Lecture 15: Two Sample Hypothesis Tests

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

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

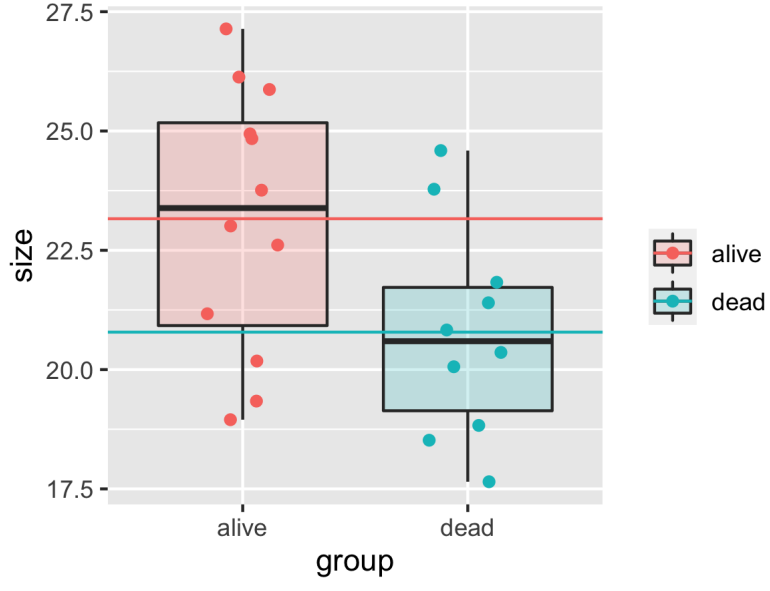
Showing 1 to 4 of 22 entries Previous 1 2 3 4 5 6 Next

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 μ_{dead}, and $\bar{X}_{\text{alive}}$ is expected to be close to μ_{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 is so far from the null hypothesis that we decide to reject it?

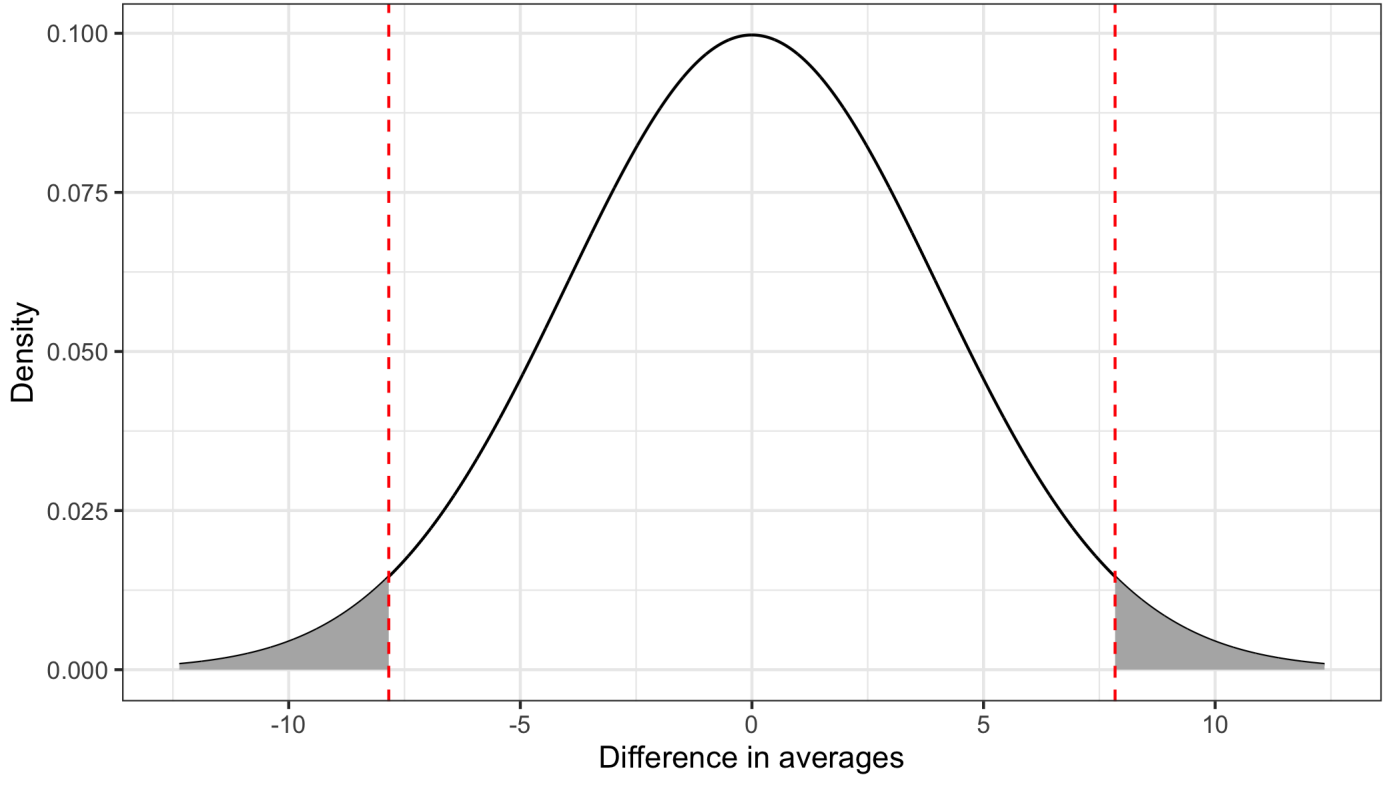
By finding the probability of observing something more extreme if we were to repeat the experiment, **assuming the null hypothesis is true**!

