

Module 2AB:

Inference for a Normal Population

Different flavours of t tests

Degrees of freedom can be a tricky subject

- The amount of independent data point or information that you have
- You ‘use up’ independent information points when you estimate parameters from your actual data --- because now your estimates are forced to be more like your data points.
- Each test calculates different numbers of estimates, so they account for how they use up independent information a bit differently.
- χ^2 Contingency tests:

			Observed totals
A	B	row total-(A+B)	A+B+C
D	E	row total-(D+E)	D+E+F
G	H	row total-(G+H)	G+H+I
Col Total-(A+D+G)	Col Total-(B+E+H)	Row total – (J+K)	J+K+L
A+D+G+J	B+E+H+K	C+F+I+L	A+B+C+D+E+F+G+H+I+J+K+L

Degrees of freedom can be a tricky subject

- The amount of independent data point or information that you have
- χ^2 Contingency tests:

			Observed totals
A	B	row total-(A+B)	A+B+C
D	E	row total-(D+E)	D+E+F
G	H	row total-(G+H)	G+H+I
Col Total-(A+D+G)	Col Total-(B+E+H)	Row total – (J+K)	J+K+L
A+D+G+J	B+E+H+K	C+F+I+L	A+B+C+D+E+F+G+H+I+J+K+L

- χ^2 Goodness of fit tests:

Number of categories – 1 – number of estimated parameters

There is a total count that constrains your estimates

Usually estimating mean from observed data
So, the mean will be constrained by the observed data

Agenda: Hypothesis testing for means using t tests

1. Why do we use Student t-tests instead of Z scores?

2. One sample t tests

- Assumptions
- Assumptions not met; median and rank → Signed test

3. Paired t test

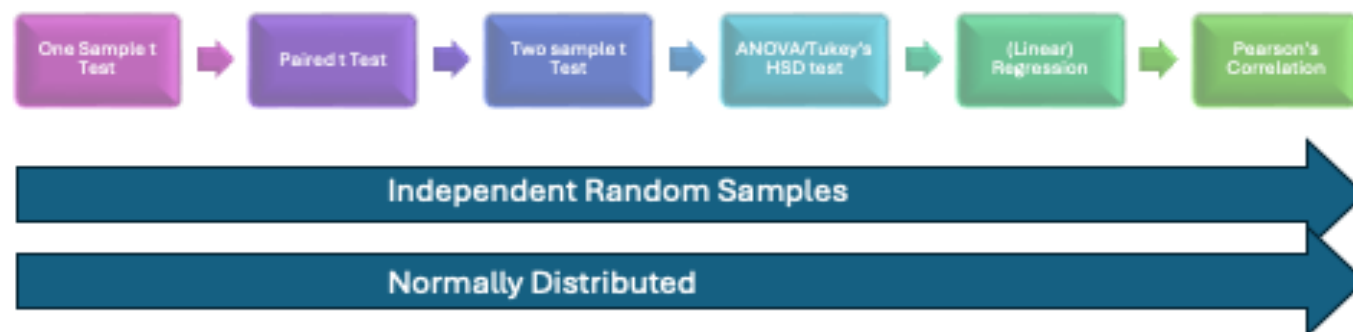
- Assumptions

4. Two sample t test

- Assumptions
- When variances aren't equal → **Welch's approximate t test**
- Assumptions not met; median and rank → Mann Whitney U test

We won't have time to cover everything so here is another reference that outlines the different t tests:

<https://statisticsbyjim.com/hypothesis-testing/t-tests-1-sample-2-sample-paired-t-tests/>



How to test:

- Visual Assessment (Boxplot, histograms, qqplot, qqnorm)
- Shapiro Wilk test

Homoscedasticity (variances are equal)

How to test

- Visual Assessment
- If normal; $n=2 \rightarrow \text{var.test}$
- If normal; $n>2 \rightarrow \text{Bartlett test}$
- If not norm $\rightarrow \text{Levene's test}$
- If $n=2$ and variances are significantly different:

Welch's
Approximate t test

Non-parametric analogs of the parametric tests above. Instead of testing population **means**, they test population **medians**, and they use **ranks**.



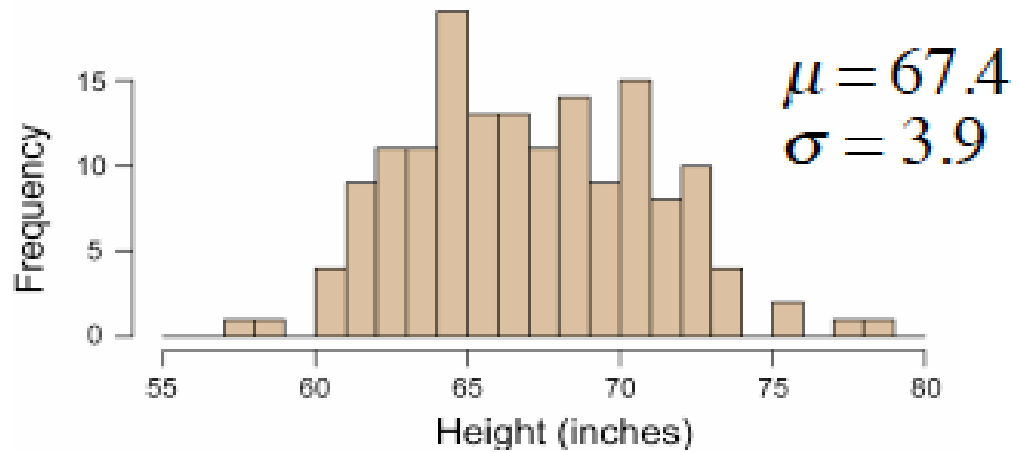
Inference about means:

As a reminder from Biostatistics I....

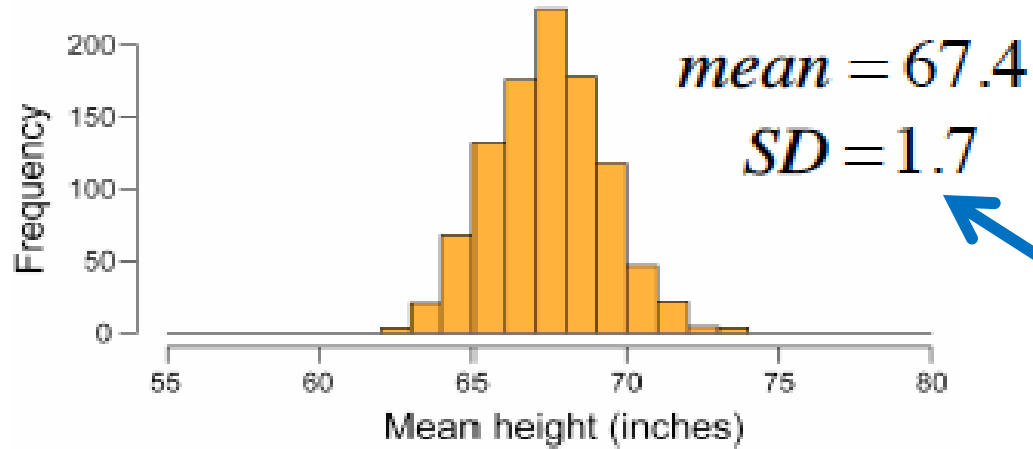
- To make statistical statements, we need to describe the sampling distribution of an estimator.
 - The sampling distribution is the probability distribution of all values of an estimate that we might obtain when sampling a population
 - *When the variable, Y , is normally distributed or n is large, the sampling distribution for $E(Y)$ is normal**

* thank-you Central Limit Theorem

Inference for a Normal Population



Mean heights of samples of size 5
(1000 samples)



The central limit theorem has two constraints:

1. It depends on a large sample size ($n > 30$ -ish)
2. To use it, we need to know σ^2 (i.i.d.), but we seldom do.

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$

Inference about Means

- Because \bar{Y} is normally distributed, we can convert the distribution to the **standard normal distribution**:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

- This gives a probability distribution of the difference between a **sample** mean and **the population** mean (scaled)

Inference about Means

But we don't know σ !

Now what?

- we *do* know **s** , the standard deviation of our sample, which estimates **σ** .

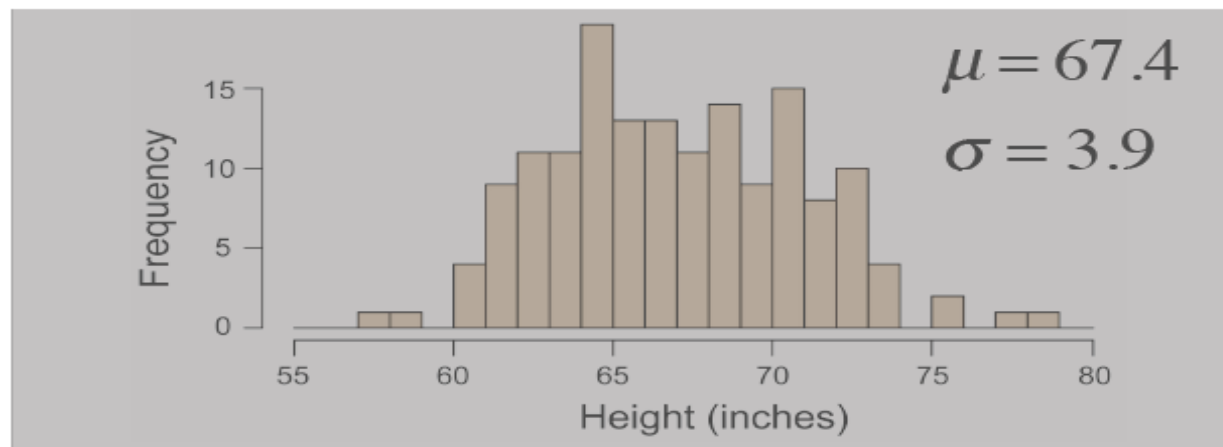
- We can use **s** to get: $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$

