Or, similarly, we would only reject $\mu = 1.5$ about $63\%$ of the time.

What can we do to do better? Increase sample size!

# Power/Type II Error Rate

What sample size do we need to be able to distinguish between $\mu = 1.5$ and $\mu = 1.3$ most of the time? Say, $80\%$ of the time?

That is, what should $n$ be such that $P(\text{reject } H_0 : \mu = 1.5 \mid \mu = 1.3) = 0.8$?

More general, if we are testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ at significance level $\alpha$, what sample size is needed to make sure that $P(\text{reject } H_0 \mid \mu = \mu_A) = 1 - \beta$?

Turns out, this is approximately

$$n \approx \left( \frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu_A} \right)^2$$

In our specific example, $\sigma = 0.34$, $\alpha = 0.05$, $\beta = 0.2$, $\mu_0 = 1.5$, and $\mu_A = 1.3$.

# Power/Type II Error Rate

We can find $z_{0.025}$ and $z_{0.2}$:

```
Z <- Normal()
quantile(Z, c(1-0.025, 1-0.2))
```

```
## [1] 1.9599640 0.8416212
```

So, $n \approx \left(\frac{0.34(1.96+0.84)}{1.5-1.3}\right)^2 = 22.6576.$

To have $80\%$ power to reject $1.5$ as the true mean if the true mean is in fact $1.3$, we would need 23 samples.

# Power/Type II Error Rate

- sample size depends only on standard deviation, $\alpha$, $\beta$, and *difference* between true and hypothesized means.
- in practice, it goes as follows:
  - researchers pick desired $\alpha$ and $\beta$
  - from previous, well-done experiments, a solid estimate of $\sigma$ is obtained
  - from expert knowledge, a "minimal interesting difference" is chosen (i.e. $\mu_0 - \mu_A$)
  - based on this, needed sample size is determined.

