Lecture 15: Two Sample Hypothesis Tests

STAT 324

Ralph Trane University of Wisconsin–Madison

Spring 2020





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.



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

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

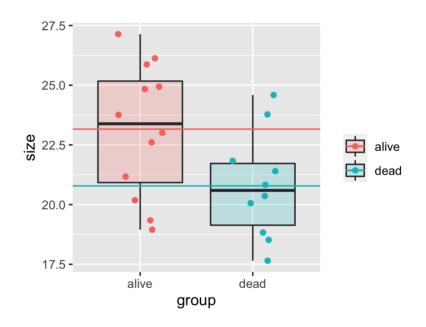
Some data:





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

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



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



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

We can rephrase the question in terms of hypotheses:

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

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

How would we go about answering this question?



IF $ar{X}_{
m alive} \sim N$ and $ar{X}_{
m dead} \sim N$, then $ar{X}_{
m alive} - ar{X}_{
m dead} \sim N$.

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

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

IF the two samples are independent of each other, $ar{X}_{
m alive}$ is independent of $ar{X}_{
m dead}$, so

$$ext{Var}(ar{X}_{ ext{alive}} - ar{X}_{ ext{dead}}) = ext{Var}(ar{X}_{ ext{alive}}) + ext{Var}(ar{X}_{ ext{dead}}) = rac{\sigma_{ ext{alive}}^2}{n_{ ext{alive}}} + rac{\sigma_{ ext{dead}}^2}{n_{ ext{dead}}}$$

So, **IF**
$$H_0$$
 is true, then $ar{X}_{
m alive} - ar{X}_{
m dead} \sim N\left(0,rac{\sigma_{
m alive}^2}{n_{
m alive}} + rac{\sigma_{
m dead}^2}{n_{
m dead}}
ight)$.



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

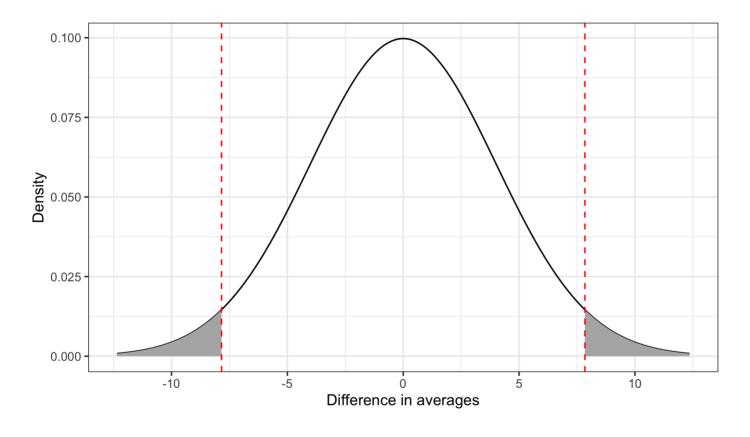
By finding the probability of observing something more extreme if we were to repeat the experiment, **assuming the null hypothesis is true**!

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

- look at the curve that is the distribution of the difference $ar{X}_{
 m alive} ar{X}_{
 m dead}$, i.e. $N\left(0,rac{\sigma_{
 m alive}^2}{n_{
 m alive}} + rac{\sigma_{
 m dead}^2}{n_{
 m dead}}
 ight)$.
- using quantiles:
 - \circ find quantiles that cut-off lpha/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"
 - \circ reject if smaller than α

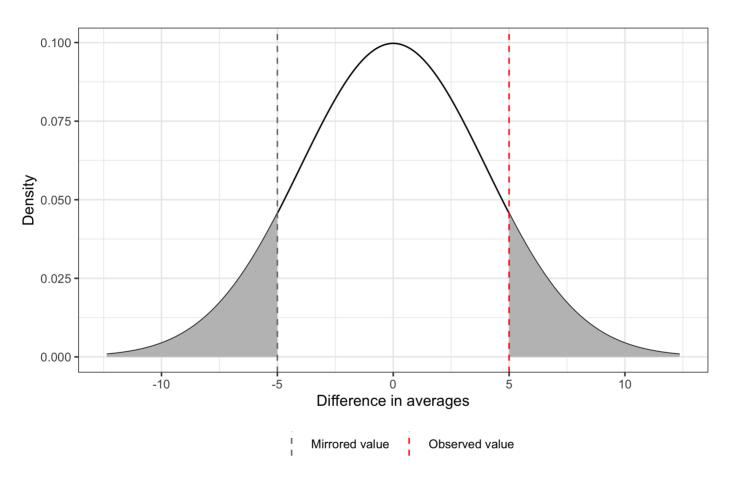


Using quantiles: reject if outside of dotted lines that cut-off lpha/2 on each side.