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 α

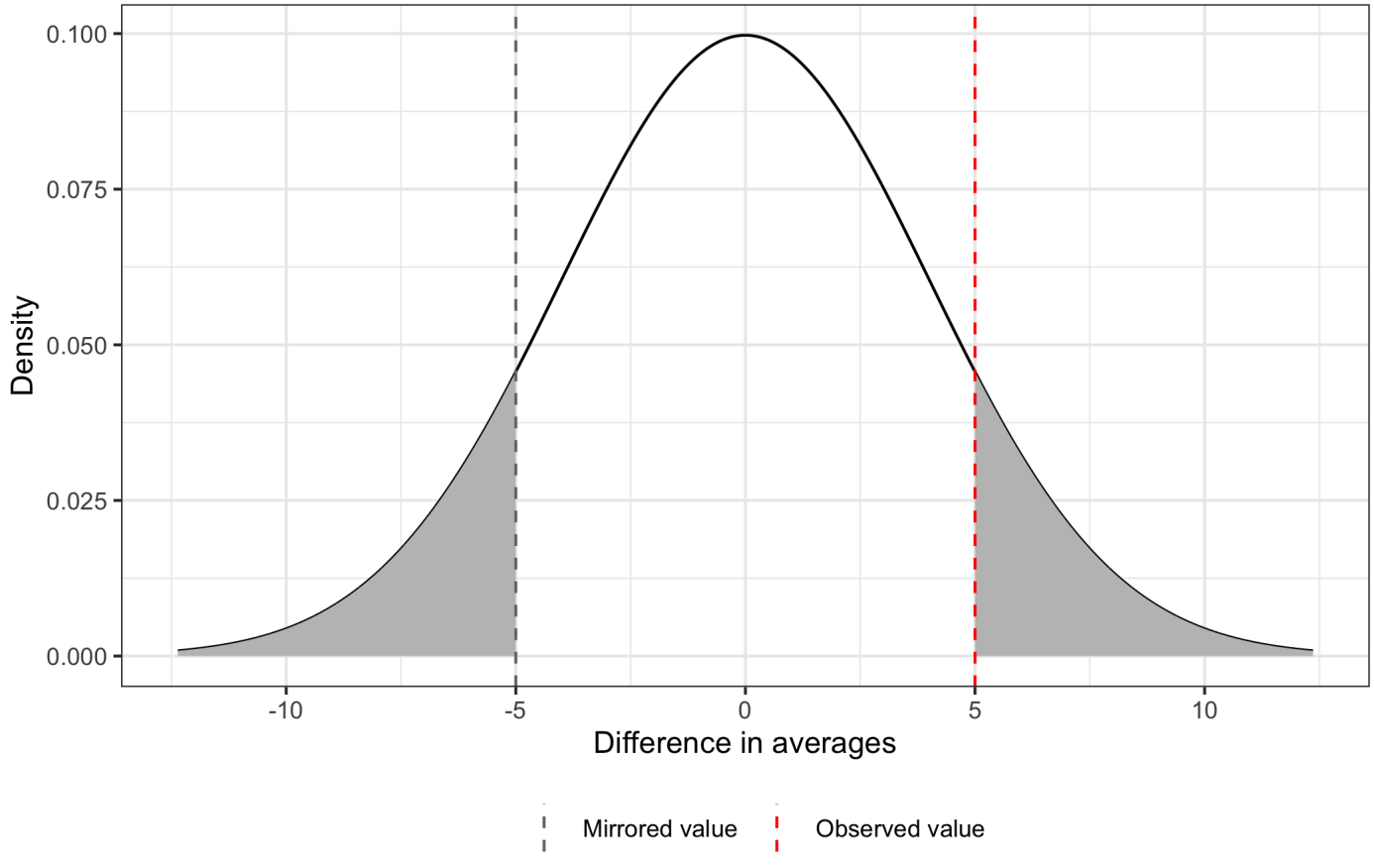
Two Independent Samples Hypothesis Test

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 α



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

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

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, σ^2_{pooled} (or simply σ_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 $\text{Var}(V) = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} = s_p^2(1/n_1 + 1/n_2)$, and our test statistic will follow a $t_{n_1+n_2-2}$ distribution:

$$T = \frac{V - v_0}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

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

$$T = \frac{V - v_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_\nu,$$

where

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

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

Two Independent Samples Hypothesis Test

Since we do not know σ_1^2 or σ_2^2, we use s_1^2 and s_2^2 to determine if we find it plausible that $\sigma_1^2 = \sigma_2^2$.

The general rule of thumb: if $0.5 < \frac{s_1}{s_2} < 2$, then we are okay with assuming $\sigma_1^2 = \sigma_2^2$.

In our lizards example:

```
std_devs <- lizards %>%
  group_by(group) %>%
  summarize(s = sd(size),
            n = n())

std_devs
```

```
## # A tibble: 2 x 3
##   group     s     n
##   <chr> <dbl> <int>
## 1 alive  2.76    12
## 2 dead   2.22    10
```

So, we will assume $\sigma_{\text{dead}}^2 = \sigma_{\text{alive}}^2$.