- This is used as an estimate of $\sigma_{\bar{Y}}$

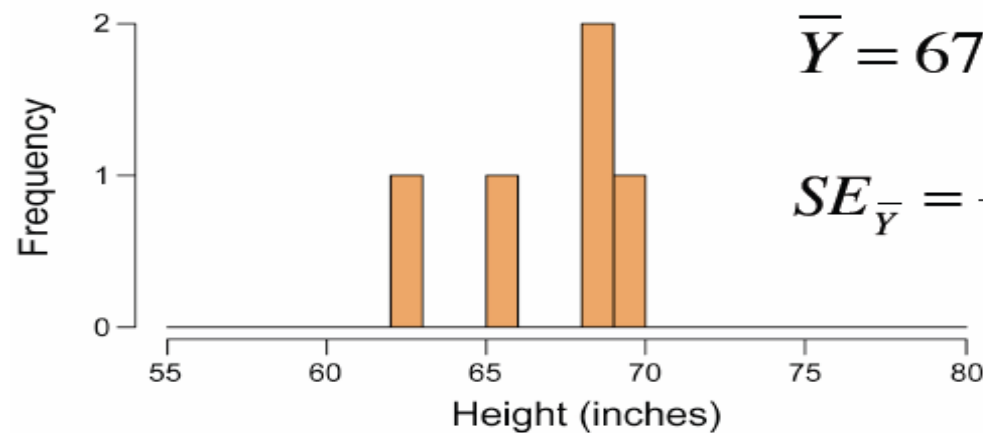
Inference for a Normal Population

In most cases, we don't know the real population distribution.

We only have a sample.



Heights of a sample of students ($n = 5$)



$$\bar{Y} = 67.1 \quad s = 3.1$$

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\bar{Y}}$

A one sample t -test is usually used instead of a one sample z -test to correct for:

A) \bar{X} used to estimate μ

B) s used to estimate σ

C) n used to estimate N (population size)

D) none of the above

The Z score:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

The Student's t Distribution:



$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$

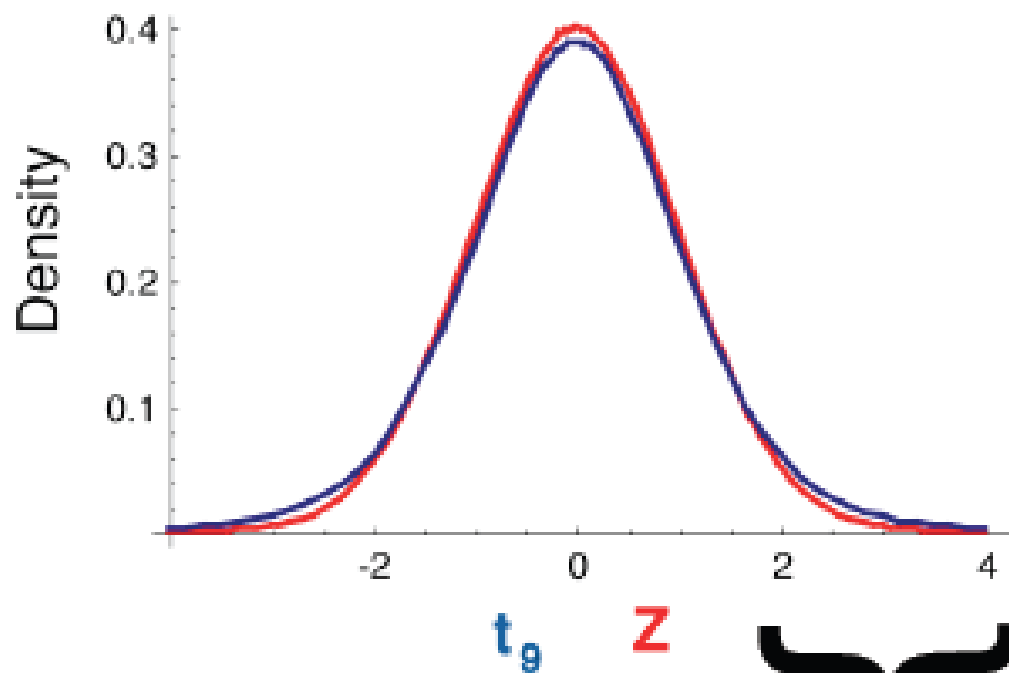
* Alias of William Gosset of the Guinness Brewing Company

	<u>t</u>	<u>z</u>
<u>Stand. Error</u>	$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$	$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
<u>dof</u>	n - 1	n
<u>Sampling Distribution</u>	t-distribution	Normal Distribution

Why is degrees of freedom n-1 instead of n?

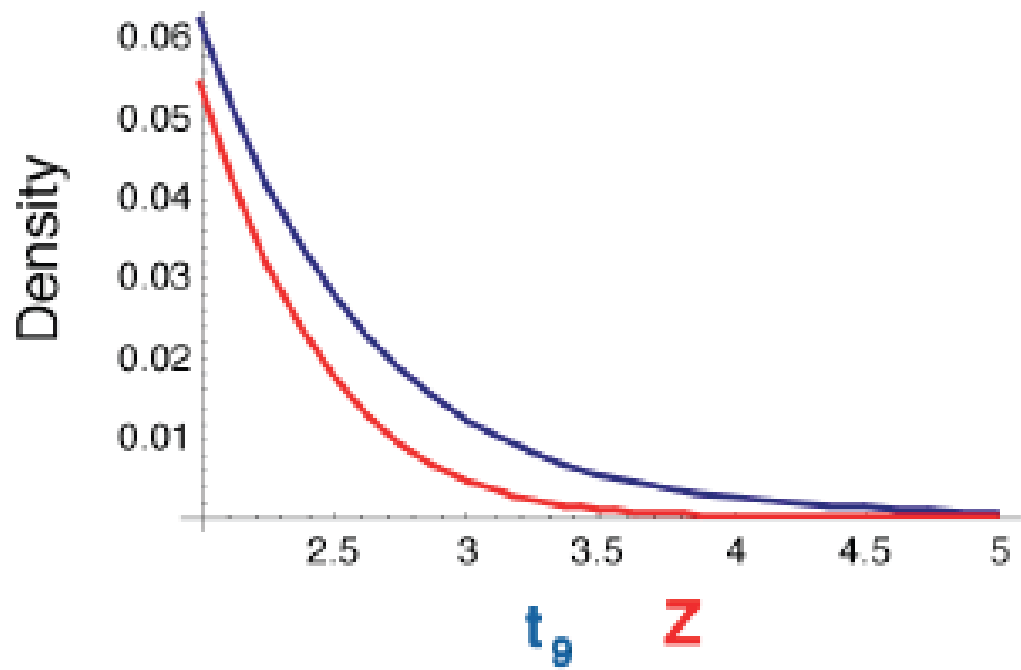
The consequences of using $SE_{\bar{Y}}$ instead of $\sigma_{\bar{Y}}$

- The value of $SE_{\bar{Y}}$ is different for each sample; it doesn't have a constant value like $\sigma_{\bar{Y}}$
 - *results in the introduction of some error*
 - t-distribution is wider than the equivalent Normal distribution
 - not as precise
 - As sample size, n , increases the t-distribution narrows and approaches the Normal Distribution
- *dof = $n - 1$ because we have 'used up' one piece of information when we estimate $\sigma_{\bar{Y}}$*



$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$



A one sample t -test is conducted on $H_0: \mu = 81.60$.

The sample has $\bar{X}=84.10$, $s = 3.10$, and $n = 25$. The t -test statistic is:

A) 0.81. $t = \frac{84.10-81.60}{3.10/\text{Sqrt}(25)}$

B) 1.80

C) 4.03

D) 20.16

https://PollEv.com/multiple_choice_polls/Mwcst98NN6ayUJscIKwQq/respond

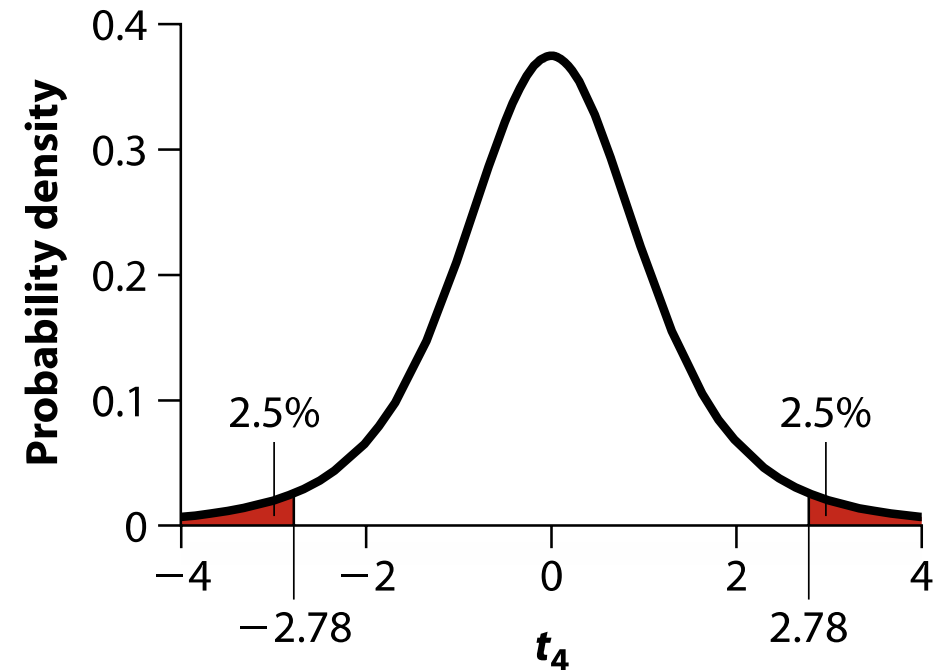
Critical values for the student's t distribution:

<https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf>

- $t_{0.05(2),df}$

– $\alpha = 0.05$ is the fraction of the area under the curve shared between the two tails of the distribution:

$2.5\% > t_{0.05(2),df}$ and $-2.5\% < -t_{0.05(2),df}$



Use the t-distribution to calculate the **confidence interval** for the mean of a normal distribution

$$-t_{\alpha(2),df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < t_{\alpha(2),df}$$

- This can be re-arranged:

$$\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}$$

Never this (why not?): ~~$\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \bar{Y} < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}$~~

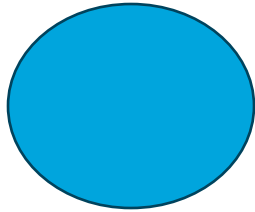
Is a 95% confidence interval always sufficient in a t-test as compared to a z-test?