Using p-value: reject if area outside dotted lines is smaller than lpha





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

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

In the two sample case, we will use

$$T = rac{V - v_0}{\widehat{ ext{SD}}(V)},$$

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

$$\widehat{ ext{SD}}(V) = \sqrt{s_{ ext{alive}}^2/n_{ ext{alive}} + s_{ ext{dead}}^2/n_{ ext{dead}}}$$

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



In general, two scenarios:

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

When this is the case, we replace both by a common number, $\sigma_{\rm pooled}^2$ (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

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



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

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 ${
m Var}(V)=rac{s_p^2}{n_1}+rac{s_p^2}{n_2}=s_p^2(1/n_1+1/n_2)$, and our test statistic will follow a $t_{n_1+n_2-2}$ distribution:

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



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

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

where

$$u = rac{\left(rac{s_1^2}{n_1} + rac{s_2^2}{n_2}
ight)^2}{rac{(s_1^2/n_1)^2}{n_1-1} + rac{(s_2^2/n_2)^2}{n_2-1}}$$

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



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 $0.5<rac{s_1}{s_2}<2$, then we are okay with assuming $\sigma_1^2=\sigma_2^2$.

In our lizards example:

```
## # 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 $\sigma_{
m dead}^2 = \sigma_{
m alive}^2$.



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</pre>
```

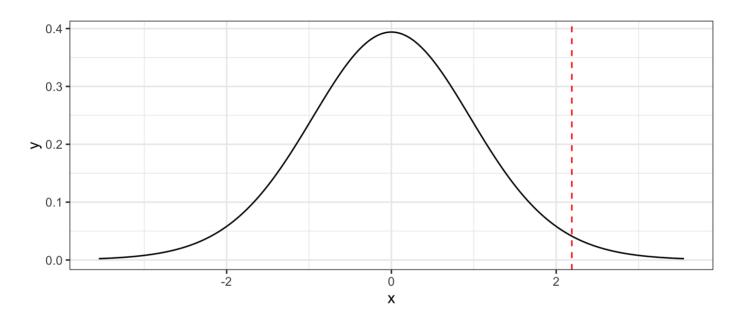
We can now calculate the observed value of the test statistic:

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

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

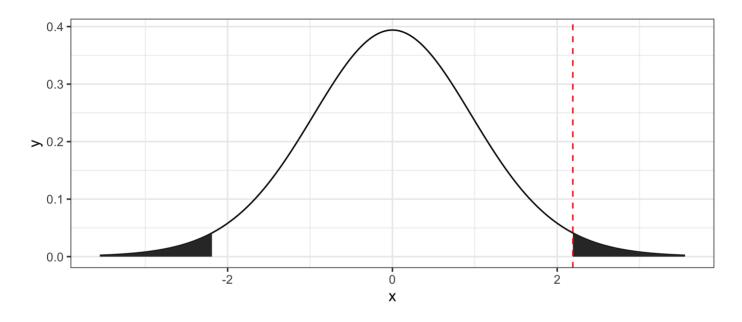


We compare this to a $t_{n_{
m alive}+n_{
m dead}-2}$ distribution to find the p-value:





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



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

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



```
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
```



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

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

DT::datatable(movies, options = list(pageLength = 4, scrollX = TRUE))
```

Show 4 coentries Search:										
	title 🛊	title_type 🖣	genre 🖣	runtime 🖣	mpaa_rating +	studio 🛊	thtr_rel_year			
1	Filly Brown	Feature Film	Drama	80	R	Indomina Media Inc.	201			
2	The Dish	Feature Film	Drama	101	PG-13	Warner Bros. Pictures	200			
3	Waiting for Guffman	Feature Film	Comedy	84	R	Sony Pictures Classics	199			



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 $H_A: \mu_{90}
eq \mu_{10}$ using lpha=0.01.

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

We can!

80 28.0

90 28.8

4

5



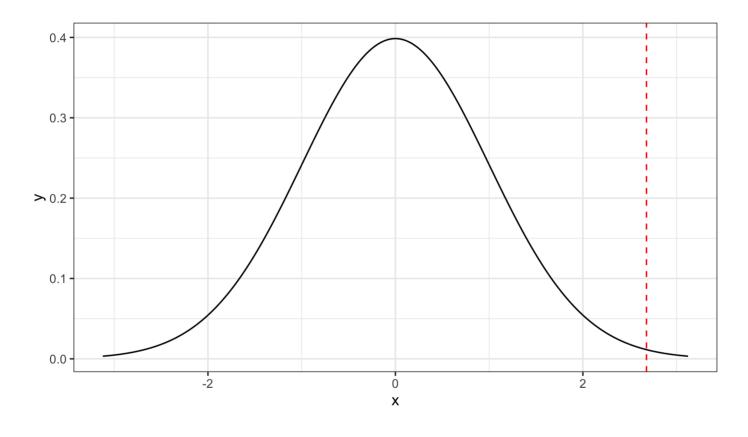
So, let's find means, s_p , and $T_{
m obs}$.