Two Independent Samples Hypothesis Test

That means we have to find s_p^2 for our test:

```
sp2 <- ((12 - 1)*2.7607339^2 + (10 - 1)*2.2212872^2)/(12 + 10 - 2)
sp2
```

```
## [1] 6.412261
```

We can now calculate the observed value of the test statistic:

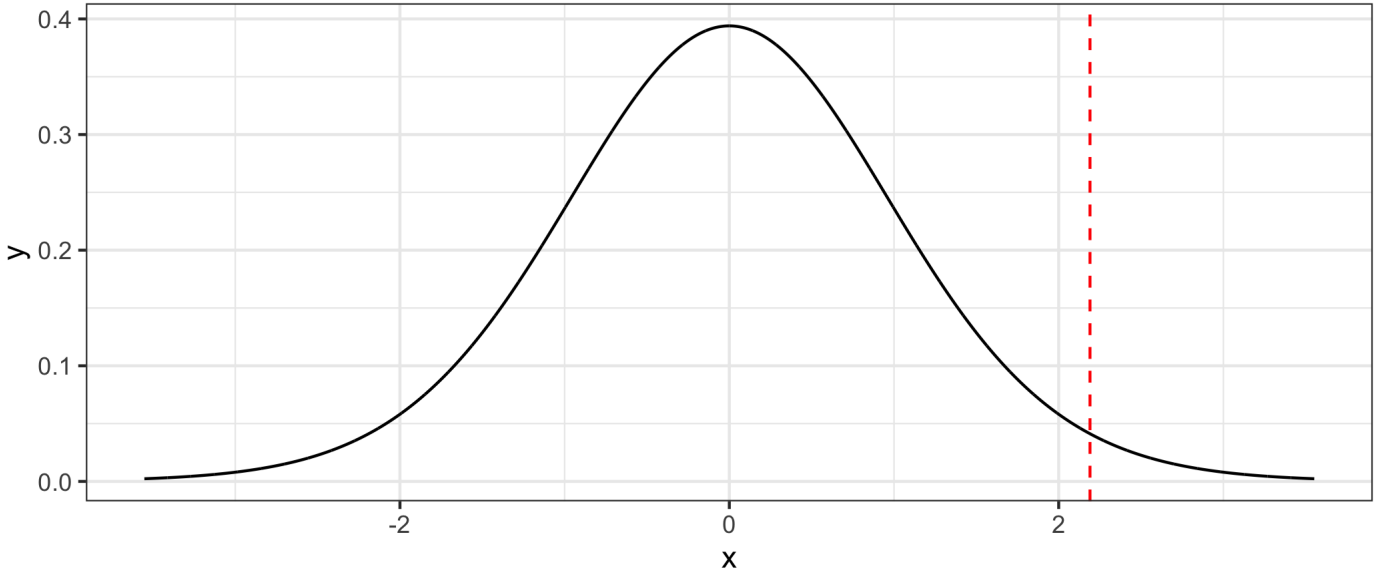
```
means_and_ns <- lizards %>%
  group_by(group) %>%
  summarize(means = mean(size),
            n = n())
means_and_ns
```

```
## # A tibble: 2 x 3
##   group means     n
##   <chr> <dbl> <int>
## 1 alive  23.2    12
## 2 dead   20.8    10
```

$$
\begin{aligned}
T_{\text{obs}} &= \frac{V - v_0}{\widehat{SD}(V)} \\
&= \frac{\bar{X}_{\text{alive}} - \bar{X}_{\text{dead}}}{s_p\sqrt{1/n_{\text{alive}} + 1/n_{\text{dead}}}} \\
&= \frac{23.1616667 - 20.785}{2.5322442\sqrt{1/12 + 1/10}} \\
&= 2.1920072
\end{aligned}
$$

