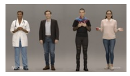
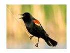
One- and two-sample hypothesis testing

One sample
- One sample (frogs) according to a single categorical variable (Left/Right)
  Binomial test

- One sample (humans) according to a single quantitative variable (temperature)
  One-sample t-test



Two samples
- Paired-design
  Paired t-test



- independent-design
  Two-sample t-test (equal variance)

- independent-design
  Two-sample t-test (unequal variance) [TODAY]

Multiple samples
- independent-design
  Analysis of Variance (equal variance) [coming soon]

1

Two-sample comparison of means

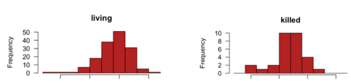
Assumptions:

- Each of the two samples is a random sample from their population.

- The variable (e.g., horn length) is normally distributed for each population.

- The standard deviation (and variance) of the variable is the same in both populations.

- The theoretical sampling distribution (i.e., assumed under the null hypothesis) of the differences between sample means is t-distributed only if the samples come from theoretical populations with the same variance (the theoretical populations have the same mean, i.e., assumed under the null hypothesis but not necessarily the same variance).



Loggerhead shrike

Horned lizard

2

Where does the assumption of equal variances for the t-distribution come from?

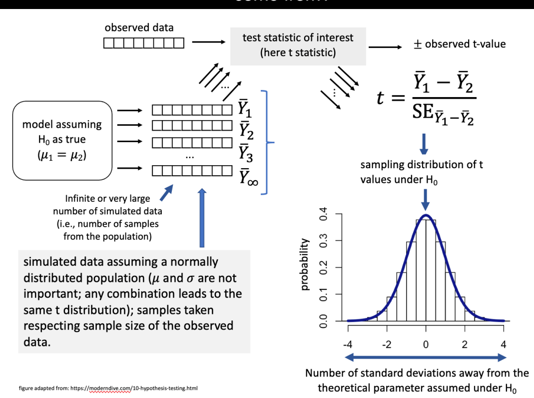


figure adapted from: https://moderndive.com/10-hypothesis-testing.html

3

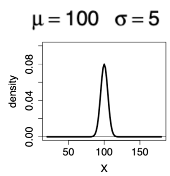
Two-sample t test when sample variances are different

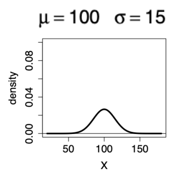
Two normally distributed populations with the same mean ($\mu = 100$) but different standard deviations ($\sigma = 5, \sigma = 15$).

So, by setting an alpha = 0.05 (i.e., the probability of rejecting the null hypothesis when is in reality true, i.e., risk rate of false positives), if we were to take 100 samples means from each of these two populations and conduct a t-test to assess their differences, we would expect that only 5 (5%) of them would be significant.

But when the null hypothesis is true (equal μ) but variances (standard deviations) are different, then the risk of false positives are higher than the alpha pre-established (i.e., the chance of type I error increases - we say that it gets inflated).

We then say that the standard t-test for the differences between two sample means are not robust against heteroscedasticity (meaning differences in variances).





4

How to know if our alpha levels will hold true?
We need test whether variances differ or not:
Two-sample comparison of variances

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

H_0: Lizards killed by shrikes and living lizard *do not differ* in their horn length variances (i.e., $\sigma_1^2 = \sigma_2^2$).

H_A: Lizards killed by shrikes and living lizard *differ* in their horn length variances (i.e., $\sigma_1^2 \neq \sigma_2^2$).

5

Intuition underlying a two-sample test of variances

Assume that the null hypothesis is true (i.e., $\sigma_1^2 = \sigma_2^2$).

Conduct infinite sampling (or computationally large number of samples) from a population that have the same variances (doesn't matter whether they have the same population means as they don't affect the variance).

Each sample should have the appropriate sample size (living lizards = 154 observations and killed lizards = 30 observations).

For each pair of samples calculate the ratio of the two variances.

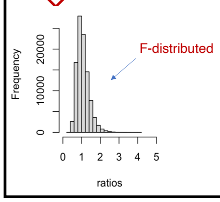
The sampling distribution of all possible variance ratios assuming our null hypothesis true will serve as the distribution in which we can compare our sample values against.

