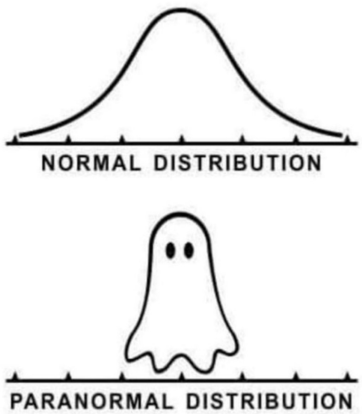


1

Distinguishing "Non-Significant" vs. "Insignificant" in Statistics

Non-Significant:
- Indicates the result does not reach the threshold for statistical significance (e.g., $p > 0.05$).
- Means there's not enough evidence to reject the null hypothesis within the set confidence level (alpha).
- Does *not* imply the absence of an effect, only that it's not statistically detectable in the sample (Type II error).

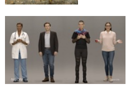
Insignificant:
- Implies a lack of importance or relevance, which is not the intended message in statistics (more data, the potential for new discoveries).
- Even non-significant results can be meaningful, especially in exploratory research.

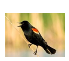
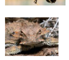
Key Takeaway:
- Use "non-significant" in statistical contexts to convey that while an effect isn't statistically supported, it could still be relevant in practice. Avoid "insignificant", as it implies lack of importance.

2

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]

3

Two-sample comparison of means

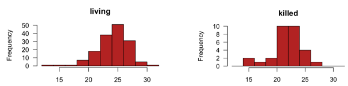
Assumptions:

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

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

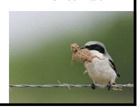
- The standard deviation (and variance) of the variable is the same in both populations.

- The theoretical sampling distribution of the differences between sample means assuming H_0 as true follows a t-distribution only if the samples are drawn from populations with equal variances. While the null hypothesis assumes the populations share the same mean, it does not require the variances to be identical. However, for the t-distribution assumption to hold, equality of variances across populations is necessary.



Horned lizard

Loggerhead shrike



4

Where does the assumption of equal variances for the t-distribution come from? The theoretical population from which the t-distribution was built is the same, i.e., same mean and same variance

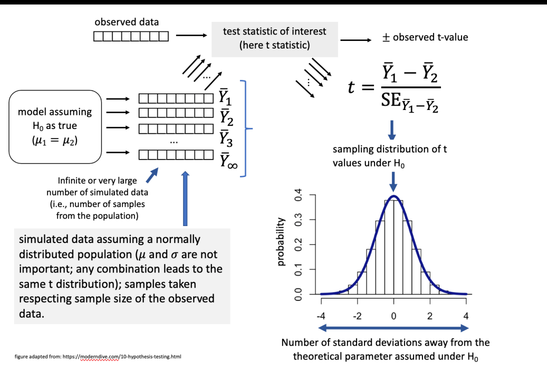


figure adapted from: https://moderndive.com/10-hypothesis-testing.html

5

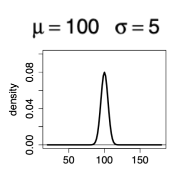
Two-sample t test when sample variances are different

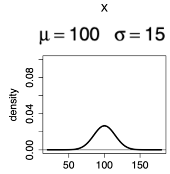
Consider two normally distributed populations with the same mean (μ = 100) but different standard deviations (σ = 5 and σ = 15).

If we set alpha (α) at 0.05 - the probability of rejecting the null hypothesis when it is actually true, representing the rate of false positives—and take 100 sample means from each population to conduct a t-test, we would expect about 5% of the tests to yield significant results purely by chance.

However, when the populations have equal means (μ) but unequal variances (standard deviations), the actual risk of false positives may differ from the pre-set alpha. This means the Type I error rate can either increase or decrease, resulting in inflated or deflated Type I error probabilities (the latter can affect statistical power).

In such cases, the standard t-test for comparing two sample means is not robust against heteroscedasticity (i.e., unequal variances), potentially leading to misleading conclusions.