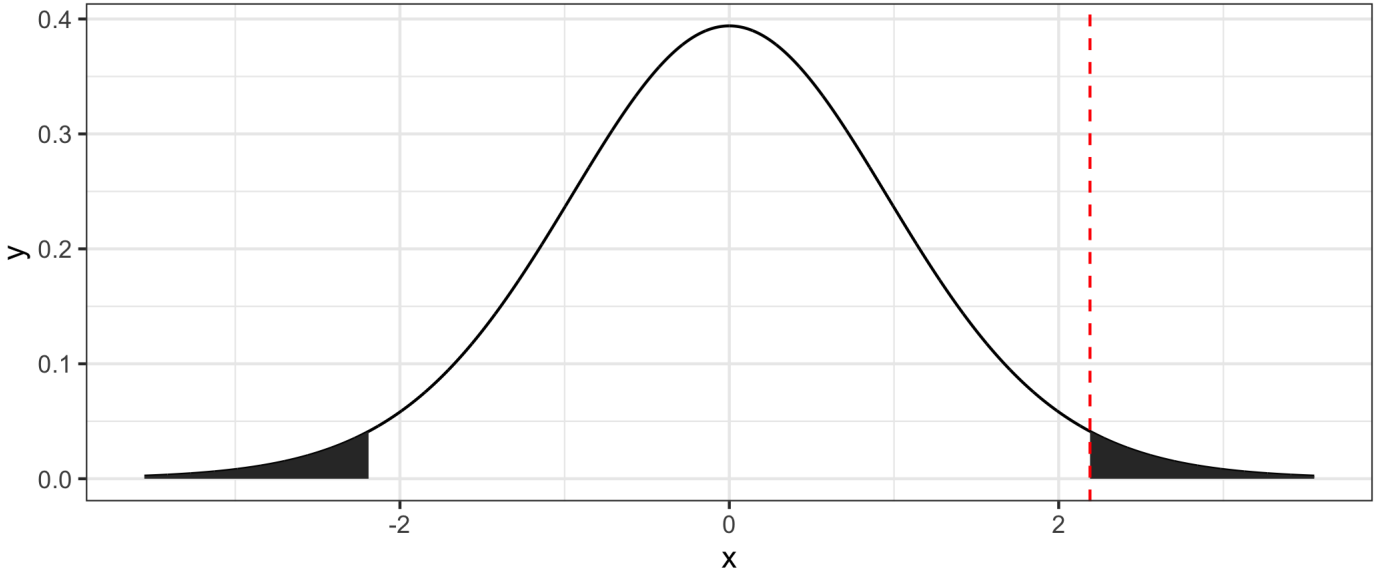
Two Independent Samples Hypothesis Test

We compare this to a $t_{n_{\text{alive}}+n_{\text{dead}}-2}$ distribution to find the p-value:



Two Independent Samples Hypothesis Test

We compare this to a $t_{n_{\text{alive}}+n_{\text{dead}}-2}$ distribution to find the p-value:




```
T_20 <- StudentsT(20)
2*(1 - cdf(T_20, 2.192))
```

```
## [1] 0.04038106
```

Two Independent Samples Hypothesis Test

```
t.test(data = lizards,
       size ~ group, var.equal = TRUE)
```

```
##
##     Two Sample t-test
##
## data:  size by group
## t = 2.192, df = 20, p-value = 0.04038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1149771 4.6383563
## sample estimates:
## mean in group alive  mean in group dead
##            23.16167            20.78500
```

Two Independent Samples Hypothesis Test

Which decade had better movies: the 90s of 2010s?

```
movies_orig <- read_csv('https://raw.githubusercontent.com/idc9/stor390/master/data/m
movies <- movies_orig %>%
  mutate(decade = as.numeric(str_sub(thtr_rel_year, start = 3, end = 3))*10)

DT::datatable(movies, options = list(pageLength = 3, scrollX = TRUE))
```

Show 3 entries Search:

	title	title_type	genre	runtime	mpaa_rating	studio	thtr_rel_year
1	Filly Brown	Feature Film	Drama	80	R	Indomina Media Inc.	2013
2	The Dish	Feature Film	Drama	101	PG-13	Warner Bros. Pictures	2001
3	Waiting for Guffman	Feature Film	Comedy	84	R	Sony Pictures Classics	1996

Showing 1 to 3 of 651 entries Previous 1 2 3 4 5 … 217 Next

Two Independent Samples Hypothesis Test

We'll use `critics_score` as our criteria for "good movie". This dataset contains a random sample of movies from IMDB.

We will test $H_0 : \mu_{90} = \mu_{10}$ against $H_A : \mu_{90} \neq \mu_{10}$ using $\alpha = 0.01$.

First, need to find out if we can assume equal variances:

```
movies %>%
  group_by(decade) %>%
  summarize(s = sd(critics_score))
```

```
## # A tibble: 5 x 2
##   decade     s
##    <dbl> <dbl>
## 1      0  28.4
## 2     10  27.1
## 3     70  25.3
## 4     80  28.0
## 5     90  28.8
```

We can!

Two Independent Samples Hypothesis Test

So, let's find means, s_p, and T_{obs}.

```
movies %>% filter(decade %in% c(90, 10))
  group_by(decade) %>%
  summarize(means = mean(critics_score),
            s = sd(critics_score),
            n = n())
```

```
## # A tibble: 2 x 4
##   decade means     s     n
##    <dbl> <dbl> <dbl> <int>
## 1     10  62.4  27.1   102
## 2     90  52.9  28.8   161
```