That sampling distribution is called the F-distribution.

6

Intuition underlying a two-sample test of variances – their ratios are F-distributed

```
samples.n154 <- replicate(100000, rnorm(n=154,mean=350,sd=100))
samples.n30 <- replicate(100000, rnorm(n=30,mean=10,sd=100))

variances.n154 <- apply(X=samples.n154,MARGIN=2,FUN=var)
variances.n30 <- apply(X=samples.n30,MARGIN=2,FUN=var)

ratios <- variances.n154/variances.n30

hist(ratios)
```



$\mu_1 = 350 \quad \mu_1 = 10$

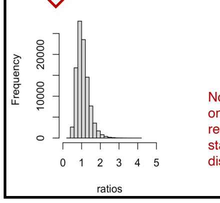
$\sigma_1 = 100 \quad \sigma_2 = 100$

Remember that the test is about variances, so we are assuming under H_0 that they are equal.

7

Let's change the population parameters

```
samples.n154 <- replicate(100000, rnorm(n=154,mean=8,sd=7.2))
samples.n30 <- replicate(100000, rnorm(n=30,mean=4,sd=7.2))

variances.n154 <- apply(X=samples.n154,MARGIN=2,FUN=var)
variances.n30 <- apply(X=samples.n30,MARGIN=2,FUN=var)

ratios <- variances.n154/variances.n30

hist(ratios,xlim=c(0,5))
```



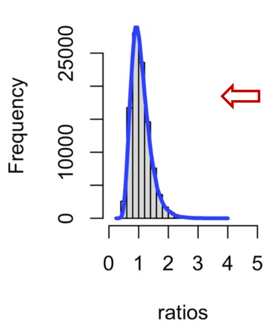
$\mu_1 = 8 \quad \mu_1 = 4$

$\sigma_1 = 7.2 \quad \sigma_2 = 7.2$

Note how the previous sampling distribution is the same as the one here. So the distribution is constant assuming H_0 is true regardless of the parameters of the populations (mean and standard deviation). So we can use it as a universal distribution for testing H_0 of homoscedasticity.

8

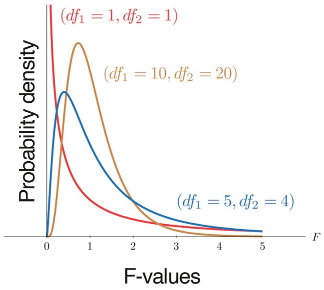
The sampling distribution of two sample ratios assuming H_0 as true follows the F-distribution




```
samples.n154 <- replicate(100000, rnorm(n=154,mean=8,sd=7.2))
samples.n30 <- replicate(100000, rnorm(n=30,mean=4,sd=7.2))

variances.n154 <- apply(X=samples.n154,MARGIN=2,FUN=var)
variances.n30 <- apply(X=samples.n30,MARGIN=2,FUN=var)

ratios <- variances.n154/variances.n30

hist(ratios,xlim=c(0,5))
```

9

The F-test for variance ratios
(also referred as to test of homogeneity of variances)



$$F = \frac{s_1^2}{s_2^2} \quad \begin{matrix} \longrightarrow df_1 \\ \longrightarrow df_2 \end{matrix}$$

Note that the F-distribution changes with the sample size (df) of the numerator (here s_1^2) and denominator (here s_2^2)

10

Two-sample comparison of variances

The F-test for variance ratios (also referred as to homogeneity of variance)

H_0: Lizards killed by shrikes and living lizard *do not differ* in their horn length variances (i.e., $\sigma_1^2 = \sigma_2^2$).

H_A: Lizards killed by shrikes and living lizard *differ* in their horn length variances (i.e., $\sigma_1^2 \neq \sigma_2^2$).

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

11

Two-sample comparison of variances

The F-test for variance ratios (also referred as to homogeneity of variance)

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30

$$F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{2.71^2}{2.63^2} = 1.06$$

Largest variance

Smallest variance

Degrees of freedom (numerator) = 30 - 1= 29

Degrees of freedom (denominator) = 154 - 1= 153

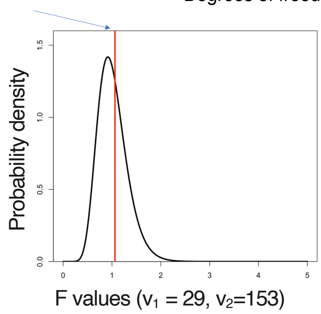
Because the F-distribution is asymmetric, we set it up as the largest variance divided by the smallest; which has a slightly different P-value than if we divide the smallest by the largest variance.