6

How can we determine if our alpha levels remain valid? We need to assess whether the variances differ: a two-sample variance comparison

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., $\sigma_1^2 \neq \sigma_2^2$).

7

Intuition underlying a two-sample test of variances

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

Now, conduct infinite sampling (or a computationally large number of samples) from populations with equal variances. The population means do not affect variance, so they do not need to be the same.

Each sample should match the appropriate sample sizes (e.g., 154 observations for living lizards and 30 observations for killed lizards).

For each sample pair, calculate the ratio of their variances.

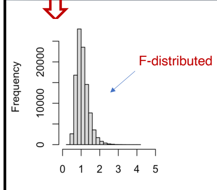
The distribution of all possible sample variance ratios under the null hypothesis will serve as the reference (null) distribution to compare our sample ratio against.

This sampling (null) distribution is called the F-distribution.

8

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)
```



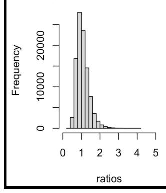
$\mu_1 = 350 \quad \mu_1 = 10$
$\sigma_1 = 100 \quad \sigma_2 = 100$

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

9

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))
```

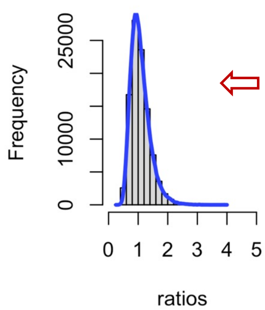


$\mu_1 = 8 \quad \mu_1 = 4$
$\sigma_1 = 7.2 \quad \sigma_2 = 7.2$

Notice that the previous sampling distribution is identical to the one shown here. This consistency holds as long as the null hypothesis (H_0) is true, regardless of the population parameters (such as mean and standard deviation). As a result, this distribution serves as a universal reference for testing the H_0 of homoscedasticity.

10

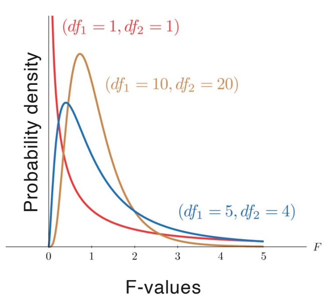
When the null hypothesis (H_0) holds, the sampling distribution of the ratio of two sample variances follows the F-distribution




```
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))
```


11

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



$$F = \frac{s_1^2 \rightarrow df_1}{s_2^2 \rightarrow df_2}$$

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

12

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

13

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

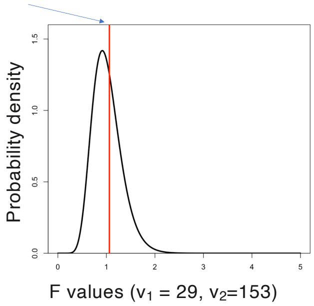
Degrees of freedom (denominator) = 154 - 1= 153

Since the F-distribution is asymmetric, we calculate it by dividing the largest variance by the smallest. This approach yields a slightly different p-value compared to dividing the smallest variance by the largest.

14

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 x Pr[F > 1.06] = **0.7832**

Multiplying the p-value by 2 makes the F-test two-tailed. Due to the asymmetry of the F-distribution, there are other ways to calculate p-values, but for simplicity, we will use this approach and multiply by 2.

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

15

$$F = 1.061762$$

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 x Pr[F > 1.06] = **0.7832**

16

There are other ways to calculate p-values, e.g.:

$$F = 1.061762$$

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

```
p_value_one_tail <- pf(1.061762, 29, 153, lower.tail = FALSE)

# Compute the two-tailed p-value
p_value_two_tail <- 2 * min(p_value_one_tail, 1 - p_value_one_tail)
```

17

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] = **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 differ. Therefore, use the two standard sample t-test for these data as the assumption of equality of variances is met!

18

Two-sample comparison of variances

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

Assumptions:

- Both samples are independently drawn at random from their respective statistical populations (live and dead).
- The variable (e.g., horn length) is normally distributed in each statistical population (live and dead).