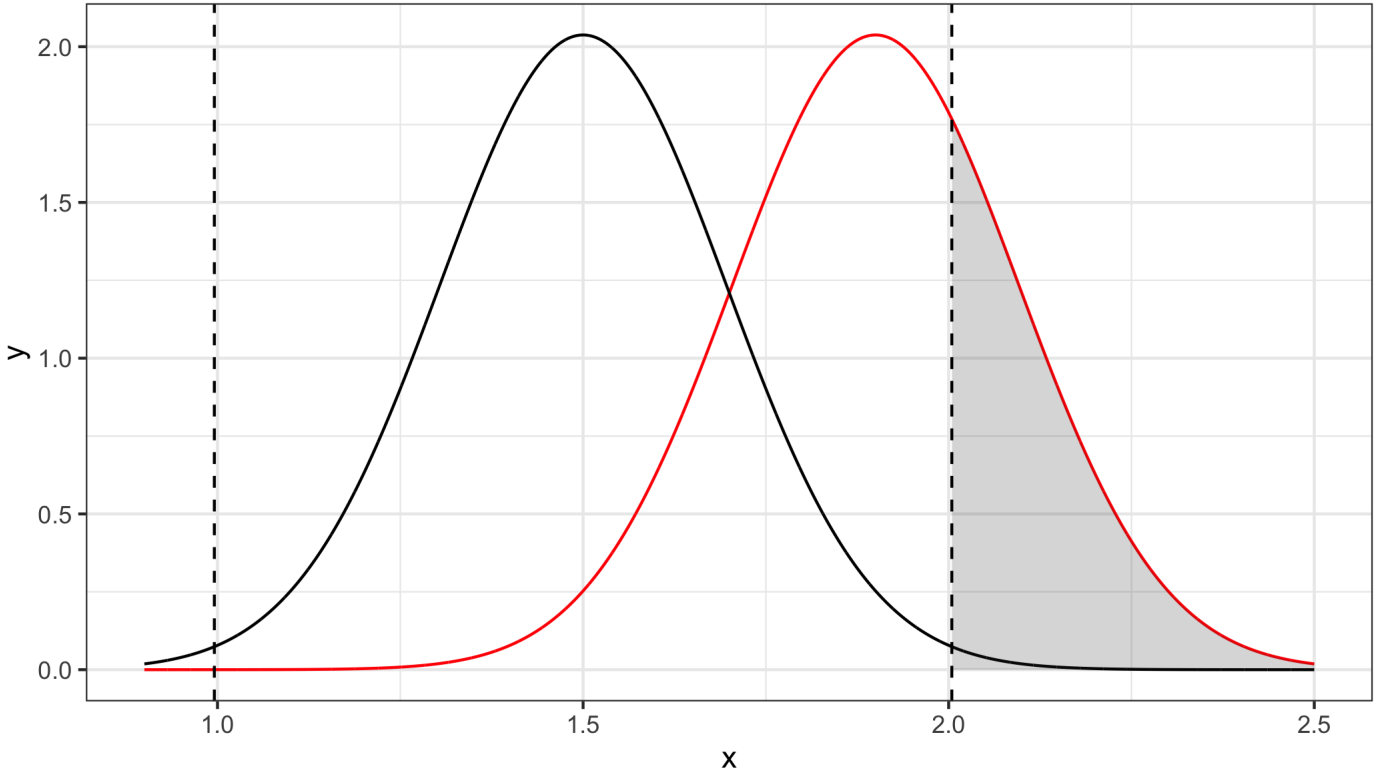
# Power/Type II Error Rate

- Based on the data at hand, the engineers at the car manufacturer want to design a new study.
- They want to test $H_0 : \mu = 1.5$ vs $H_A : \mu \neq 1.5$ using $\alpha = 0.01$
- What sample size is needed to make sure they have $80\%$ power to detect a difference of $0.4$?
    - i.e. if the true mean is $0.4$ from their null hypothesis, they want a $80\%$ chance of rejecting the null

# Power/Type II Error Rate

Based on the data, $\sigma^2 = 0.115$. So, if the null hypothesis is true, $\bar{X} \sim N(1.5, 0.115/n)$.

If the alternative is true (say, $\mu = 1.9$), then $\bar{X} \sim N(1.9, 0.115/n)$. What should $n$ be such that the area under the TRUE curve in the rejection region is 0.8)



# Power/Type II Error Rate

Using formula above:

$$n \approx \left( \frac{0.34(2.576 + 0.842)}{0.4} \right)^2 = 8.4407681$$

So we would need a sample size of $9$ to achieve the desired power to detect a difference of $0.4$.

# Power/Type II Error Rate

Sometimes, the sample size is limited by other resources such as time or money. A natural question to ask is then, what power do we have to detect a certain difference, given the sample size we can afford?

We want to test $H_0 : \mu = 1.5$ against $H_A : \mu \neq 1.5$ using $\alpha = 0.1$. With a sample size of $5$, what power do we have to detect a difference of $0.3$ if the true variance is $\sigma^2 = 0.115$?

Step 1: find rejection region.

- we do this **assuming $H_0$ is true**. I.e. what value of $\bar{x}_{\text{obs}}$ would lead us to reject $\mu = 1.5$?
- we reject when $P(\text{more extreme} \mid H_0) < 0.1$. I.e. reject when $\bar{x}_{\text{obs}}$ is smaller than the $0.05$ th quantile or larger than the $0.95$th of the distributions of $\bar{X}$

```
Xbar <- Normal(1.5, sqrt(0.115/5))
quantile(Xbar, 1-0.05)
```

```
## [1] 1.25055 1.74945
```

- so RR is $(-\infty, 1.25055]$ and $[1.74945, \infty)$.

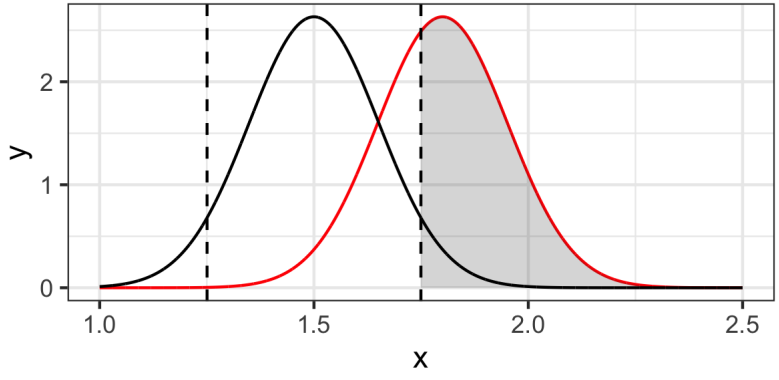
# Power/Type II Error Rate

Step 2: What is the probability of $\bar{X}$ being in the RR **IF** the *alternative* is true?