12

The F-test for variance ratios (also referred as to homogeneity of variance)

$$F = 1.06$$

Degrees of freedom (numerator) = 29 (v_1)
Degrees of freedom (denominator) = 153 (v_2)



$\Pr[F > 1.06] = 0.3916$
$2 \text{ x } \Pr[F > 1.06] = \mathbf{0.7832}$

Multiplying the p-value by 2, makes the F test two-tailed. Because the F-distribution is asymmetric, there are other ways to calculate P-values. We will keep it simple here and simply multiply by 2.

Statistical decision based on alpha = 0.05: ***do not reject H_0***

13

$$F = 1.061762$$

Degrees of freedom (numerator) = 29 (v_1)
Degrees of freedom (denominator) = 153 (v_2)

```
> pf(1.061762, 29, 153, lower.tail = FALSE)
[1] 0.3916306
```

$\Pr[F > 1.06] = 0.3916$
$2 \text{ x } \Pr[F > 1.06] = \mathbf{0.7832}$

14

The F-test for variance ratios (also referred as to homogeneity of variance)

H_0: Lizards killed by shrikes and living lizard *do not differ* in their horn length variances (i.e., $\sigma_1^2 = \sigma_2^2$).

H_A: Lizards killed by shrikes and living lizard *differ* in their horn length variances (i.e., $\sigma_1^2 \neq \sigma_2^2$).

$$F = 1.06$$
$$2 \Pr[F > 1.06] = \mathbf{0.7832}$$

Decision based on alpha = 0.05: ***do not reject H_0***

Conclusion – We have no evidence to reject the H_0 that the variances are different. Therefore, use the two standard sample t-test for these data as the assumption of equality of variances is met!

15

Two-sample comparison of variances

The F-test for variance ratios (also referred as to homogeneity of variance)

Assumptions:

- Each of the two samples is a random sample from its population.

- The variable (e.g., horn length) is normally distributed in each population.

```
samples.n154 <- replicate(100000, rnorm(n=154,mean=8,sd=7.2))
samples.n30 <- replicate(100000, rnorm(n=30,mean=4,sd=7.2))

variances.n154 <- apply(X=samples.n154,MARGIN=2,FUN=var)
variances.n30 <- apply(X=samples.n30,MARGIN=2,FUN=var)

ratios <- variances.n154/variances.n30

hist(ratios,xlim=c(0,5))
```

16

Let’s take a break – 2 minutes

17

A study in which the variance of the two samples differ and the need to apply a different type of t-test for comparing two sample means, the so called Welch’s t-test

Heteroscedasticity (differences in sample variances) is not an issue for the paired t-test because it is basically a single sample of differences).

18

A study in which the variance of the two samples differ

- Biodiversity is threatened by alien species.
- Alien species from outside their natural range may do well because they have fewer predators or parasites in the new area.
- Brook trout is a species native to eastern North America that has been introduced into streams in the West for sport fishing.
- Biologists followed the survivorship of a native species, chinook salmon, released in a series of 12 streams that either had brook trout introduced or did not (Levin et al. 2002).

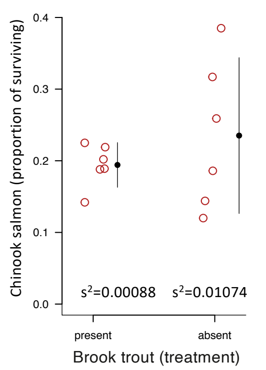
Research question: Does the presence of brook trout affect the survivorship of salmon?

19

A study in which the variance of the two samples differ

Research question: Does the presence of brook trout affect the survivorship of salmon?

Note the differences in variances



20

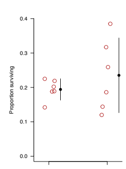
Two-sample comparison of variances

Research question: Does the presence of brook trout affect the survivorship of the salmon?

We need first to test the differences in variance to determine which type of t test we should use. If variances are different we can't use the standard t test but rather the Welch's t-test.