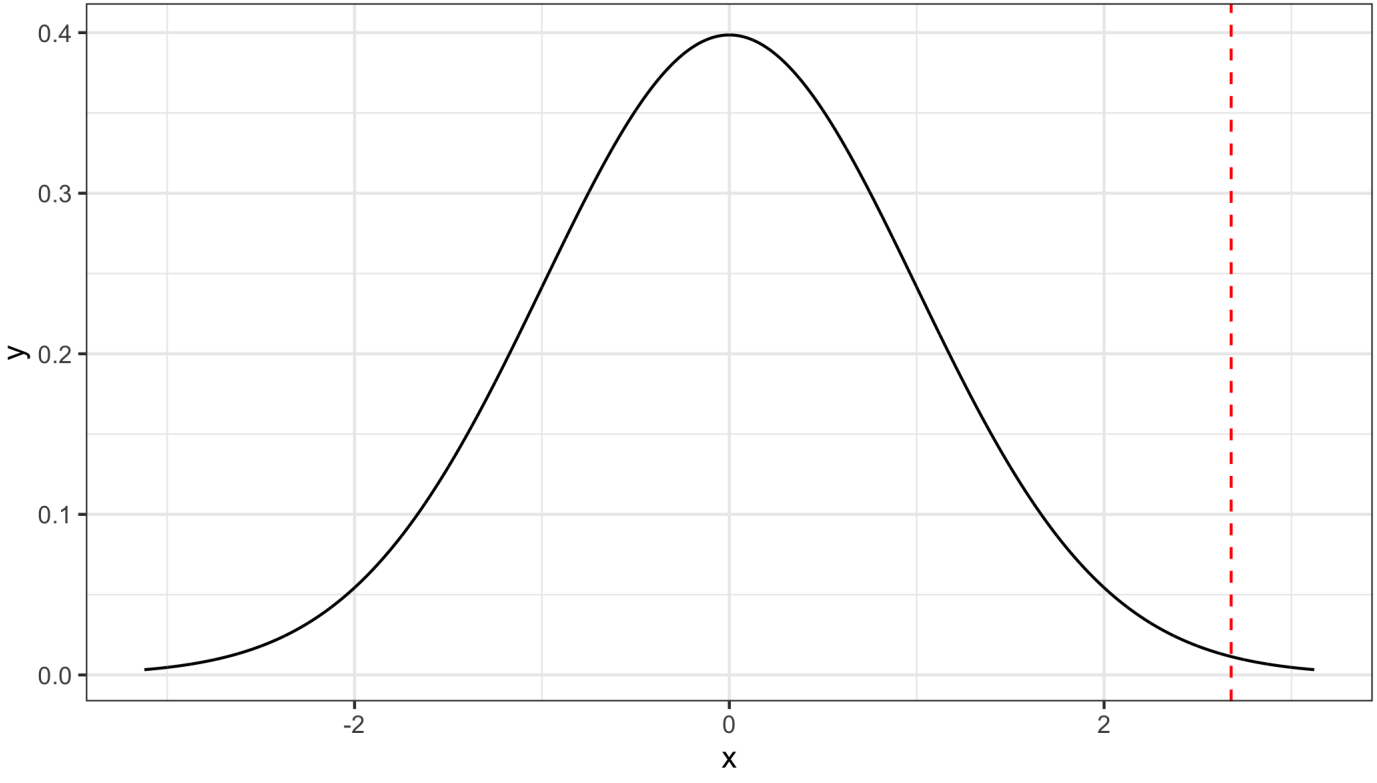
```
sp2 <- ((102 - 1)*27.1^2 + (161 - 1)*28.
sp2
```

```
## [1] 792.3129
```

$$
\begin{aligned}
T_{\text{obs}} &= \frac{V - v_0}{\widehat{SD}(V)} \\
&= \frac{\bar{x}_{10} - \bar{x}_{90}}{s_p\sqrt{1/n_{10} + 1/n_{90}}} \\
&= \frac{62.42 - 52.88}{\sqrt{792.31}\sqrt{1/102 + 1/161}} \\
&= 2.6781565
\end{aligned}
$$