- A) Yes, the area under the tails are equal and therefore have the same likelihood of encompassing the true parameter
- B) No, the Z-test has fatter tails and therefore the 99% confidence interval is often more appropriate
- C) No, the t-test has fatter tails and therefore the 99% confidence interval is often more appropriate
- D) Yes, the size of the tails has no effect on finding the true parameter of the mean

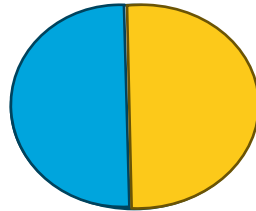
https://PollEv.com/multiple_choice_polls/QcJCy2IxSpioFpD3F2iVm/respond

We will look at the following t-tests:

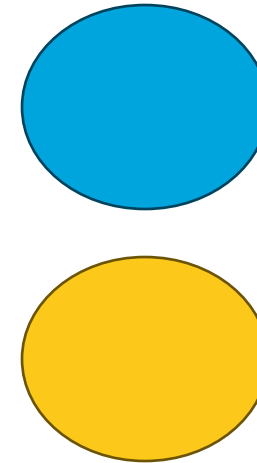
1. Comparing one mean:
 - a. One-sample t-test
2. Comparing two means:
 - a. Paired t-test
 - b. Two-sample t-test



one sample



paired



two sample

*Each of the above tests have **slightly different assumptions** which allow our conclusions to be supported. We will investigate what happens when these assumptions are violated and how robust our various t-tests are to violations.*

Researchers are studying the body weight of mice to understand the impact of a high-fat diet on genetically modified (GM) mice. They can collect the following data: Body Weight.

One-Sample t-test:

Two-Sample t-test:

Paired t-test:

Applications of one sample t-test

Researchers are studying the body weight of mice to understand the impact of a high-fat diet on genetically modified (GM) mice. They can collect the following data: body weight.

One-Sample t-test: *Does the body weight of the GM mice differ significantly from a known population mean weight of non-GM mice?*

Two-Sample t-test: *Does the body weight of GM mice on a high-fat diet differ from the body weight of a GM mice on a standard diet?*

Paired t-test: *They measure the body weights of a group of GM mice **before** and **after** they are switched from a normal diet to a high fat diet to see if there's a significant change in weight within the same group.*

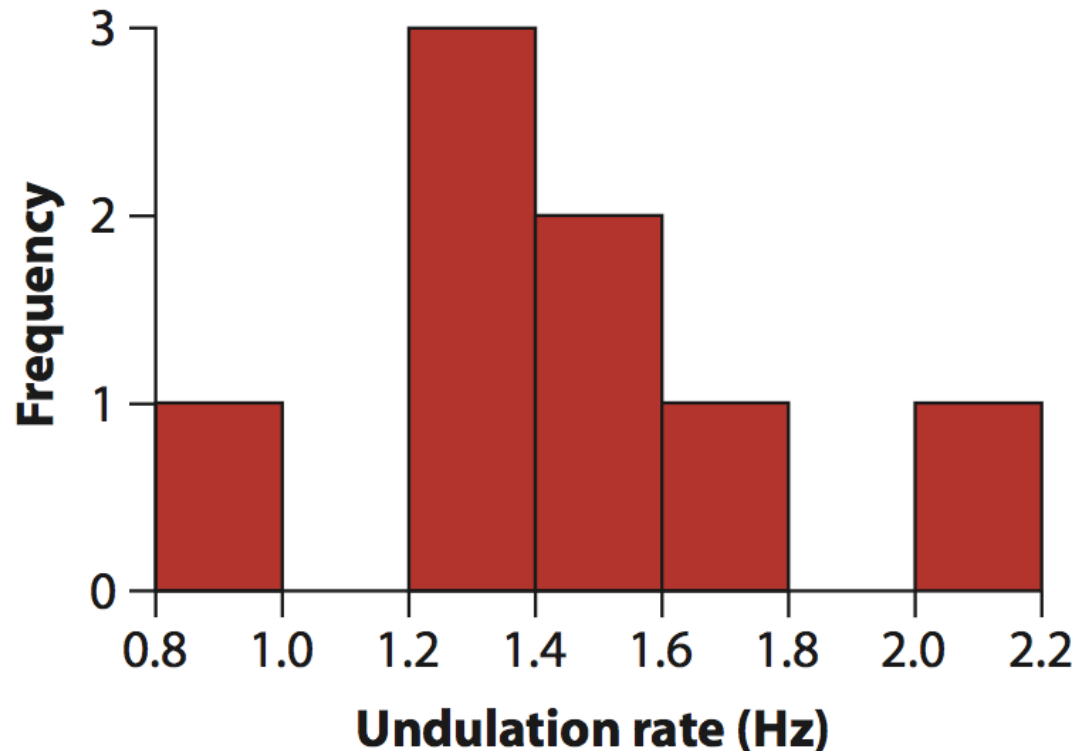
Example: What is the **95% confidence interval** for the mean undulation rate of the paradise flying snakes. The rate of undulation (hz) from 8 snakes:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

Applications of one sample t-test

Example: What is the 95% confidence interval for the mean undulation rate of the paradise flying snakes. The rate of undulation (hz) from 8 snakes:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6



$$\bar{Y} = 1.375 \quad dof = n - 1 = 7$$

$$s = 0.324$$

$$n = 8$$

$$t_{\alpha(2),df} = t_{0.05(2),7} = 2.36$$

Inference for a Normal Population

Example: What is the 95% confidence interval for the mean undulation rate of the paradise flying snakes. The rate of undulation (hz) from 8 snakes:

0.9,1.4,1.2,1.2,1.3,2.0,1.4,1.6

Answer:

$$\bar{Y} \pm t_{0.05(2),7} SE_{\bar{Y}} = 1.375 \pm 0.115(2.36)$$

1.10 < μ < 1.65 (95% Confidence Interval)

$$\bar{Y} \pm t_{0.01(2),7} SE_{\bar{Y}} = 1.375 \pm 0.115(3.50)$$

0.97 < μ < 1.78 (99% Confidence Interval)

The one-sample t test:

Compares the mean of a random sample from a normal population with the population mean proposed in a null hypothesis

H_0 : True mean equals μ_0

H_A : True mean *does not* equal μ_0

Assumptions:

- The variable is normally distributed
- The sample is a random sample

HOW WOULD YOU TEST THE FOLLOWING:

Do high school seniors from a particular school who attend a summer math camp score above the state mean on the math subtest of the state's standardized achievement test ? ($n=15$)

- a. Test one mean against a hypothesized constant.
- b. Test the difference between two means (independent samples).
- c. Test the difference in means between two paired or dependent samples.
- d. Use a chi-squared test of association.

https://PollEv.com/multiple_choice_polls/wSrD7fROl6rmRpSumY2d8/respond

One Sample t-Test

H_0 : True mean equals μ_0

H_A : True mean *does not* equal μ_0

Test Statistic:

$$t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

One Sample t Test

Example: The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

One Sample t Test

Example: The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

Step 1:

H_0 : No change in mean fitness, $\mu = 0$

H_A : Change in mean fitness after 100 gens, $\mu \neq 0$

One Sample t Test

Example: The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

H_0 : No change in mean fitness, $\mu = 0$

H_A : Change in mean fitness after 100 gen, $\mu \neq 0$

Step 2: conduct the test (and investigate assumptions)

$$\begin{aligned}\bar{Y} &= 0.2724 \\ s &= 0.600 \\ n &= 5 \\ dof &= 4\end{aligned}$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{0.2724 - 0}{0.600/\sqrt{5}} = 1.02$$

One Sample t Test

Example: The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

H_0 : No change in mean fitness, $\mu = 0$

H_A : Change in mean fitness after 100 gen, $\mu \neq 0$

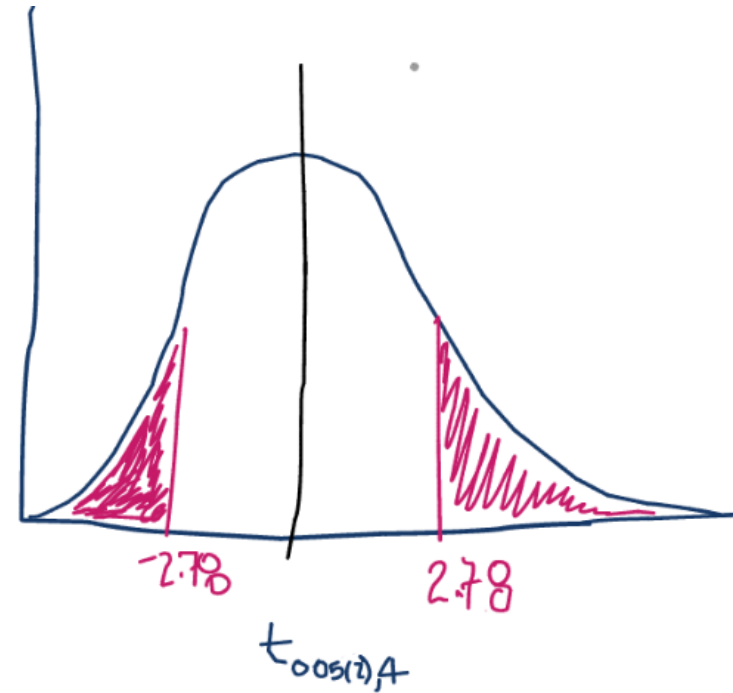
$$\bar{Y} = 0.2724$$

$$s = 0.600$$

$$n = 5$$

$$df = 4$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{0.2724 - 0}{0.600/\sqrt{5}} = 1.02$$