```
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
##
   <dbl> <dbl> <dbl> <int>
##
        10 62.4 27.1
## 1
                           102
## 2 90 52.9 28.8
                           161
sp2 \leftarrow ((102 - 1) \times 27.1^2 + (161 - 1) \times 28.
sp2
## [1] 792.3129
```

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

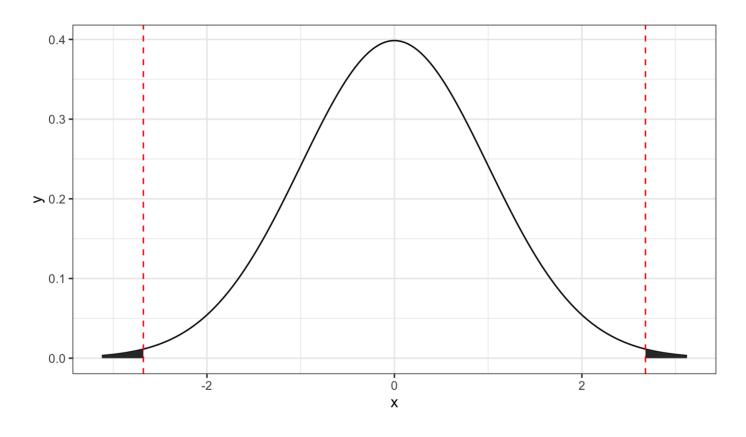


If H_0 is true, $T \sim t_{n_{10}+n_{90}-2}.$





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





```
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
```



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?

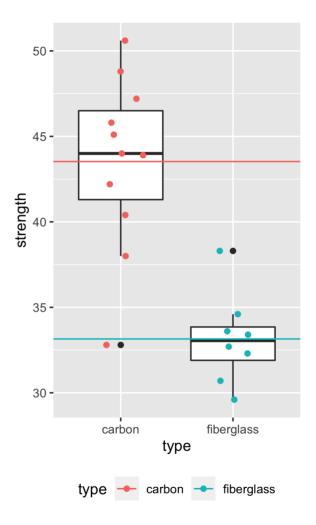
I.e., want to test $H_0: \mu_{ ext{fiber}} = \mu_{ ext{carbon}}$ vs. $H_A: \mu_{ ext{fiber}}
eq \mu_{ ext{carbon}}$.

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



<pre>DT::datatable(rebars, options = list(pageLength = 5))</pre>												
Show 5 entries Search:												
		type						st	rength			
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				







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

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!



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
```



"By hand":

$$egin{aligned} T_{
m obs} &= rac{V - v_0}{\widehat{
m SD}(V)} \ &= rac{ar{x}_{
m fiber} - ar{x}_{
m carbon}}{\sqrt{s_{
m fiber}^2/n_{
m fiber} + s_{
m carbon}^2/n_{
m carbon}}} \ &= rac{33.15 - 43.53}{\sqrt{2.63^2/8 + 5.06^2/11}} \ &= -5.8096699 \end{aligned}$$



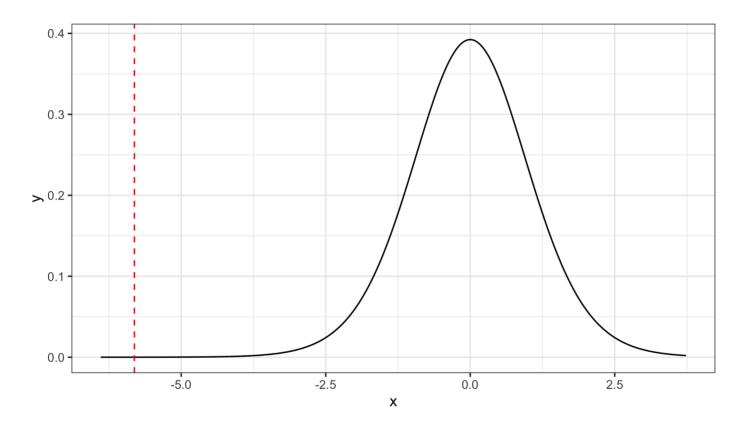
Next, we need to determine the degrees of freedom. Since we cannot assume equal variances, we need to use what is known as Satterthwaite's approximation:

$$u = rac{\left(rac{s_1^2}{n_1} + rac{s_2^2}{n_2}
ight)^2}{rac{(s_1^2/n_1)^2}{n_1 - 1} + rac{(s_2^2/n_2)^2}{n_2 - 1}} \ = rac{\left(rac{2.63^2}{8} + rac{5.06^2}{11}
ight)^2}{rac{(2.63^2/8)^2}{8 - 1} + rac{(5.06^2/11)^2}{11 - 1}} \ = 15.7119324$$

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



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





Since $H_A: \mu_{\mathrm{fiberglass}} \neq \mu_{\mathrm{carbon}}$, 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.



Summary

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

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

$$oldsymbol{oldsymbol{oldsymbol{s}}} oldsymbol{oldsymbol{s}} \widehat{\mathrm{SD}}(V) = s_p \sqrt{1/n_1 + 1/n_2}, ext{where}$$
 $s_p^2 = rac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

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

If we cannot assume equal variances,

•
$$\widehat{\mathrm{SD}}(V) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

• if H_0 is true, $T\sim t_
u$ where

$$u = rac{\left(rac{s_1^2}{n_1} + rac{s_2^2}{n_2}
ight)^2}{rac{(s_1^2/n_1)^2}{n_1 - 1} + rac{(s_2^2/n_2)^2}{n_2 - 1}}$$