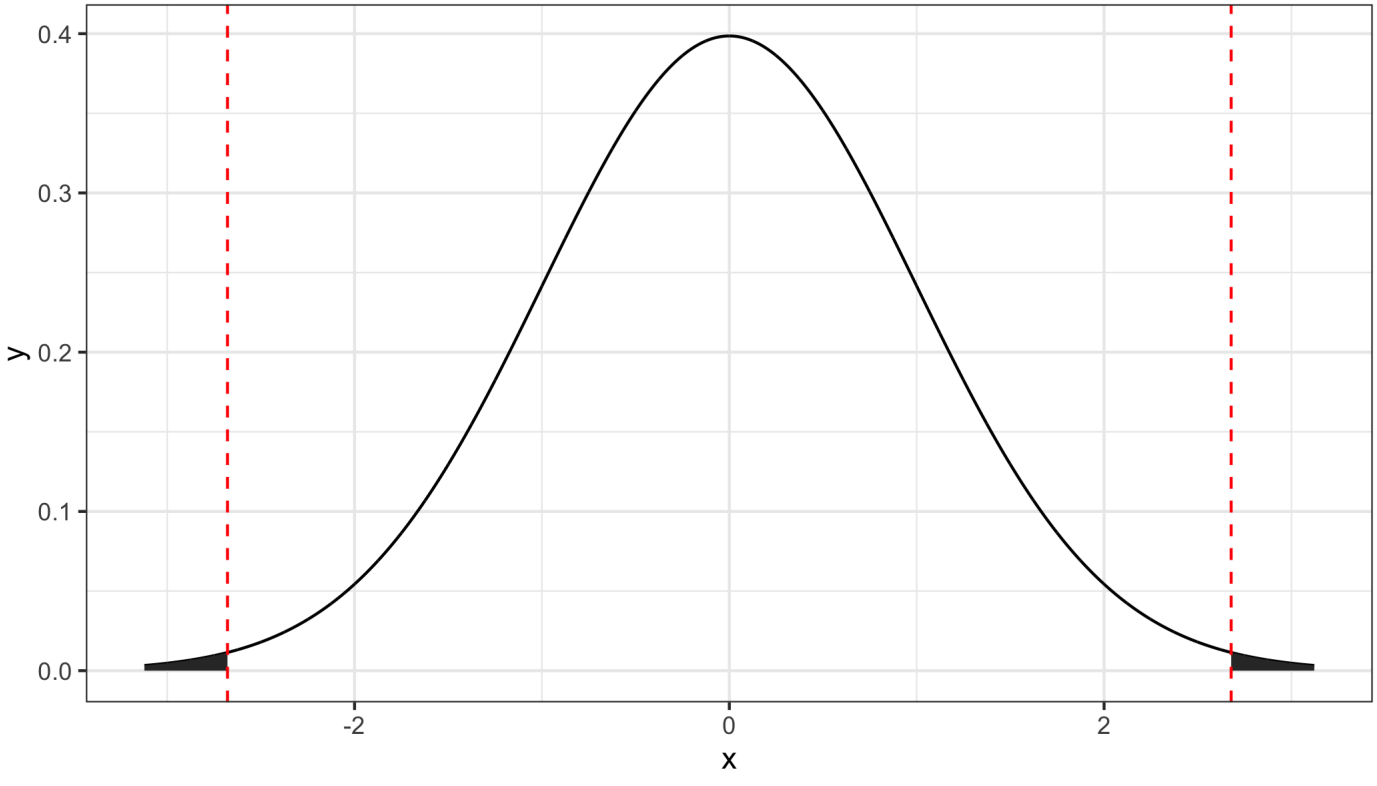
Two Independent Samples Hypothesis Test

If H_0 is true, $T \sim t_{n_{10}+n_{90}-2}$.



Two Independent Samples Hypothesis Test

Since we are testing against $H_A : \mu_{90} \neq \mu_{10}$, we need to consider both tails:



Two Independent Samples Hypothesis Test

```
T_261 <- StudentsT(261)

2*(1-cdf(T_261, 2.6781565))

## [1] 0.00787261
```

Since the p-value is smaller than α, we reject the null hypothesis. There seems to be a difference in mean critic score between 1990s movies and 2010s movies.

The quick way:

```
t.test(data = filter(movies, decade %in% c(10, 90)),
       critics_score ~ decade, var.equal = TRUE)
```

```
##
##     Two Sample t-test
##
## data:  critics_score by decade
## t = 2.6796, df = 261, p-value = 0.007839
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.531214 16.560371
## sample estimates:
## mean in group 10 mean in group 90
##         62.42157         52.87578
```

Two Independent Samples Hypothesis Test

Concrete used for roadways or buildings is often reinforced with a material that is placed inside the setting concrete. A common example of this is called 'rebar' which is short for 'reinforcing bar' and is usually made out of steel. You can often see rebar poking out of the concrete of demolished buildings. It is desireable that the reinforcing material is strong and corrosion resistant. Steel is strong, but tends to corrode over time, so experiments were conducted to test two corrosion resistant materials, one made of fiberglass and the other made of carbon.

Eight beams with fiberglass reinforcement, and 11 beams with carbon reinforcement were poured, and each was then subjected to a load test, which measures the strength of the beam. Strength is measured in kN (kiloNewtons), which is a measure of the force required to break the beam.

Research question: is there any difference in strength?

I.e., want to test $H_0 : \mu_{\text{fiber}} = \mu_{\text{carbon}}$ vs. $H_A : \mu_{\text{fiber}} \neq \mu_{\text{carbon}}$.

(Or, equivalently, $H_0 : \mu_{\text{fiber}} - \mu_{\text{carbon}} = 0$ vs. $H_0 : \mu_{\text{fiber}} - \mu_{\text{carbon}} \neq 0$.)

Two Independent Samples Hypothesis Test

```
DT::datatable(rebars, options = list(pageLength = 5))
```

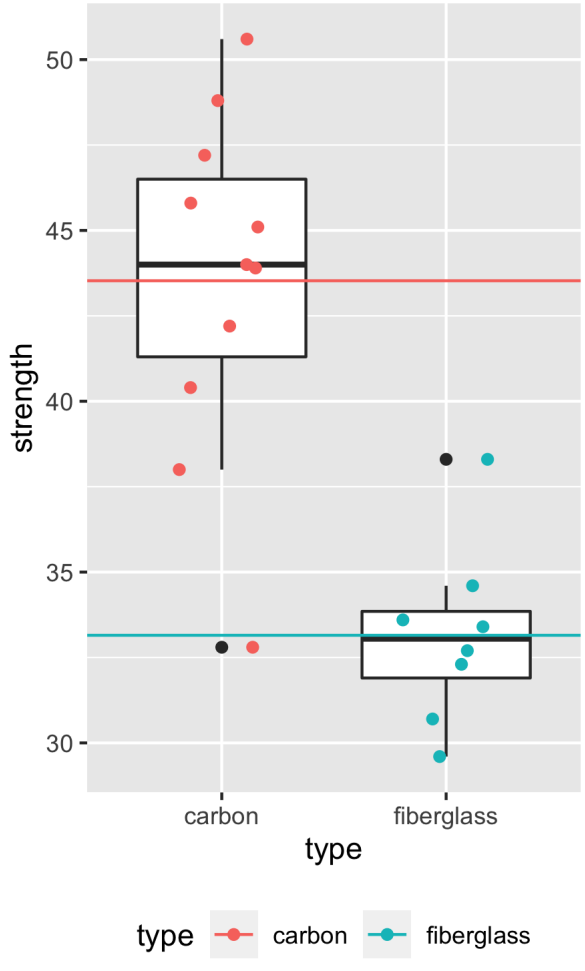
Show 5 entries Search:

	type	strength
1	fiberglass	38.3
2	fiberglass	29.6
3	fiberglass	33.4
4	fiberglass	33.6
5	fiberglass	30.7

Showing 1 to 5 of 19 entries Previous 1 2 3 4 Next

Two Independent Samples Hypothesis Test

```
ggplot(rebars,
       aes(x = type, y = strength)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, height = 0,
              aes(color = type)) +
  geom_hline(data = rebars %>%
               group_by(type) %>%
               summarize(means = mean(st
             aes(yintercept = means,
                 color = type)) +
  theme(legend.position = "bottom")
```



Two Independent Samples Hypothesis Test

First, need to decide if we can assume equal variances:

```
rebars %>%
  group_by(type) %>%
  summarize(s = sd(strength))
```

```
## # A tibble: 2 x 2
##   type            s
##   <chr>       <dbl>
## 1 carbon       5.06
## 2 fiberglass   2.63
```

Since the standard deviation of the strength of carbon bars is more than twice that of the fiberglass bars, we cannot assume that the variances are the same!

