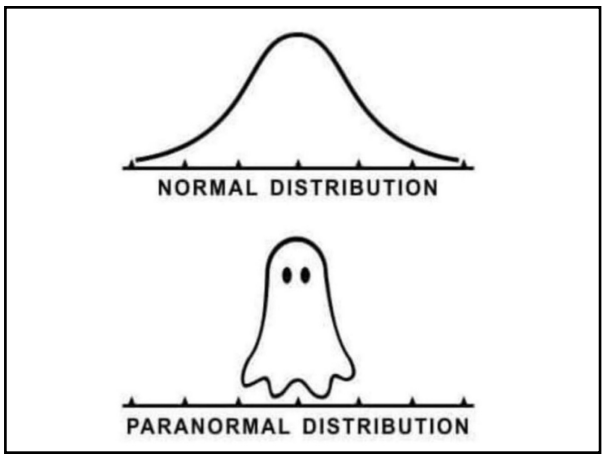



1

**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; rather, it may reflect that the effect is not statistically detectable given the sample (i.e., a potential Type II error).

**Insignificant:**

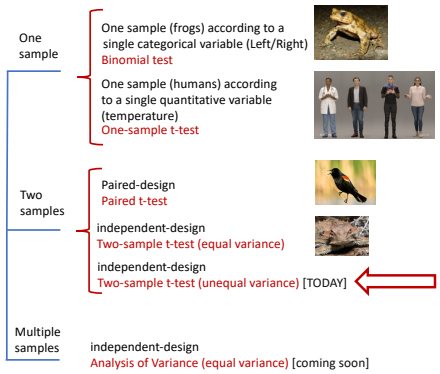
- Implies a lack of importance or relevance, which is not the intended message in statistics (more data, the potential for new discoveries).
- Even non-significant results can be meaningful, especially in exploratory research.

**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




3

## Two-sample comparison of means

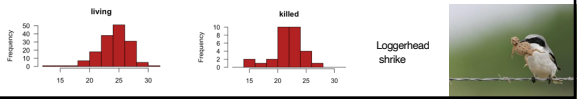
**Assumptions:**

*- 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.




Loggerhead shrike

Horned lizard



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

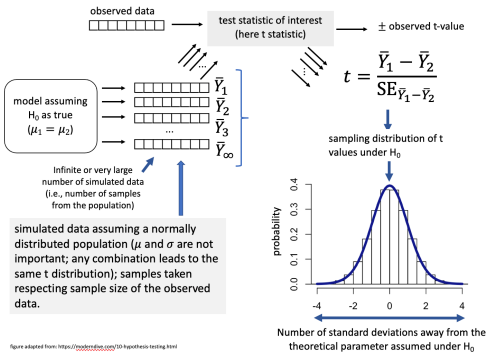



figure adapted from: https://moderndive.com/10-hypothesis-testing.html

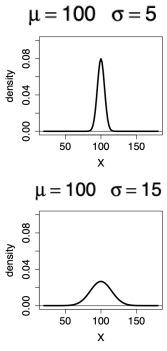
5

## Two-sample t test when sample variances are different

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

When the assumption of equal variances is violated (heteroscedasticity), the actual Type I error rate of tests that assume equal variances (e.g., the classical Student's two-sample t-test) can deviate from the nominal α. This deviation can go in either direction: the test can become liberal (inflated Type I error) or conservative (deflated Type I error).

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




6

## How can we determine whether the nominal α level remains valid? This requires evaluating whether the assumption of equal variances holds

| 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, perform repeated sampling (i.e., a very large number of samples) from populations with equal variances. Because the population mean does not influence the variance, the means do not need to be equal.

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

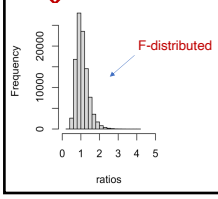
## Intuition underlying a two-sample test of variances – their ratios are F-distributed

```
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)
```




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

$\boldsymbol{\sigma_1 = 100}$ $\boldsymbol{\sigma_2 = 100}$

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

9

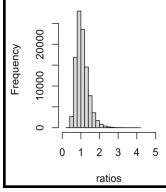
## Let's change the population parameters

```
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))
```



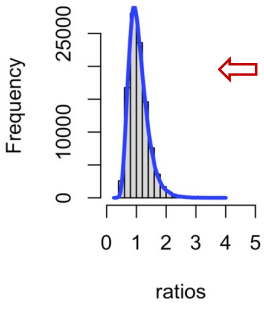

$\mu_1 = 8 \quad \mu_1 = 4$

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

Notice that the previous sampling distribution is identical to the one shown here. Therefore, this distribution provides a universal reference for testing the null hypothesis ($H_0$) of homoscedasticity.