One Sample t Test

Example The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

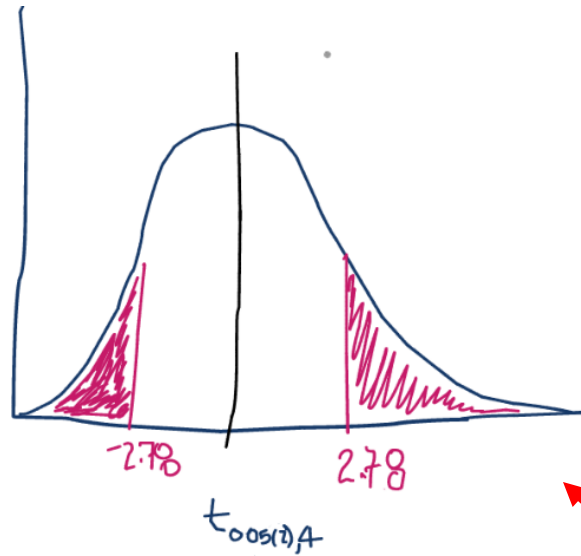
0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

H_0 : No change in mean fitness, $\mu = 0$

H_A : Change in mean fitness after 100 gen, $\mu \neq 0$

$$\begin{aligned}\bar{Y} &= 0.2724 \\ s &= 0.600 \\ n &= 5 \\ df &= 4\end{aligned}$$



$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{0.2724 - 0}{0.600/\sqrt{5}} = 1.02$$

How unusual is this data?

One Sample t Test

Example The consequences of mutation on fitness after 100 generations in $\phi 6$ using 5 lines:

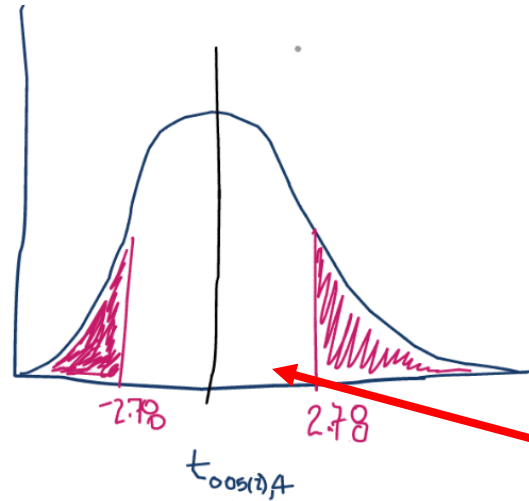
0.063, -0.062, 0.064, -0.043, 1.34

Has the mean fitness changed?

H_0 : No change in mean fitness, $\mu = 0$

H_A : Change in mean fitness after 100 gen, $\mu \neq 0$

$$\begin{aligned}\bar{Y} &= 0.2724 \\ s &= 0.600 \\ n &= 5 \\ df &= 4\end{aligned}$$



$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{0.2724 - 0}{0.600/\sqrt{5}} = 1.02$$

How unusual is this data?

Answer: Not very. Based on this sample of mean fitness's, we fail to reject H_0 .

We could also calculate the confidence interval for mean fitness after 100 generations...

$$\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}$$

For 95%:

$$-0.47 < \mu < 1.02$$

Assumptions:

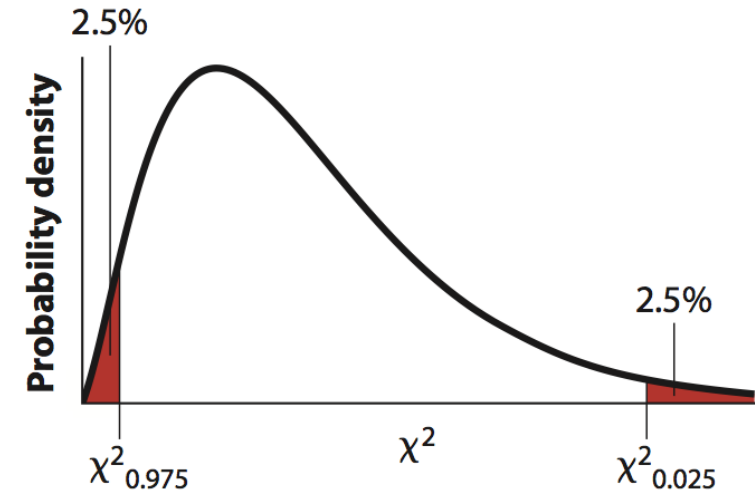
- Random sample
- Normally (ish) distributed variable

Estimating the Standard Deviation and Variance of a Normal Population:

- Y is a normally distributed variable then the sampling distribution of:

$$(n-1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$$

is the χ^2 distribution with dof = n-1



- This can be rearranged to give CI

$$\frac{df * s^2}{\chi_{\frac{\alpha}{2}, df}^2} < \sigma^2 < \frac{df * s^2}{\chi_{1-\frac{\alpha}{2}, df}^2}$$

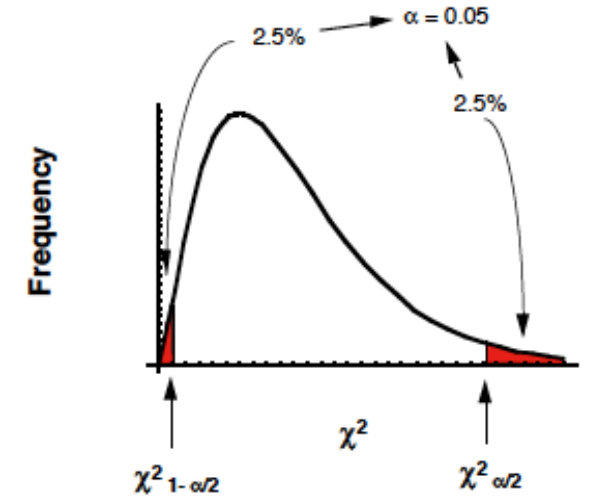
$$\sqrt{\frac{df * s^2}{\chi_{\frac{\alpha}{2}, df}^2}} < \sigma < \sqrt{\frac{df * s^2}{\chi_{1-\frac{\alpha}{2}, df}^2}}$$

Inference for a Normal Population

Example: What is the 95% confidence interval for the **mean** undulation rate of the paradise flying snakes. The rate of undulation (hz) from 8 snakes:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

Answer: $\bar{Y} \pm t_{0.05(2),7} SE_{\bar{Y}} = 1.375 \pm 0.115(2.36)$
 $1.10 < \mu < 1.65$ (95% CI)



We can also add CI for variance!

$$\frac{df * s^2}{\chi^2_{\frac{\alpha}{2}, df}} < \sigma^2 < \frac{df * s^2}{\chi^2_{1-\frac{\alpha}{2}, df}}$$

$$df = n - 1 = 7$$

$$s^2 = (0.324)^2 = 0.105$$

$$\frac{df * s^2}{\chi^2_{\frac{\alpha}{2}, df}} < \sigma^2 < \frac{df * s^2}{\chi^2_{1-\frac{\alpha}{2}, df}}$$

$$\frac{7(0.324)^2}{16.01} < \sigma^2 < \frac{7(0.324)^2}{1.69}$$

$$0.0459 < \sigma^2 < 0.435$$

df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
X	1.6E-6	3.9E-5	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0	0.01	0.02	0.05	0.1	5.99	7.38	9.21	10.6	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.3	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.6	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

$$\chi^2_{\frac{\alpha}{2}, df} = \chi^2_{0.025, 7} = 16.01$$

$$\chi^2_{1-\frac{\alpha}{2}, df} = \chi^2_{0.975, 7} = 1.69$$

One Sample t Test

Example A professor wants to test if her introductory class has a good grasp of basic concepts. 6 students are randomly chosen and given a proficiency test. The professor wants the class to be able to score above 70 on the test.

62, 92, 75, 68, 83, 95

Can the professor be at least 90% certain that the mean score for the class on the test would be 70%?

- A. Yes
- B. No

Example A professor wants to test if her introductory class has a good

grasp of basic concepts. 6 students are randomly chosen and given proficiency test. The professor wants the class to be able to score above 70 on the test.

62, 92, 75, 68, 83, 95

Can the professor be at least 90% certain that the mean score for the class on the test would be 70%?

Step 1: $H_0: \mu \leq 70$; $H_A: \mu > 70$

Step 2: Use a one sample t test. Calculate the relevant values from our data: sample mean = 79.17

sample stand. dev. = 13.17

$$\text{t-value: } t = \frac{79.17 - 70}{13.17 / \sqrt{6}} = \frac{9.17}{5.38} = 1.71$$

Step 3: compare to the critical value of **t** that has **5 df** and alpha = 0.10 (one-tailed so you don't divide alpha by 2); $t_{0.10,5} = 1.476$

Step 4: Since $1.71 > 1.476$, we reject H_0 and with 90% confidence state that the true class mean on the math test would be at least 70%. The 90% Confidence Interval:

$$79.17 - 1.476 * 5.38 < \mu < 79.17 + 1.476 * 5.38$$

$$71.23 < \mu < 87.11 \text{ (note that 70 is not included)}$$

Comparing Two Means

A tiny bit of review:

Which statement is NOT true about one-sample t-test?