Two Independent Samples Hypothesis Test

The simple step first: find the observed test statistic:

```
rebars %>%
  group_by(type) %>%
  summarize(means = mean(strength),
            s = sd(strength),
            n = n()) %>% print() %>% # code for first table ends here
  ungroup() %>%
  summarize(T_obs = diff(means)/(sqrt(sum(s^2/n))))
```

```
## # A tibble: 2 x 4
##   type       means     s     n
##   <chr>      <dbl> <dbl> <int>
## 1 carbon      43.5  5.06    11
## 2 fiberglass  33.2  2.63     8

## # A tibble: 1 x 1
##   T_obs
##   <dbl>
## 1 -5.80
```

Two Independent Samples Hypothesis Test

"By hand":

$$
\begin{aligned}
T_{\text{obs}} &= \frac{V - v_0}{\widehat{\text{SD}}(V)} \\
&= \frac{\bar{x}_{\text{fiber}} - \bar{x}_{\text{carbon}}}{\sqrt{s^2_{\text{fiber}}/n_{\text{fiber}} + s^2_{\text{carbon}}/n_{\text{carbon}}}} \\
&= \frac{33.15 - 43.53}{\sqrt{2.63^2/8 + 5.06^2/11}} \\
&= -5.8096699
\end{aligned}
$$

Two Independent Samples Hypothesis Test

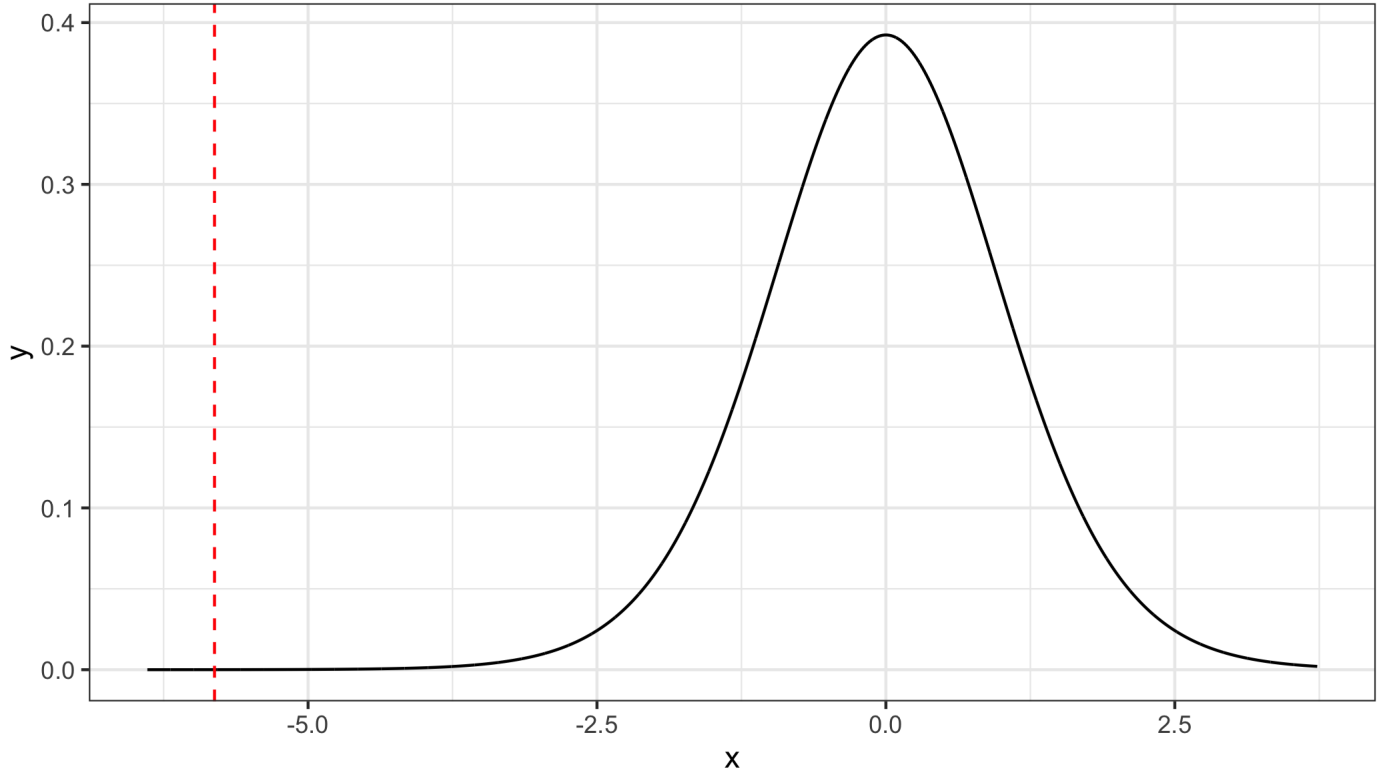
Next, we need to determine the degrees of freedom. Since we cannot assume equal variances, we need to use what is known as Satterthwaite's approximation:

$$\begin{aligned}\nu &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \\ &= \frac{\left(\frac{2.63^2}{8} + \frac{5.06^2}{11}\right)^2}{\frac{(2.63^2/8)^2}{8-1} + \frac{(5.06^2/11)^2}{11-1}} \\ &= 15.7119324\end{aligned}$$

When used to determine the correct df, we always round down. So, $\nu = 15$.