H_0: The variance of the proportion of chinook surviving is the same in streams with and without brook trout (i.e., $\sigma_1^2 = \sigma_2^2$).

H_A: The variance of the proportion of chinook surviving differs in streams with and without brook trout(i.e., $\sigma_1^2 \neq \sigma_2^2$).



21

Two-sample comparison of variances

Research question: Does the presence of brook trout affect the survivorship of the salmon?

We need first to test the differences in variance to determine which type of t test we should use. If variances are different we can't use the standard t test but rather the Welch's t-test.

$$\mathrm{F} = \frac{\sigma_1^2}{\sigma_2^2} = \frac{0.01074}{0.00088} = 12.17$$

Largest variance

Smallest variance

Degrees of freedom (numerator) = 6 - 1= 5

Degrees of freedom (denominator) = 6 - 1= 5

$$\Pr[\mathrm{F} > 12.17] = 0.007945$$
$$2 \Pr[\mathrm{F} > 12.17] = \mathbf{0.01589}$$

Decision based on alpha = 0.05: ***reject H_0 in favour of H_A.***

22

Two-sample comparison of variances

H_0: The variance of the proportion of chinook surviving is the same in streams with and without brook trout (i.e., $\sigma_1^2 = \sigma_2^2$).

H_A: The variance of the proportion of chinook surviving differs in streams with and without brook trout(i.e., $\sigma_1^2 \neq \sigma_2^2$).

$$2 \Pr[\mathrm{F} > 12.17] = \mathbf{0.01589}$$

Decision based on alpha = 0.05: ***reject H_0 in favour of H_A.***

23

Welch's t-test: comparing two sample means when their variances are different

Since variances are different we need to use the the Welch's t-test to test for differences between the two treatments (samples)

H_0: The mean proportion of chinook surviving is the same in streams with and without brook trout (i.e., $\mu_1 = \mu_2$).

H_A: The mean proportion of chinook surviving differs in streams with and without brook trout(i.e., $\mu_1 \neq \mu_2$).

Group	Sample mean	Variance	Sample size
Brook trout present	0.194	0.00088	6
Brook trout absent	0.235	0.01074	6

24

Welch's t-test: comparing two sample means when their variances are significantly different

The Welch's test is not based same t test statistic as the standard t-test for two sample means. The standard error is not based on the pooled variances (weighted variances by their sample sizes)

$$t = \frac{(Y_1 - Y_2)}{\mathrm{SE}_{Y_1 - Y_2}} \quad \begin{cases} \mathrm{SE}_{Y_1 - Y_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ s_p^2 = \dfrac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} \end{cases}$$

Standard t-test for comparing two-sample means

$$t = \frac{(Y_1 - Y_2)}{\mathrm{SE}_{Y_1 - Y_2}} \quad \begin{cases} \mathrm{SE}_{Y_1 - Y_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \end{cases}$$

Welch's modified t-test

25

Welch's t-test: comparing two sample means when their variances are significantly different

And the degrees of freedom for the Welch's test is also calculated in a more complex way.

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

Group	Sample mean	Variance	Sample size
1) Brook trout present	0.194	0.00088	6
2) Brook trout absent	0.235	0.01074	6

26

Welch's t-test: comparing two sample means when their variances are significantly different

The Welch's test t statistic is then:

$$t = \frac{0.194 - 0.235}{\sqrt{\frac{0.00088}{6} + \frac{0.01704}{6}}} = 0.93148$$

Group	Sample mean	Variance	Sample size
1) Brook trout present	0.194	0.00088	6
2) Brook trout absent	0.235	0.01074	6

27

Differences in degrees of freedom between the standard t-test and the modified Welch's test test for comparing two sample means that are heteroscedastic.

$$df_{Welch} = \frac{\frac{1}{6} + \frac{\frac{0.01704}{0.00088}}{6}}{\frac{1}{36(6-1)} + \frac{\left(\frac{0.01704}{0.00088}\right)^2}{36(6-1)}} = 5.8165$$

$$df_{standard\ t-test} = (6-1) + (6-1) = 10$$

$$t = 0.93148$$

Group	Sample mean	Variance	Sample size
1) Brook trout present	0.194	0.00088	6
2) Brook trout absent	0.235	0.01074	6

28

Remember from an early slide in this lecture:

But when the null hypothesis is true (equal μ) but variances (standard deviations) are different, then the risk of false positives are higher than the alpha pre-established.