- A. It is used to compare the sample mean of a variable to hypothesized value.
- B. The test only assumes that data are random sample from the population.
- C. The test statistic is t.
- D. If the null hypothesis is true, then t should have a t-distribution with $n - 1$ df

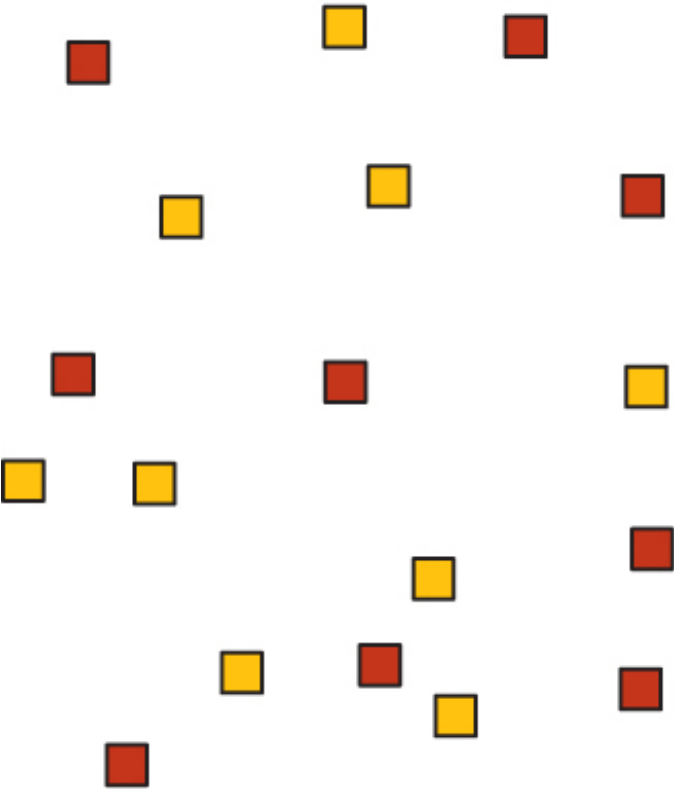
What does the two-sample t-test method do?

- a) Compares the means of a numerical variable between two dependent groups
- b) Compares the standard error of a numerical variable between two independent groups
- c) Compares the means of a numerical variable between two independent groups
- d) Compares the means of a numerical variable between two similar populations

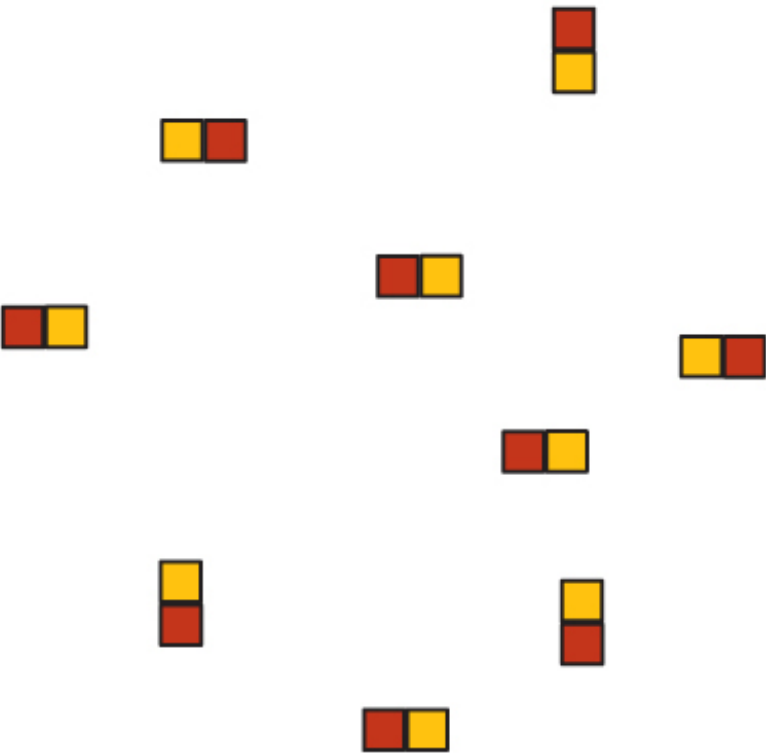
- Tests with one categorical and one numerical variable
 - Goal: compare the mean of a numerical variable among different groups
- Examine two major study design comparisons:
 - Paired Design
 - Two-Sample Design

Paired vs. Two Sample Comparisons

Two-sample



Paired



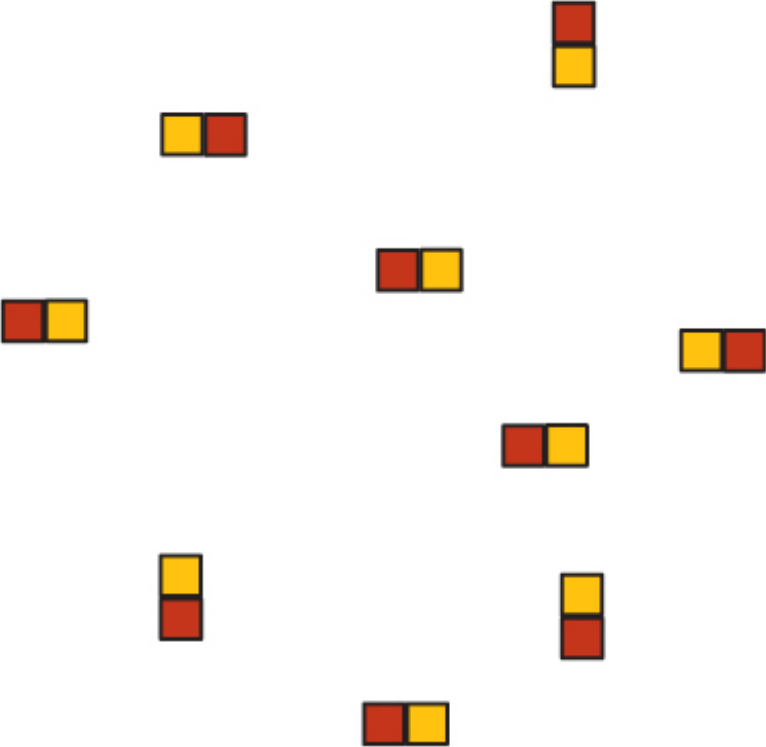
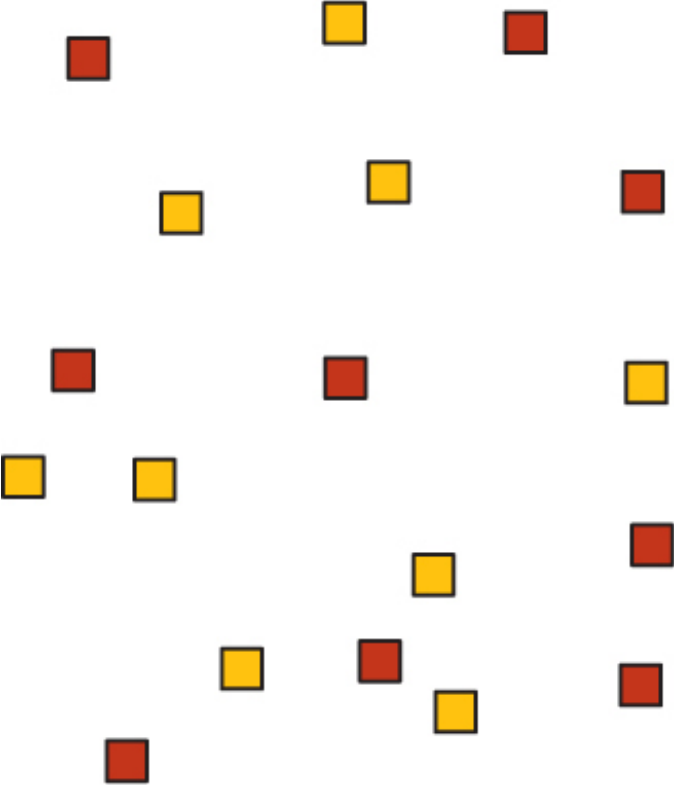
Easier

Paired vs. Two Sample Comparisons

Two-sample

Better

Paired



Do scores on a test of science achievement differ for female and male 8th grade students who attend gender-segregated schools?

- a.** Test one mean against a hypothesized constant.
- b.** Test the difference between two means (independent samples).
- c.** Test the difference in means between two paired or dependent samples.
- d.** Use a chi-squared test of association.

Do boy/girl twins who attend the same school differ in reading achievement?

- a.** Test one mean against a hypothesized constant.
- b.** Test the difference between two means (independent samples).
- c.** Test the difference in means between two paired or dependent samples.
- d.** Use a chi-squared test of association.

Paired Design:

- Data from two groups are paired
- Each member of the pair shares everything in common with the other except for the tested categorical variable

→ Reduces effects of (hidden) confounding variables

- One-to-one correspondence between the individuals in the two groups

Examples:

- **Before** and **After** treatments
- **Identical twins:** one with treatment and one without
- One arm given treatment (sunscreen) the other arm is not on the **same individual**
- Testing effects of treatment (smoking) in a sample of patients, each of which is compared to a non-treatment (nonsmoker) closely **matched** by age, weight, ethnic background and socioeconomic condition*

Paired Design:

- The sampling unit *is the pair: one member with a treatment and a second with a different treatment*
- two measurements must be reduced to a single number which is the mean of the difference between the two measurements
 - Ex: If there are 46 individuals grouped into 23 pairs and $n = ?$

Paired Design:

- Strategy: Compares the mean of the differences to a value given in the null hypothesis
- Often tests the null hypothesis that the mean difference of paired measurements is equal to “0”

For each pair, calculate the difference. The paired t-test is then simply the one-sample t-test on the differences

Paired Design:

- Strategy: Compares the mean of the differences to a value given in the null hypothesis
- Often tests the null hypothesis that the mean difference of paired measurements is equal to “0”

For each pair, calculate the difference. The paired t-test is then simply the one-sample t-test on the differences

Assumptions:

- **Difference** between members of each pair have a normal distribution
 - Distribution of the single measurements on each sampling unit do not need to be normally distributed – only the differences must be normally distributed
- Pairs are chosen at random

Example: No Smoking Day in Great Britain on the second Wednesday of March. Compared to the previous Wednesday, does (voluntarily) not smoking for a day affect injury rate?