- if the alternative is a difference of $0.3$ from the null, then $\mu = 1.2$ or $\mu = 1.8$. Because of the symmetry of the normal distribution, which you pick won't change the result
- so, say that the alternative is true such that $\mu = 1.8$. Then what is $P(\bar{X} \text{ in } RR)$?
- $P(\bar{X} < 1.25055 \mid \mu = 1.8) + P(\bar{X} > 1.74945 \mid \mu = 1.8)$

```
X_HA <- Normal(1.8, sqrt(0.115/5))
cdf(X_HA, 1.25055) + (1 - cdf(X_HA, 1.74945))
```

```
## [1] 0.6306981
```



# Two Independent Samples Hypothesis Test

The horned lizard Phrynosoma mcalli is named for the fringe of spikes around the back of the head. It was thought that the spikes may provide the lizard protection from its primary predator, the loggerhead shrike, Lanius ludovicanus, though there was not much existing quantitative evidence to support this. Researchers were interested in comparing two populations: the population of dead lizards known to be killed by shrikes, and the population of live lizards from the same geographic location. Random samples were taken from each population. The longest spike was measured on each sampled lizard, in mm.

# Two Independent Samples Hypothesis Test

The fundamental question: is there, overall, a difference between the longest spike in the two populations?

In terms of means: is $\mu_{\text{dead}} = \mu_{\text{alive}}$ or not?

Some data:

```
DT::datatable(lizards, options = list(pageLength = 4))
```

Show 4 entries Search:

| | group | size |
|---|---|---|
| 1 | dead | 17.65 |
| 2 | dead | 20.83 |
| 3 | dead | 24.59 |
| 4 | dead | 18.52 |

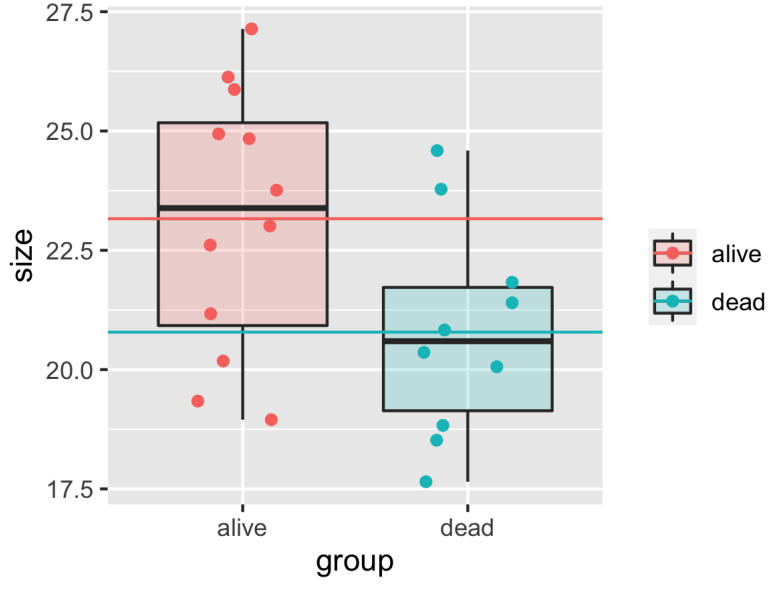
Showing 1 to 4 of 22 entries Previous 1 2 3 4 5 6 Next

# Two Independent Samples Hypothesis Test

The fundamental question: is there, overall, a difference between the longest spike in the two populations?

In terms of means: is $\mu_{\text{dead}} = \mu_{\text{alive}}$ or not?

```
ggplot(lizards,
       aes(x = group, y = size)) +
  geom_boxplot(aes(fill = group),
               alpha = 0.2) +
  geom_hline(data = lizards %>%
               group_by(group) %>%
               summarize(mean = mean(siz
             aes(yintercept = mean, colo
  geom_jitter(height = 0, width = 0.2,
              aes(color = group)) +
  labs(color = "", fill = "")
```




Are the lines so far apart that we reject the idea that the underlying true means are the same?

# Two Independent Samples Hypothesis Test

Since $\bar{X}_{\text{dead}}$ is expected to be close to $\mu_{\text{dead}}$, and $\bar{X}_{\text{alive}}$ is expected to be close to $\mu_{\text{alive}}$, $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}$ is expected to be close to $\mu_{\text{alive}} - \mu_{\text{dead}}$.