19

Let's take a break – 1 minute

20

When the variances of two samples differ, it is necessary to use a different type of t-test for comparing their means—commonly known as Welch's t-test.

Heteroscedasticity (differences in sample variances) is not a concern for the paired t-test, as it analyzes a single sample of differences between paired observations.

21

A study where the two samples are drawn from populations with different variances.

- 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).

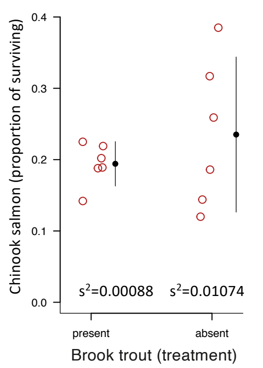
Research question: Does the presence of brook trout affect the survivorship of salmon?

22

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



23

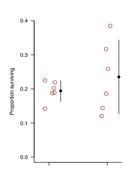
Two-sample comparison of variances

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

We first need to test for differences in variance to determine which type of t-test to use. If the variances differ, the standard t-test is not appropriate, and we should use Welch's t-test instead.

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., $\sigma_1^2 \neq \sigma_2^2$).



24

Two-sample comparison of variances

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

We first need to test for differences in variance to determine which type of t-test to use. If the variances differ, the standard t-test is not appropriate, and we should use Welch's t-test instead.

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

Largest variance

Smallest variance

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

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

Pr[F > 12.17] = 0.007945
2 Pr[F > 12.17] = **0.01589**

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

25

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[F > 12.17] = **0.01589**

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

26

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

Since variances differ, 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

27

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

Welch's t-test uses a different test statistic than the standard t-test for two sample means. Unlike the standard t-test, it does not rely on pooled variances (i.e., variances weighted by sample sizes) to calculate the standard error.

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

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

28

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

29

Welch's t-test is used to compare the means of two independent samples assumed to be drawn from populations with unequal variances

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

30

Differences in degrees of freedom between the standard t-test and the modified Welch's t-test arise when comparing sample means from heteroscedastic populations.

$$\left.\begin{array}{l} df_{Welch} = \dfrac{\frac{1}{6} + \frac{\frac{0.01704}{0.00088}}{6}}{\frac{1}{36(6-1)} + \frac{\left(\frac{0.01704}{0.00088}\right)^2}{36(6-1)}} = 5.8165 \\ \\ df_{standard\ t-test} = (6-1) + (6-1) = 10 \end{array}\right\} t = 0.93148$$

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

31

Non-Whole degrees of freedom in the Welch's Test

Welch's t-Test and degrees of freedom

- In Welch's t-test (and other testss), degrees of freedom can be non-whole numbers.
- This happens because Welch's test uses an *adjusted formula* to better handle differences in group variances, rather than assuming equal variances.

Why non-whole numbers?

- The adjustment in Welch's formula results in a fractional degree of freedom, reflecting the sample sizes and variances of both groups more accurately.
- This fractional degree of freedom improves the precision of the test without requiring complex statistics.

Key takeaway

- Non-whole degrees of freedom in Welch's test help provide a more accurate result by accounting for unequal variances between groups.

32

Remember from an earlier slide in this lecture:

When the null hypothesis is true (equal μ) but the variances (standard deviations) differ, the risk of false positives exceeds the pre-established alpha level (in general).

This is because the standard t-test is not robust against heteroscedasticity (differences in variances between samples).

With smaller degrees of freedom, the p-value for Welch's t-test tends to be larger than that of the standard t-test.

As a result, Welch's t-test adjusts the p-value, making it more difficult to reject the null hypothesis. This adjustment ensures that the risk of committing a Type I error (false positive) aligns with the original significance level (alpha).

33

Why Type I errors are considered worse than Type II errors

Helpful Analogy:

Type I Error: "Crying wolf" when there's no wolf
(rejecting H_0 when is true; claiming that there is an effect when there is none).

Type II Error: Missing the wolf when it's there.
(not rejecting H_0 when is false; not claiming that there is an effect when there is one).