Year	Injuries Previous Wednesday	Injuries “No Smoking” Wednesday
1987	516	540
1988	610	620
1989	581	599
1990	586	639
1991	554	607
1992	632	603
1993	479	519
1994	583	560
1995	445	515
1996	522	556

Example: No Smoking Day in Great Britain on the second Wednesday of March. Compared to the previous Wednesday, does (voluntarily) not smoking for a day affect injury rate?

Step 1: formulate null hypothesis

H_0 : There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d = 0$

H_A : There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d \neq 0$

Example: No Smoking Day in Great Britain on the second Wednesday of March. Compared to the previous Wednesday, does (voluntarily) not smoking for a day affect injury rate?

Step 1: formulate null hypothesis

H_0 : There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d = 0$

H_A : There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d \neq 0$

Step 2: Identify test statistic and determine if assumptions are met

- paired t-test
- Assumptions: d_i are normally distributed
- Calculate \bar{d} , $SE_{\bar{d}}$

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}}$$

- Remember: n is the number of pairs, which are the independent sampling units

Paired Design: No Smoking Wednesday Injuries

H_0 : There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d = 0$

H_0 : There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d \neq 0$

Test statistic & assumptions: paired t-test

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}}$$

Year	Injuries Previous Wed.	Injuries “No Smoking” Wed.	d_i
1987	516	540	24
1988	610	620	10
1989	581	599	18
1990	586	639	53
1991	554	607	53
1992	632	603	-29
1993	479	519	40
1994	583	560	-23
1995	445	515	70
1996	522	556	34

Paired Design: No Smoking Wednesday Injuries

H₀: There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday,

$$\mu_d = 0$$

H₀: There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday,

$$\mu_d \neq 0$$

Test statistic & assumptions:

- paired t-test

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}}$$

Year	Injuries Previous Wed.	Injuries “No Smoking” Wed.	d _i
1987	516	540	24
1988	610	620	10
1989	581	599	18
1990	586	639	53
1991	554	607	53
1992	632	603	-29
1993	479	519	40
1994	583	560	-23
1995	445	515	70
1996	522	556	34

$$\bar{d} = 25$$

$$SE_{\bar{d}} = 10.22$$

$$n = 10$$

$$dof = 9$$

Paired Design: No Smoking Wednesday Injuries

H_0 : There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d = 0$

H_0 : There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d \neq 0$

Test statistic & assumptions:

- paired t-test

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}}$$

$$\begin{aligned}\bar{d} &= 25 \\ SE_{\bar{d}} &= 10.22 \\ dof &= 9\end{aligned}$$

$$t = \frac{25 - 0}{10.22} = 2.45$$

Critical Values and Significance levels:

$$\alpha = 0.05$$

$$t_{0.05(2), 9} > \mathbf{2.26}$$

We can reject the H_0 : there is a difference in injury rate between days on which people smoke and those on which people don't

Paired Design: No Smoking Wednesday Injuries

H_0 : There is no difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d = 0$

H_a : There is a difference in number of injuries experienced on “No smoking” Wednesday and a regular Wednesday, $\mu_d \neq 0$

Test statistic:

- paired t-test

$$t = \frac{25 - 0}{10.22} = 2.45$$

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}}$$

$$\begin{aligned}\bar{d} &= 25 \\ SE_{\bar{d}} &= 10.22 \\ df &= 9\end{aligned}$$

Critical Values and Significance levels:

$$\alpha = 0.05$$

$$t_{0.05(2), 9} > 2.26$$

We can reject the H_0 : there is a difference in injury rate between Days on which people smoke and those on which people don't

$$\bar{d} - t_{\alpha(2), df} SE_{\bar{d}} < \mu_d < \bar{d} + t_{\alpha(2), df} SE_{\bar{d}}$$

$$1.9 < \mu_d < 48.1$$

Two-sample Design:

- **Assumptions:**

- **Random Sample** is **normally distributed** in both populations --> sampling distribution for difference between sample means is also normal
- **Standard deviation** is the same in both populations --> if this is not true, use Welch's approximate t-test instead*

- Strategy:

- Unlike in a paired t-test, there are two variables from two entirely different populations. Instead of one variable describing the difference, \bar{d} , you have two: $\bar{Y}_1 - \bar{Y}_2$
- Standard Error of $\bar{Y}_1 - \bar{Y}_2$ is **pooled!**

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- *two-sample t-test is robust to violations of assumptions if n is similar between the two groups.*

Two-sample Design:

Pooled sample variance:

- Weighted average; the average of the variances of the samples weighted by their degrees of freedom

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

- tangent: what is the “pooled” variance doing?
 - allowing us to access and use the additional information that is in our sample
 - We will something similar in ANOVA

BEHRENS-FISHER problem

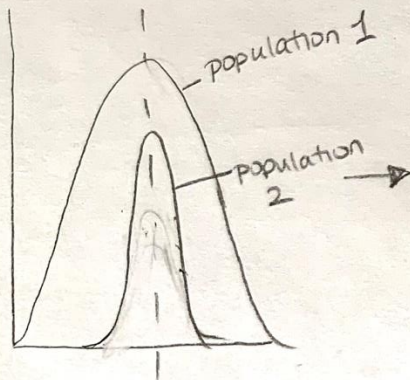
• When variances of two populations are not equal

• we can illustrate the problem with two extreme situations

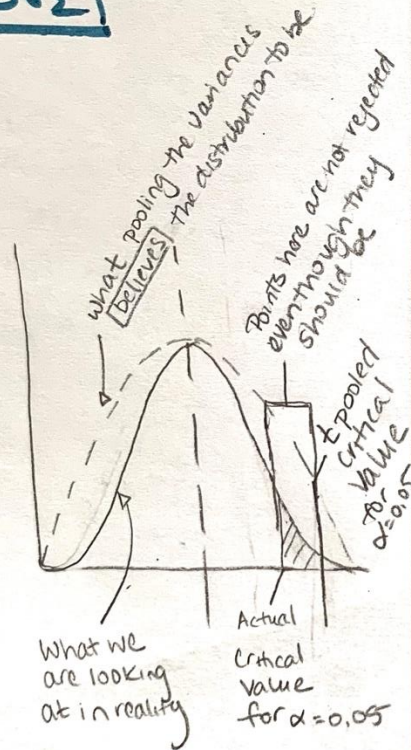
• Two sample t-test is not robust when $S_1^2 \neq S_2^2$ and $n_1 \neq n_2$

Situation 1

larger sample has larger variance:
 $n_1 \gg n_2 ; \sigma_1 > 3\sigma_2$

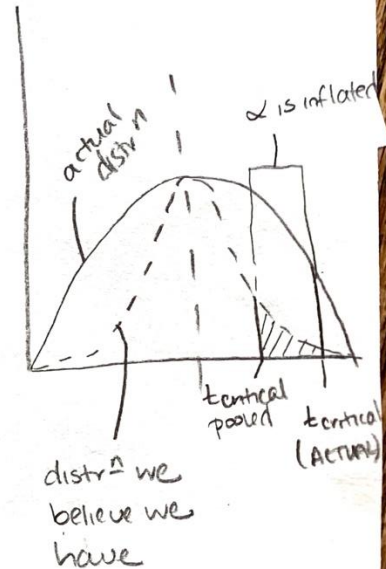
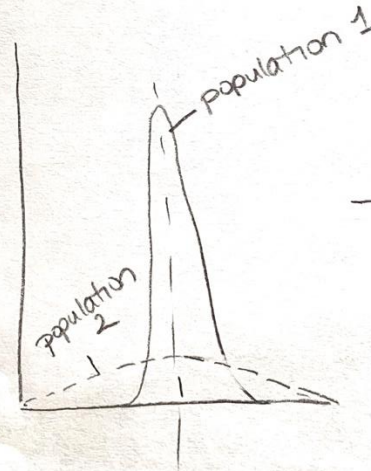


• pooled variance allows the larger sample, with its much larger variance, to contribute more



Situation 2

larger sample has smaller variance:
 $n_1 \gg n_2 ; \sigma_2 > 3\sigma_1$

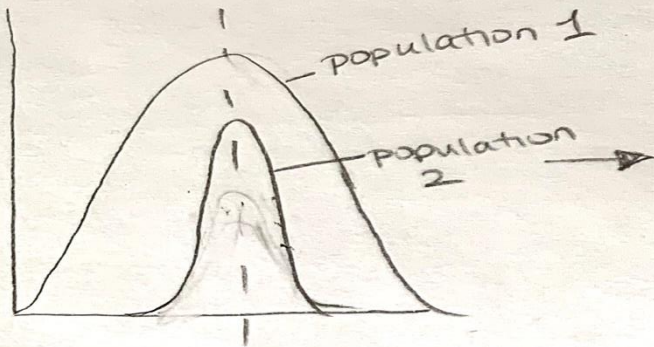


BEHRENS- problem

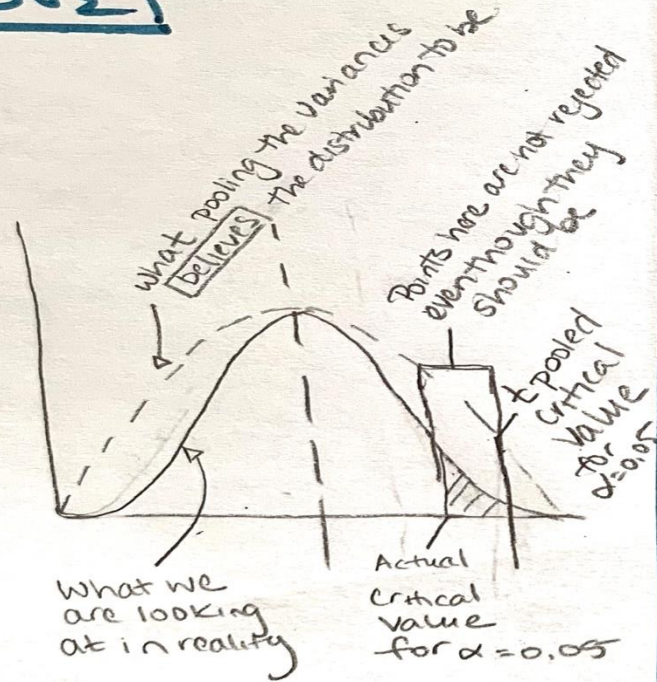
• Two sample t-test is n
Situation 1

larger sample has larger variance:

$$n_1 \gg n_2 ; \sigma_1 > 3\sigma_2$$



• pooled variance. allows
the larger sample, with
its much larger variance,
to contribute more
than it actually does