We can rephrase the question in terms of hypotheses:

$$H_0 : \mu_{\text{alive}} - \mu_{\text{dead}} = 0 \qquad \text{vs.} \qquad H_A : \mu_{\text{alive}} - \mu_{\text{dead}} \neq 0$$

So the question is, is our observed difference in averages ( $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}$ ) so far from 0 that we no longer think that $\mu_{\text{alive}} - \mu_{\text{dead}} = 0$ (i.e. we reject the idea that the means are the same)?

How would we go about answering this question?

# Two Independent Samples Hypothesis Test

**IF** $\bar{X}_{\text{alive}} \sim N$ and $\bar{X}_{\text{dead}} \sim N$, then $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}} \sim N$.

**IF** $H_0$ is true, then $E(\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}) = E(\bar{X}_{\text{alive}}) - E(\bar{X}_{\text{dead}}) = \mu_{\text{alive}} - \mu_{\text{dead}} = 0$.

So, **IF** $H_0$ is true, then $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}} \sim N(0, ??)$.

**IF** the two samples are independent of each other, $\bar{X}_{\text{alive}}$ is independent of $\bar{X}_{\text{dead}}$, so

$$\text{Var}(\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}) = \text{Var}(\bar{X}_{\text{alive}}) + \text{Var}(\bar{X}_{\text{dead}}) = \frac{\sigma^2_{\text{alive}}}{n_{\text{alive}}} + \frac{\sigma^2_{\text{dead}}}{n_{\text{dead}}}$$

So, **IF** $H_0$ is true, then $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}} \sim N\left(0, \frac{\sigma^2_{\text{alive}}}{n_{\text{alive}}} + \frac{\sigma^2_{\text{dead}}}{n_{\text{dead}}}\right)$.

# Two Independent Samples Hypothesis Test

So, how do we judge if what we see are so far from the null hypothesis that we decide to reject it?

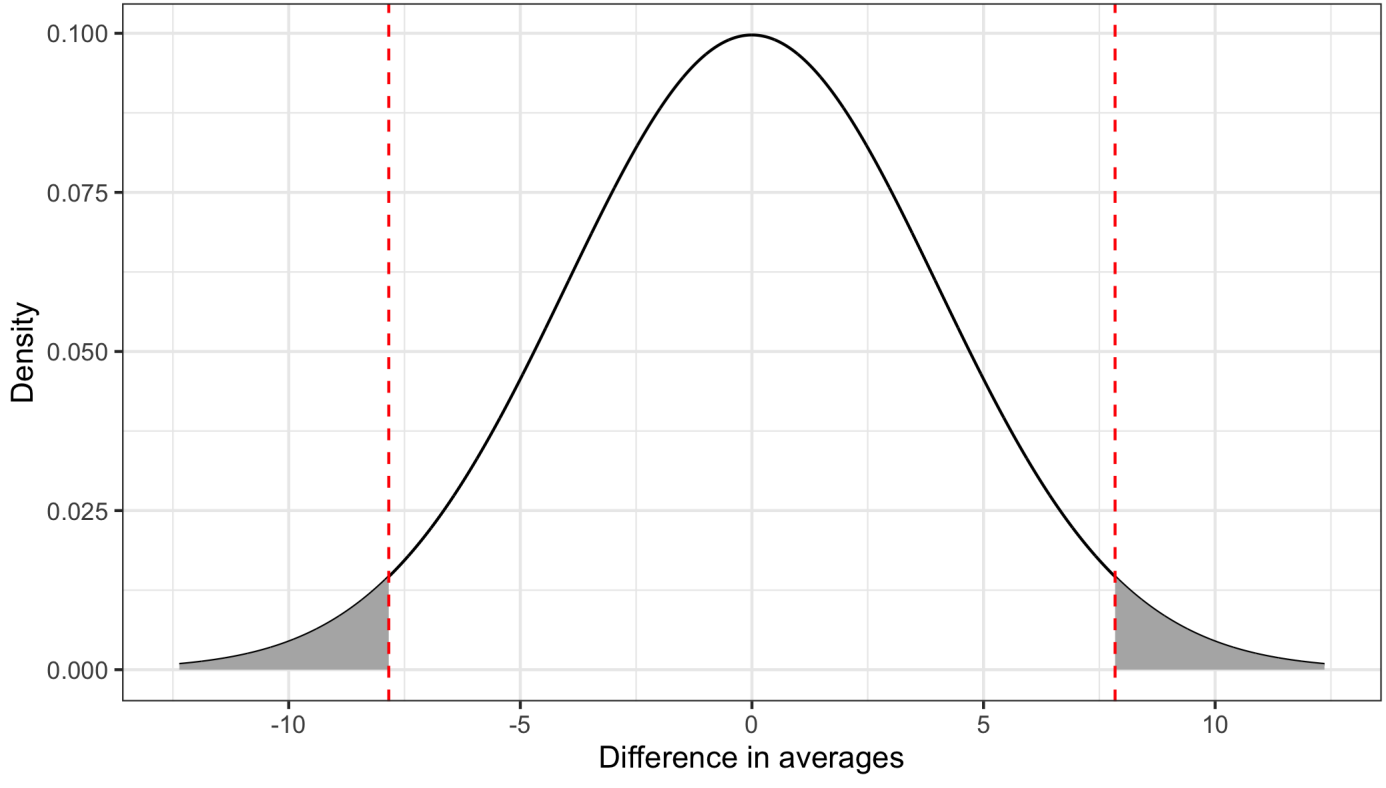
By finding the probability of observing something more extreme if we were to repeat the experiment, **assuming the null hypothesis is true**!