Two Independent Samples Hypothesis Test

So, if H_0 is true, $T \sim t_{15}$.



Two Independent Samples Hypothesis Test

Since $H_A : \mu_{\text{fiberglass}} \neq \mu_{\text{carbon}}$, we find the p-value as twice the area to the left of the observed value:

```
T_15 <- StudentsT(15)
2*cdf(T_15, -5.8096699)
```

```
## [1] 3.43928e-05
```

The p-value is very, very small, so we reject the null hypothesis. It seems that there is a difference in mean strength between the two materials.

Two Independent Samples Hypothesis Test

Summary

A Two Sample T-test with Independent Samples uses the test statistic $T = \frac{V - v_0}{\widehat{\text{SD}}(V)}$, where $V = \bar{X}_1 - \bar{X}_2$ and v_0 is the "null value" ($H_0 : V = v_0$; usually $v_0 = 0$).

If $0.5 < \frac{s_1}{s_2} < 2$, we assume equal variances $\sigma_1^2 = \sigma_2^2$. In this case,

- $\widehat{\text{SD}}(V) = s_p \sqrt{1/n_1 + 1/n_2}$, where

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

- if H_0 is true, $T \sim t_{n_1+n_2-2}$.

If we cannot assume equal variances,

- $\widehat{\text{SD}}(V) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$
- if H_0 is true, $T \sim t_\nu$ where

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

Two Independent Sample T Confidence Interval

As always, simply performing a test does not give us a lot of information. It simply answers the question "is this one value plausibly the true value?", where we would much rather ask "what values could be the true value?"

Fortunately, we can construct confidence intervals for the difference in means from the two sample t-test.

1. Find the distribution of $T = \frac{V - v_0}{\widehat{\text{SD}}(V)}$ **assuming** the null hypothesis is true.

2. Find cut-offs (i.e. quantiles) in said distribution such that we have α outside the cut-offs.

3. Create the confidence interval as $V_{\text{obs}} \pm t_{\text{df},\alpha/2}\widehat{\text{SD}}(V)$.

Two Independent Sample T Confidence Interval

Example: Fiber vs Carbon

We already established that the variances are not equal, and found $\nu = 15$. So, we can find a 90% CI as follows:

```
rebars %>%
  group_by(type) %>%
  summarize(mean = mean(strength),
            s = sd(strength),
            n = n())
```

```
## # A tibble: 2 x 4
##   type         mean     s     n
##   <chr>       <dbl> <dbl> <int>
## 1 carbon       43.5  5.06    11
## 2 fiberglass   33.2  2.63     8
```

```
quantile(T_15, 1-0.1/2)
```

```
## [1] 1.75305
```

$$
\begin{aligned}
&V_{\text{obs}} \pm t_{15,0.1/2}\widehat{\text{SD}}(V) \\
&= \bar{X}_{\text{fiber}} - \bar{X}_{\text{carbon}} \pm t_{15,0.1/2}\sqrt{\frac{s^2_{\text{fiber}}}{n_{\text{fiber}}} + \frac{s^2_{\text{carbon}}}{n_{\text{carbon}}}} \\
&= 33.15 - 43.53 \pm 1.75\sqrt{2.63^2/8 + 5.06^2/11} \\
&= -10.38 \pm 3.13
\end{aligned}
$$

So the `(90\%)` CI for the difference in means is `([-13.51, -7.25])`.

Two Independent Sample T Confidence Interval

Example: IMDB Movies

We saw that we can reasonably assume equal variances, and found the pooled variance. We can now find a 99% CI.

```
movies %>%
  filter(decade %in% c(90, 10)) %>%
  group_by(decade) %>%
  summarize(means = mean(critics_score),
            s = sd(critics_score),
            n = n())
```

```
## # A tibble: 2 x 4
##   decade means     s     n
##    <dbl> <dbl> <dbl> <int>
## 1     10  62.4  27.1   102
## 2     90  52.9  28.8   161
```

```
((102-1)*27.1^2 + (161-1)*28.8^2)/(102 +
```

```
## [1] 792.6659
```

```
quantile(T_261, 1 - 0.01/2)
```

```
## [1] 2.594797
```

$$
\begin{aligned}
&V_{\text{obs}} \pm t_{261,0.01/2}\widehat{\text{SD}}(V) \\
&\quad = \bar{X}_{10} - \bar{X}_{90} \pm t_{261,0.01/2}\sqrt{s_p^2\left(\frac{1}{n_{10}} + \frac{1}{n_{90}}\right)} \\
&\quad = 62.4 - 52.9 \pm 2.59\sqrt{792.31 \cdot (1/102 + 1/161)} \\
&\quad = 9.5 \pm 9.23
\end{aligned}
$$

So the `(99\%)` CI for the difference in means is `([0.27, 18.73])`.