FISHER



• When variances of two populations are not equal

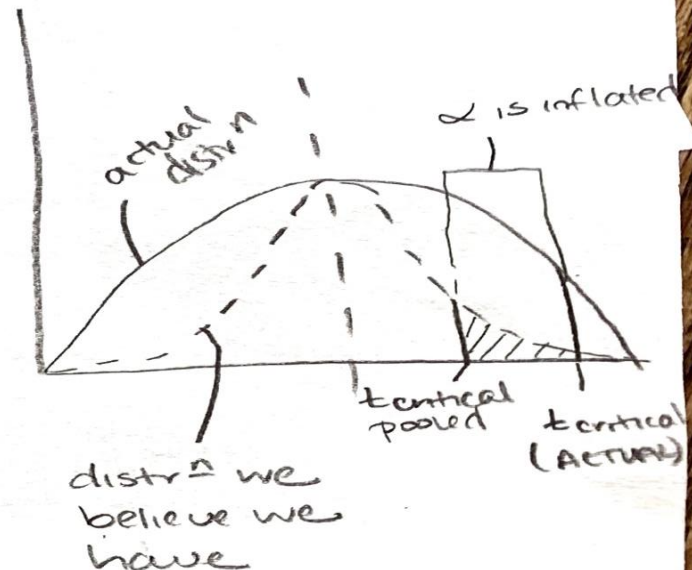
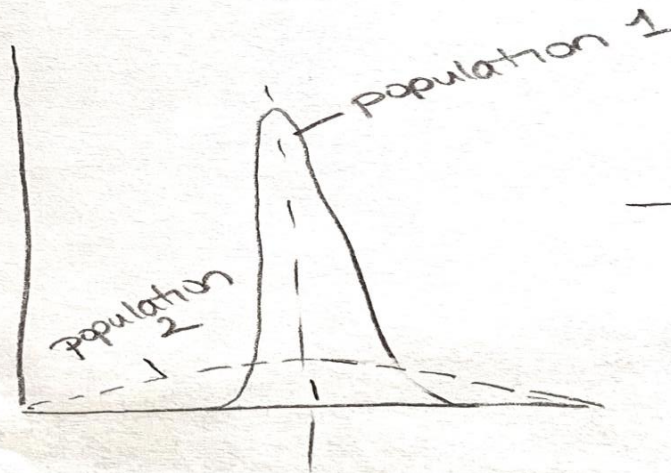
• we can illustrate the problem with two extreme situations

not robust when $S_1^2 \neq S_2^2$ and $n_1 \neq n_2$

Situation 2

larger sample has smaller variance

$$n_1 \gg n_2 ; \sigma_2 > 3\sigma_1$$



What is the total degree of freedom and the standard error of difference between $\text{mean}_1 = 32$ and $\text{mean}_2 = 30$, given that $n_1 = 10$, $n_2 = 15$; $s_1 = 2$ and $s_2 = 3$. (The variable in both population is normally distributed)

- A. $df = 23$; $SE = 1.13$
- B. $df = 24$; $SE = 1.13$
- C. $df = 23$; $SE = 1.08$
- D. $df = 24$; $SE = 1.08$
- E. $df = 23$; $SE = 0.66$

、
 $SE = \sqrt{sp^2(1/n_1 + 1/n_2)}$

Two-sample Design:

Student's t-distribution of two-sample design:

- Compares the means of a numerical variable between two populations

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

Total degrees of freedom:

$$df = df_1 + df_2 = n_1 + n_2 - 2$$

- Two means are estimated, so subtract 2

Two-sample Design:

Example: 2 genotypes of lettuce: *susceptible* and *resistant*. Do these genotypes differ in fitness in the absence of aphids.

The proxy for fitness that is measured are number of buds.



Two-sample Design:

Example: 2 genotypes of lettuce: *susceptible* and *resistant*. Do these genotypes differ in fitness in the absence of aphids.

	<i>Susceptible</i>	Resistant
Mean number of buds	720	582
SD of number of buds	223.6	277.3
Sample size	15	16

Both distributions are normally distributed

Two-sample Design:

Example: 2 genotypes of lettuce: *susceptible* and *resistant*. Do these genotypes differ in fitness in the absence of aphids.

	<i>Susceptible</i>	<i>Resistant</i>
Mean number of buds	720	582
SD of number of buds	223.6	277.3
Sample size	15	16

H_0 : There is no difference between the number of buds in susceptible and resistant plants ($\mu_1 = \mu_2$)

H_A : There is a difference between the number of buds in susceptible and resistant plants ($\mu_1 \neq \mu_2$)

Two-sample Design:

Example: 2 genotypes of lettuce: *susceptible* and *resistant*. Do these genotypes differ in fitness in the absence of aphids.

	<i>Susceptible</i>	<i>Resistant</i>
Mean number of buds	720	582
SD of number of buds	223.6	277.3
Sample size	15	16

H_0 : There is no difference between the number of buds in susceptible and resistant plants ($\mu_1 = \mu_2$)

H_A : There is a difference between the number of buds in susceptible and resistant plants ($\mu_1 \neq \mu_2$)

t-test:

$$df = 15 + 16 - 2 = 29$$

$$\alpha = 0.05$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{14(223.6)^2 + 15(277.3)^2}{14 + 15} = 63909.9$$

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{63909.9 \left(\frac{1}{15} + \frac{1}{16} \right)} = \sqrt{8255.02} = 90.86$$

Comparing Two Means

Two-sample Design:

Example: 2 genotypes of lettuce: *susceptible* and *resistant*. Do these genotypes differ in fitness in the absence of aphids.

	<i>Susceptible</i>	<i>Resistant</i>
Mean number of buds	720	582
SD of number of buds	223.6	277.3
Sample size	15	16

H₀: There is no difference between the number of buds in susceptible and resistant plants ($\mu_1 = \mu_2$)

H_A: There is a difference between the number of buds in susceptible and resistant plants ($\mu_1 \neq \mu_2$)

Two sample t-test:

Assumptions have been met

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE_{\bar{Y}_1 - \bar{Y}_2}} = \frac{(720 - 582)}{90.86} = 1.52$$

$$\alpha = 0.05$$

Critical value: $t_{0.05(2),29} = 2.05$: $t < 2.05$ so Fail to reject. These data are not sufficient to say there is resistance

Confidence Interval: Two-sample Design:

$$(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha(2), df} SE_{\bar{Y}_1 - \bar{Y}_2} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha(2), df} SE_{\bar{Y}_1 - \bar{Y}_2}$$

$$138 - 2.05(90.86) < \mu_1 - \mu_2 < 138 + 2.05(90.86) \\ -48.21 < \mu_1 - \mu_2 < 324.26$$

Note: this interval includes 0 which supports our conclusion (FTR)

Assumptions of parametric tests:

- Random Samples
- Populations are normally distributed
- for two sample t-test: Populations have equal(ish) variances
 - if not → **Welch's approximate t-test**
 - How do we tell when populations don't have equal variances?

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

Methods:

- 1. The F-Test of equal variances**
- 2. Levene's test for homogeneity of variances**

Comparing Variances:

Hypotheses:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

Methods:

1. The F-Test of equal variances

$$F = \frac{s_1^2}{s_2^2}$$

Basis of ANOVA 

- If variances are equal, this should be **1**
- Put larger sample variance on top
 - forces it to be a one tailed test
- two different degrees of freedom:
 - **df = n_i - 1**
- very sensitive to assumption that both populations are normally distributed

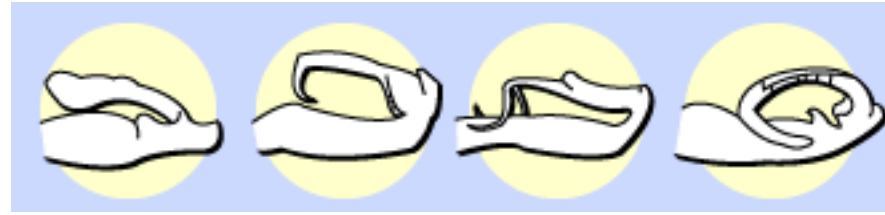
2. Levene's test for homogeneity of variances

Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$



Example: Variation in insect genitalia (the above picture is from damsel fly) between polygamous species and monogamous species of insects: Lock and key or sexual selection? Goran Arnqvist expresses this data as a 'morphometric dimension' and we know that it is a normally distributed variable.

	Polygamous	Monogamous
Mean	-19.3	10.25
Sample Variance	243.9	2.27
Sample Size	7	9

Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

	Polygamous	Monogamous
Mean	-19.3	10.25
Sample Variance	243.9	2.27
Sample Size	7	9

Example: Variation in insect genitalia between polygamous species and monogamous species.

$$s_1^2 = 243.9$$

$$s_2^2 = 2.27$$

$$F = \frac{s_1^2}{s_2^2} = \frac{243.9}{2.27} = 107.4$$

Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

	Polygamous	Monogamous
Mean	-19.3	10.25
Sample Variance	243.9	2.27
Sample Size	7	9

Example: Variation in insect genitalia between polygamous species and monogamous species.

$$F = \frac{s_1^2}{s_2^2} = \frac{243.9}{2.27} = 107.4$$

- $df_1 = 7 - 1 = 6$

- $df_2 = 9 - 1 = 8$

$$F_{0.025,6,8} = 4.7$$

Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

Example: Variation in insect genitalia between polygamous species and monogamous species.