10

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

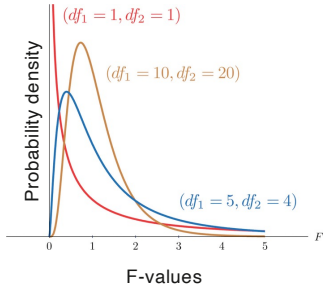



Note (not critical for BIOL-322): If the null hypothesis is **not true** (i.e., the population variances are unequal), then the ratio of sample variances **no longer follows the standard F-distribution**. Instead, it follows a *scaled* F-type distribution:

$$\frac{\sigma_1^2}{\sigma_2^2} \times \mathrm{F}$$

11

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




$$\mathrm{F} = \frac{s_1^2}{s_2^2} \quad \begin{matrix} \longrightarrow \mathrm{df}_1 \\ \longrightarrow \mathrm{df}_2 \end{matrix}$$

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

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

Because the F-distribution is asymmetric, we typically place the larger variance in the numerator so that the F-statistic is ≥ 1 and easier to interpret.

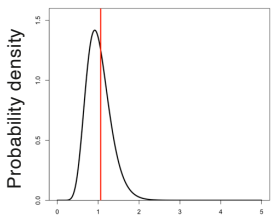
Using the reverse ratio leads to a different numerical value, but it corresponds to the opposite tail of the same distribution, resulting in an equivalent test and conclusion.

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 ($v_2$)




F values ($v_1$ = 29, $v_2$=153)

Pr[F > 1.06] = 0.3916
2 x Pr[F > 1.06] = **0.7832**

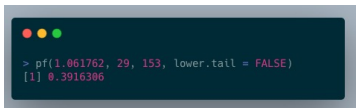
Multiplying the p-value by 2 gives the exact p-value provided that the largest variance is in the numerator.

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 ($v_2$)




Pr[F > 1.06] = 0.3916
2 x Pr[F > 1.06] = **0.7832**

Multiplying the p-value by 2 gives the exact p-value provided that the largest variance is in the numerator.

16

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!

17

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).

18

19

When the variances of two samples are unequal, a different version of the t-test should be used to compare their means, namely, the Welch's t-test.

In contrast, heteroscedasticity is not an issue for the paired t-test, because it operates on a single sample of differences between paired observations rather than on two separate samples.

20

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).

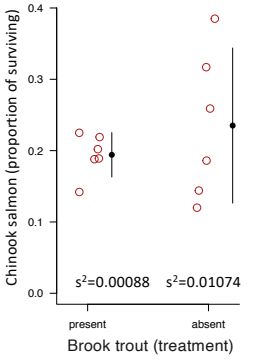
Research question: Does the presence of brook trout affect the survivorship of salmon?

21

## A study in which the variance of the two samples differ

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

Note the differences in variances




22

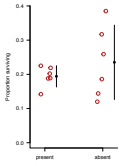
## Two-sample comparison of variances

***Research question:*** Does the presence of brook trout affect the survivorship of chinook 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 salmon 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 salmon surviving differs in streams with and without brook trout(i.e., $\sigma_1^2 \neq \sigma_2^2$).



23

## Two-sample comparison of variances

Research question: Does the presence of brook trout affect the survivorship of chinook 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 $H_A$.***

24

## 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 $H_A$***

25

## 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 |

26

## 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.

$$t = \frac{(Y_1 - Y_2)}{\mathrm{SE}_{Y_1 - Y_2}}$$

$$\mathrm{SE}_{Y_1 - Y_2} = \sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

Standard t-test for comparing two-sample means use a common variance estimator (i.e., pooled variance).

$$t = \frac{(Y_1 - Y_2)}{\mathrm{SE}_{Y_1 - Y_2}}$$

$$\mathrm{SE}_{Y_1 - Y_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Welch's test does not assume equal population variances, so each sample provides its own estimate of variability.

27

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.

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

Welch's test estimates variability separately for each sample instead of pooling them, so the overall uncertainty depends on how reliable each variance estimate is.

When sample sizes are small or variances are very different, this uncertainty increases, effectively reducing the degrees of freedom in comparison to the standard two-sample t-test (*df*=n1 + n2 – 2).

| 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

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 |

29

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.

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

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

$$t = 0.93148$$

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

30

## 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 tests), 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.

31

Remember from an earlier slide in this lecture:

When the null hypothesis for means is true (equal $\mu$) but the variances 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 adjusted p-value ensures that the risk of committing a Type I error (false positive) aligns with the original significance level (alpha).