34

Why Type I errors are considered worse than Type II errors

Type I Error (False Positive): Rejecting a true null hypothesis (claiming an effect when there is none).

Potential Impacts:

- **Wastes time and resources:** Pursuing a non-existent effect.
- **Can cause harm:** Approving an ineffective drug or treatment.
- **Loss of credibility:** Damages trust in scientific findings.

Why Type I errors are often considered worse:

False Hope or Danger: Imagine a new drug is approved but it doesn't work—this could lead to serious consequences.

More Difficult to Detect: Once published, Type I errors may persist longer in the scientific record.

Damage to Reputation: Especially in fields where public safety or health is involved.

35

Welch's t-test is used to compare the means of two independent samples assumed to be drawn from populations with unequal variances

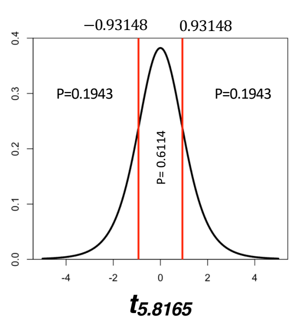
$t = 0.93148$

$df = 5.8165$

two tailed t-test
Pr[t < -0.931] + Pr[t > 0.931] =
2 x Pr[t > abs(0.931)] = **0.3886**

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

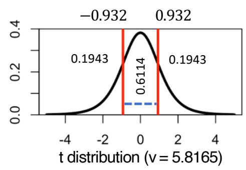
Conclusion: There is insufficient evidence to conclude that the mean proportion of Chinook survival differs between streams with and without brook trout (i.e., $\mu_1 \neq \mu_2$).



36

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

Because Welch's t-test has smaller degrees of freedom, it becomes more difficult to reject the null hypothesis, ensuring that the risk of false positives remains equal to the desired significance level (alpha).

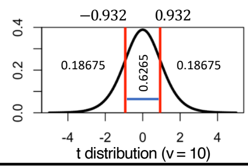


df = 5.8165 (**Welch's t test**)

$t = 0.93148$

2 Pr[t > abs(0.931)] = **0.3886**

2 x 0.1943



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

$t = 0.93148$

2 Pr[t > abs(0.931)] = **0.3735**

$n_1 = 6$ $n_2 = 6$

2 x 0.18675

37

Welch's t-test is used to compare the means of two independent samples assumed to be drawn from populations with unequal variances

Assumptions:

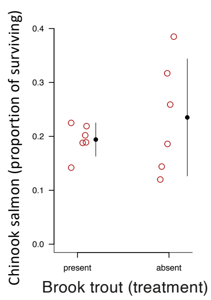
- Each sample is independently and randomly drawn from its respective statistical population.

The variable of interest (e.g., horn length, survival proportion) follows a normal distribution within each population.

38

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

Conclusion: There is insufficient evidence to conclude that the mean proportion of Chinook survival differs between streams with and without brook trout (i.e., $\mu_1 \neq \mu_2$).

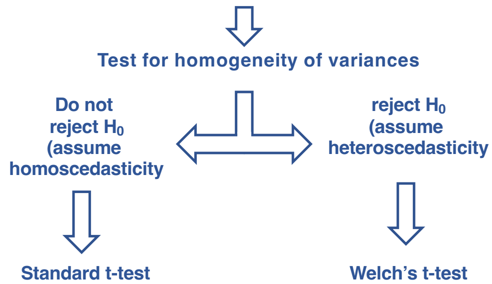


Relevant issues:

- The sample size may be too small to detect meaningful differences.
- Differences in variances reduce the degrees of freedom, which significantly lowers the statistical power.
- Even if the means are not truly different (i.e., H_0 is true; something we cannot confirm), it is crucial to recognize that the variances are statistically different. This difference may have important implications, particularly in areas such as conservation.

39

Two-sample t test for comparing means



Note that I used the terms 'assume' homoscedasticity and 'assumed' heteroscedasticity because we cannot know for certain whether the variances of the two samples are truly different. All we have is the outcome of the F-test, which either rejects or fails to reject the null hypothesis.

40