**IF** $H_0$ is true, **and we know** $\sigma_{\text{alive}}, \sigma_{\text{dead}}$, this is pretty straight forward:

- look at the curve that is the distribution of the difference $\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}$, i.e. $N\left(0, \frac{\sigma^2_{\text{alive}}}{n_{\text{alive}}} + \frac{\sigma^2_{\text{dead}}}{n_{\text{dead}}}\right)$.
- using quantiles:
  - find quantiles that cut-off $\alpha/2$ on each side.
  - reject if observed value of the difference is outside the cut-offs
- using p-value:
  - find probability of something "more extreme"
  - reject if smaller than $\alpha$

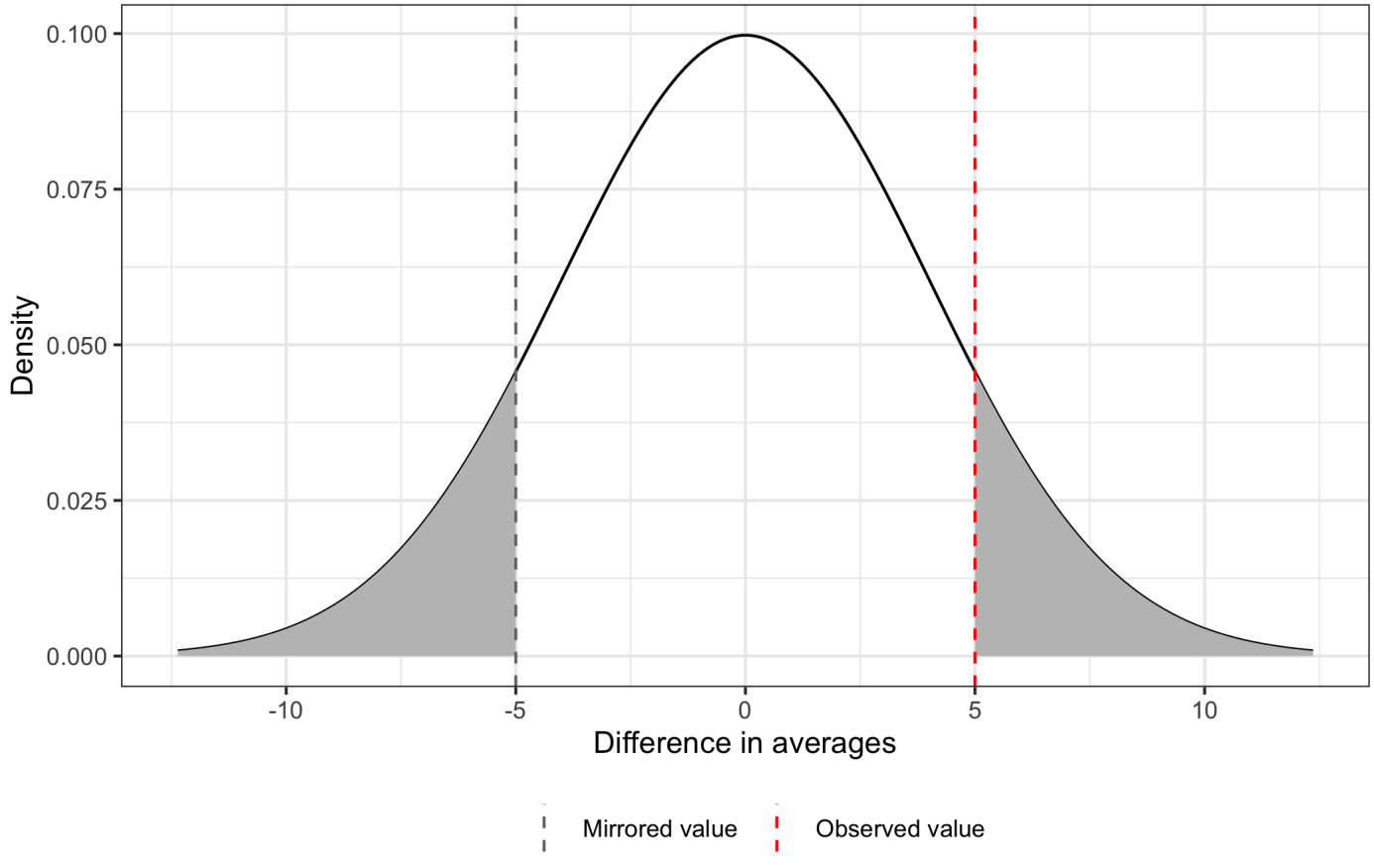
# Two Independent Samples Hypothesis Test