We then say that the standard t-test for the differences between two sample means are not robust against heteroscedasticity (meaning differences in variances).

By having a smaller degrees of freedom, the p-value for the Welch's test will be greater than the standard t-test.

As such, the Welch's test adjust the p-value making it harder (bigger) to reject the null hypothesis, thus making the risk of committing a type I error (false positive) the same as the original pre-determined alpha (significance level).

29

Welch's t-test: comparing two sample means when their variances are significantly different

$t = 0.93148$

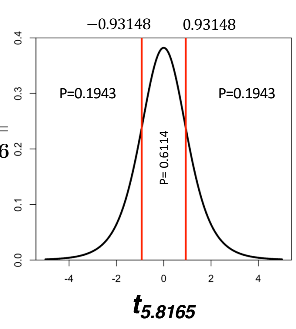
$df = 5.8165$

two tailed t-test

$\Pr[t < -0.931] + \Pr[t > 0.931] =$
$2 \times \Pr[t > \text{abs}(0.931)] = \mathbf{0.3886}$

Decision based on alpha = 0.05: ***do not reject H_0***

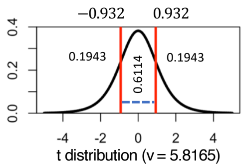
Conclusion: We lack evidence to state that the mean proportion of chinook surviving differs in streams with and without brook trout (i.e., $\mu_1 \neq \mu_2$).



30

standard t-test (equal variances) *versus* the Welch's t-test (different variances)

By the Welch's test having smaller degrees of freedom, it makes it harder to reject the null hypothesis thus making the risk of false positives to be equal to alpha.

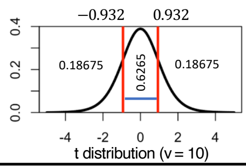


df = 5.8165 (**Welch's t test**)

$t = 0.93148$

$2\ \mathrm{Pr}[\mathrm{t} > \mathrm{abs}(0.931)] = \mathbf{0.3886}$

2 x 0.1943



df = *10* (**regular t-test**)

$t = 0.93148$

$2\ \mathrm{Pr}[\mathrm{t} > \mathrm{abs}(0.931)] = \mathbf{0.3735}$

$n_1 = 6$ $n_2 = 6$

2 x 0.18675

31

Welch's t-test: comparing two sample means when their variances are significantly different

Assumptions:

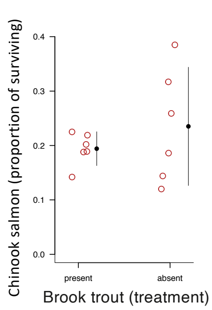
- Each of the two samples is a random sample from their populations.

- The variable (e.g., horn length, proportion of survival) is normally distributed in each population.

32

Welch's t-test: comparing two sample means when their variances are significantly different

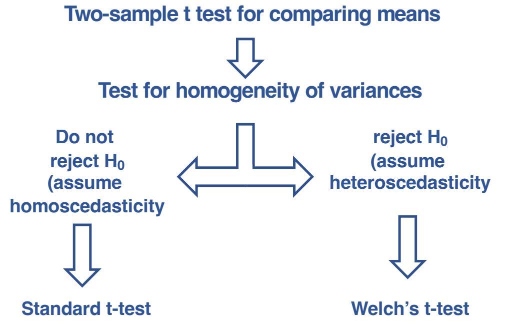
Conclusion: We lack evidence to state that the mean proportion of chinook surviving differs in streams with and without brook trout (i.e., $\mu_1 \neq \mu_2$).



Relevant issues:

- Perhaps sample size is too small to detect real differences.
- The difference in variances reduce the degrees of freedom and reduces statistical power too much.
- Even if the means are truly not different (i.e., H_0 is true; something we don't know), it is important to note the variances are statistically different, which cab have important implications (e.g., conservation).

33



Note that I used "assume" homoscedasticity or "assumed" heteroscedasticity because we don't know if in reality if the variances of the two samples are truly different. All we have is evidence from the F-test that reject or do not reject the null hypothesis.

34