32

Type I error rates: regular *versus* Welsh t-test for two samples under homoscedasticity

```
n.sim  <- 100000
alpha  <- 0.05

# Same means -> H0 is true
mu1 <- 10
mu2 <- 10

# Unequal sample sizes
n1 <- 30
n2 <- 10

sd1 <- 1
sd2 <- 1

res <- replicate(n.sim, {
  x <- rnorm(n1, mean = mu1, sd = sd1)
  y <- rnorm(n2, mean = mu2, sd = sd2)
  c(
    regular = t.test(x, y, var.equal = TRUE)$p.value,
    welch   = t.test(x, y, var.equal = FALSE)$p.value
  )
})

p_regular <- res["regular", ]
p_welch   <- res["welch", ]

type1_regular <- mean(p_regular < alpha)
type1_welch   <- mean(p_welch < alpha)

> type1_regular
[1] 0.04995
> type1_welch
[1] 0.05212
```

33

Type I error rates: regular *versus* Welsh t-test for two samples under heteroscedasticity

```
n.sim  <- 100000
alpha  <- 0.05

# Same means -> H0 is true
mu1 <- 10
mu2 <- 10

# Unequal sample sizes
n1 <- 30
n2 <- 10

# Variances differ by a factor of 3
sd1 <- 1
sd2 <- 3

res <- replicate(n.sim, {
  x <- rnorm(n1, mean = mu1, sd = sd1)
  y <- rnorm(n2, mean = mu2, sd = sd2)
  c(
    regular = t.test(x, y, var.equal = TRUE)$p.value,
    welch   = t.test(x, y, var.equal = FALSE)$p.value
  )
})

p_regular <- res["regular", ]
p_welch   <- res["welch", ]

type1_regular <- mean(p_regular < alpha)
type1_welch   <- mean(p_welch < alpha)

> type1_regular
[1] 0.21277
> type1_welch
[1] 0.0523
```

34

## Type I *versus* Type II errors

A Type I error is like raising a false alarm ("there is a wolf when there isn't")… while a Type II error is missing a real danger ("there is a wolf, but we don't act"). Which error is worse depends on the situation and its consequences.

Whether one is worse than the other depends on the consequences: in some situations (e.g., approving an ineffective drug), false positives are more serious, whereas in others (e.g., missing a real disease), false negatives may be more harmful.

A Type I error is like concluding that a species is present in an area when it is actually absent—perhaps leading to unnecessary protection measures or misallocation of conservation resources.

A Type II error is like failing to detect a species that is truly present—potentially resulting in a lack of protection and increasing the risk of its decline or local extinction.

35

Why is there such emphasis on ensuring that the *Type I error* rate (α) is properly controlled? For example, Welch's t-test maintains the correct *Type I error* rate when variances differ, and procedures addressing multiple testing or p-hacking aim to prevent an inflated risk of false discoveries.

Because in most statistical frameworks, we want to control the rate of false discoveries when claiming effects.

When you reject $H_0$, you are making a positive claim ("there is an effect," "the species is present," "this treatment works"). If that claim is wrong (*Type I error*), it can: create false knowledge, propagate through future studies, and lead to poor decisions or wasted resources.

In contrast, a *Type II error* is more conservative: you simply fail to make a claim. While this can also be costly (e.g., missing a real species), it does not introduce false conclusions into the scientific record.

36

## 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.

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:

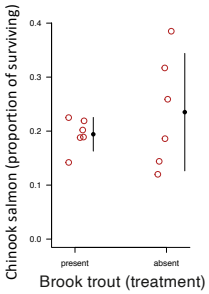
*- Each sample is independently and randomly drawn from its respective statistical population.*

*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

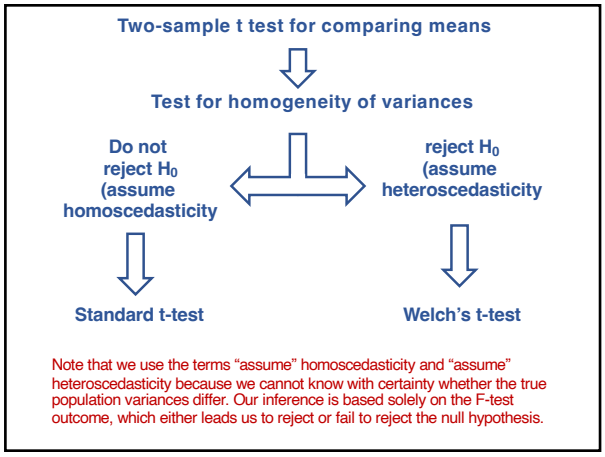
Conclusion: There is insufficient evidence to conclude that the mean proportion of Chinook survival differs between streams with and without brook trout (i.e., $\mu_1 \neq \mu_2$).




***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 verify directly), it is important to recognize when the variances differ statistically. Such differences can be meaningful in their own right and may have important implications, particularly in contexts such as conservation.

39


Two-sample t test for comparing means
Test for homogeneity of variances
Do not reject $H_0$ (assume homoscedasticity
reject $H_0$ (assume heteroscedasticity
Standard t-test
Welch's t-test
Note that we use the terms "assume" homoscedasticity and "assume" heteroscedasticity because we cannot know with certainty whether the true population variances differ. Our inference is based solely on the F-test outcome, which either leads us to reject or fail to reject the null hypothesis.

40