Using quantiles: reject if outside of dotted lines that cut-off $\alpha/2$ on each side.



# Two Independent Samples Hypothesis Test

Using p-value: reject if area outside dotted lines is smaller than $\alpha$



# Two Independent Samples Hypothesis Test

Problem: we never know $\sigma_{\text{dead}}, \sigma_{\text{alive}}$!!

In the one sample case, we got around this by considering $\frac{\bar{X}-\mu_0}{\widehat{\text{SD}}(\bar{X})} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$, which we know is $t_{n-1}$.

In the two sample case, we will use

$$T = \frac{V - v_0}{\widehat{\text{SD}}(V)},$$

where $V = \bar{X}_{\text{dead}} - \bar{X}_{\text{alive}}$, and (in the most general case)

$$\widehat{\text{SD}}(V) = \sqrt{s^2_{\text{dead}}/n_{\text{dead}} + s^2_{\text{alive}}/n_{\text{alive}}}$$

As usual, **IF** $V \sim N$, and $H_0 : v = v_0$, then $T \sim t_{\text{some appropriate df}}$. Things get a bit more tricky here, though, as deciding the approriate df is not trivial.

# Two Independent Samples Hypothesis Test

In general, two scenarios:

**Scenario 1**: $\sigma_1^2 = \sigma_2^2$.

When this is the case, we replace both by a common number, $\sigma^2_{\text{pooled}}$ (or simply $\sigma_p^2$ for convenience).

Adding this extra bit of information means we can do slightly better in trying to estimate the variance in the two groups. Our best guess for the pooled variance is

$$\hat{\sigma}_p^2 = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

# Two Independent Samples Hypothesis Test

$$\hat{\sigma}_p^2 = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Intuition:

- this is a *weighted average* of our two best guesses
- we have a best guess for group 1, best guess for group 2, so surely the "truth" must be somewhere between the two.
- the group with more data (i.e. more information) gets more weight
- if the means in the two groups were the same, the pooled standard deviation is actually the same as just treating the two groups as one.
    - cannot do this when means are different because of definition of standard deviation

We now have that $\text{Var}(D) = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} = s_p^2(1/n_1 + 1/n_2)$, and our test statistic will follow a $t_{n_1+n_2-2}$ distribution:

$$T = \frac{D - d_0}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

# Two Independent Samples Hypothesis Test

**Scenario 2**: $\sigma_1^2 \neq \sigma_2^2$.

In this case, we do not gain any insights, and so there's no adjustments we can make to the test statistic.

It turns out that

$$T = \frac{D - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_\nu,$$

where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

In either case, we can find the distribution of $T$, and use this to either reject or not reject the null hypothesis!