Why is the critical value $\alpha/2$?

- by putting the larger variance in the numerator, we are forcing F to be greater than 1
- By the null hypothesis there is a 50:50 chance of either s^2 being greater so we want the higher tail to include just $\alpha/2$
- *NOTE:* This is different than when the F ratio is used in ANOVA – because of the design of the ratio, MS_{group} will always be in numerator so it is just α

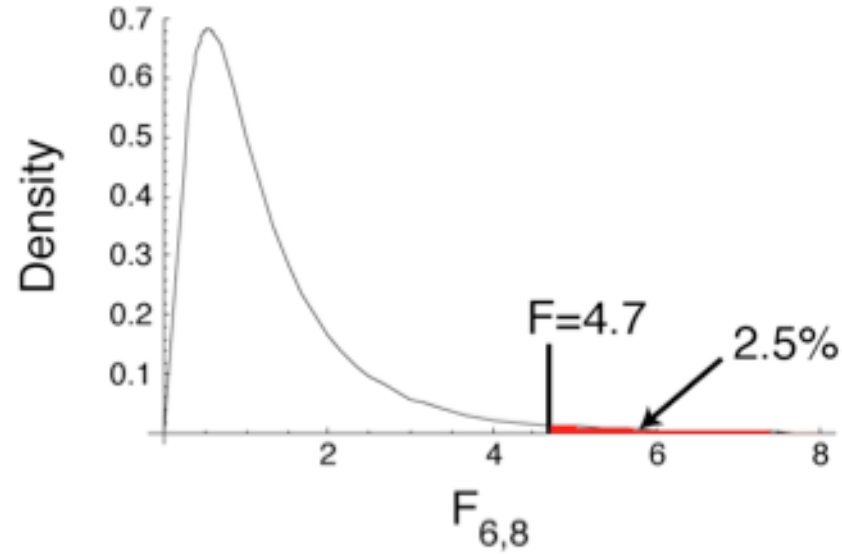
Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

Example: Variation in insect genitalia between polygamous species and monogamous species.



Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

Example: Variation in insect genitalia between polygamous species and monogamous species.

Conclusion:

- The $F = 107.4$ from the data is greater than the critical value for F with 6 and 8 dof so we can reject the null hypothesis that the variances of the two groups are equal.
- The variance in insect genitalia is much greater for polygamous species than monogamous species.

You could confirm this by determining the Confidence Interval of each of the two variances and seeing if they overlap...

$$\underline{6 \cdot 243.9} < \sigma^2_1 < \underline{6 \cdot 243.9}; \quad \underline{8 \cdot 2.27} < \sigma^2_2 < \underline{8 \cdot 2.27}$$

$$X^2_{0.025,6}$$

$$X^2_{0.975,6}$$

$$X^2_{0.025,8}$$

$$X^2_{0.975,8}$$

$$101.27 < \sigma^2_1 < 1180.16; \quad 1.04 < \sigma^2_2 < 8.33$$

Comparing Variances:

Hypotheses:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_A: \sigma^2_1 \neq \sigma^2_2$$

Methods:

1. The F-Test of equal variances

2. Levene's test for homogeneity of variances

- more robust than F-test
- calculations are complex so you should know that it exists and why you would use it but you would probably use R or Python library to conduct the test.

3. Bartlett's test, Hartley's etc.

- they each have specific assumptions that must be met.

Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have vastly unequal variances
- Often used as the 'default' method by programs such as R since it accounts for the Behrens-Fisher problem when variances are very unequal and still gives correct type I error rate when variances are equal (ish).



Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances

• Experimental Design:

- 20 randomly selected burrowing owl nests
- Randomly divided into two groups of 10 nests each
- One group given extra dung; the other not
- Counted number of dung beetles in owls' diets

Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances

• Experimental Design:

- 20 randomly selected burrowing owl nests
- Randomly divided into two groups of 10 nests each
- One group given extra dung; the other not
- Counted number of dung beetles in owls' diets

• Summary of beetles consumed:

	Dung added	No Dung added
Mean	4.8	0.51
s	3.26	0.89

Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances

- Summary of beetles consumed:

	Dung added	No Dung added
Mean	4.8	0.51
s	3.26	0.89

- Hypotheses:

- H_0 : Owls catch the same number of dung beetles with or without extra dung ($\mu_1 = \mu_2$)
- H_A : Owls do not catch the same number of dung beetles with or without extra dung ($\mu_1 \neq \mu_2$)

Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances
- Complicated formula, but it works via tweaking (rounding down) the DEGREES OF FREEDOM

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

Welsh's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances
- Summary of beetles consumed:

	Dung added	No Dung added
Mean	4.8	0.51
s	3.26	0.89

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4.8 - 0.51}{\sqrt{\frac{3.26^2}{10} + \frac{0.89^2}{10}}} = 4.01$$

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\frac{s_1^2}{n_1}}{n_1 - 1} + \frac{\frac{s_2^2}{n_2}}{n_2 - 1}} = \frac{\frac{3.26^2}{10} + \frac{0.89^2}{10}}{\frac{\frac{3.26^2}{10}}{9} + \frac{\frac{0.89^2}{10}}{9}} = 10.33$$

More Comparing Two Means

Welch's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4.8 - 0.51}{\sqrt{\frac{3.26^2}{10} + \frac{0.89^2}{10}}} = 4.01$$

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\frac{s_1^2}{n_1}}{n_1 - 1} + \frac{\frac{s_2^2}{n_2}}{n_2 - 1}} = \frac{\frac{3.26^2}{10} + \frac{0.89^2}{10}}{\frac{3.26^2}{9} + \frac{0.89^2}{9}} = 10.33$$

(you might notice it looks an awful lot like the other t statistic confidence intervals)

$$(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha(2), df} SE_{\bar{Y}_1 - \bar{Y}_2} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha(2), df} SE_{\bar{Y}_1 - \bar{Y}_2}$$

Welsh's approximate t-test:

- it is used when comparing means of two populations that are normally distributed but have unequal variances

• Conclusion:

$$t_{0.05(2),10} = 2.23$$

Since $t = 4.01 > 2.23$, we can reject the null hypothesis with $P < 0.05$.

$$(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha(2),df} SE_{\bar{Y}_1 - \bar{Y}_2} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha(2),df} SE_{\bar{Y}_1 - \bar{Y}_2}$$

$$4.29 - 2.23*(1.069) < \mu_1 - \mu_2 < 4.29 + 2.23(1.069)$$

$$1.91 < \mu_1 - \mu_2 < 6.67$$

The above range does not contain “0” so it supports our rejection of the null hypothesis AND it is greater than “0” so extra dung near burrowing owls' nests increases the number of dung beetles eaten.

Four general ways to address violations:

- **Ignore**

- Sometimes you can use a method even if assumptions are violated
- Thank the Central Limit Theorem for robustness
 - *means of large samples are normally distributed*
 - **tests that are based on** comparing sample **means** will be robust when they have sufficient sample size
 - **this doesn't help with F-test etc.**
- Especially true if sample sizes are large ($n_i \gg 50$) and violations are not extreme
 - **sample size must increase to accommodate how extreme violations are between groups**
 - especially if two samples both differ in opposite directions
- Even with CLT we can't always ignore:
 - **outliers**
 - **frequency distribution between groups is very different**

Four general ways to address violations:

- Ignore
 - sometimes you can use a method even if assumptions are violated
 - especially true if sample sizes are large and violations are not extreme
- Transform
 - attempt to force normality and other assumptions onto data
 - We will investigate various tools
 - Usually boils down to: **take the log of the data**
 - *Changes each measurement in the same way (1 to 1 correspondence) so that you can **transform back to get original data without ambiguity***
 - **monotonic relationship with original values**
 - remember to transform back the upper and lower limits for a Confidence Interval
 - work that does not always pay off but you at least maintain power of the test that you are using (since nonparameteric tests usually reduce power)

Four general ways to address violations:

- Ignore
 - sometimes you can use a method even if assumptions are violated
 - especially true if sample sizes are large and violations are not extreme
- Transform
 - attempt to force normality and other assumptions onto data
 - We will investigate various tools
 - work that does not always pay off
- Use Non-parametric method
 - classes of methods that do not require assumption of normality
 - not cost free! Often lose power etc

$$\text{Power} = 1 - P[\text{FTR } H_0 | H_0 \text{ is incorrect}] = P[\text{reject} | H_0 \text{ is incorrect}]$$

Four general ways to address violations:

- Ignore
 - sometimes you can use a method even if assumptions are violated
 - especially true if sample sizes are large and violations are not extreme
- Transform
 - attempt to force normality and other assumptions onto data
 - We will investigate various tools
 - work that does not always pay off
- Use Non-parameter method
 - classes of methods that do not require assumption of normality
 - not cost free! Often lose power etc
- Computationally Intensive Methods
 - Simulation
 - Bootstrap
 - randomization/permutation test

Mann-Whitney U-Test

- Compares the central tendencies of two groups using ranks
 - uses ranks of measurements to test whether or not frequency distribution of two groups are the same
 - Small samples lead to little power
 - All group samples are random samples
 - Distribution of the variable has the same shape in every population
- Nonparametric version of two-sample t-test

Mann-Whitney U-Test

Method:

1. Declare hypotheses
2. Rank all individuals from both groups together in order (smallest to largest)
3. Sum ranks for all individuals in each group
4. Calculate test statistic, U

Mann-Whitney U-Test

Method:

1. Declare hypotheses
2. Rank all individuals from both groups together in order (smallest to largest): R_1, R_2, \dots, R_n
3. Sum ranks for all individuals in each group
4. Calculate test statistic, U

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1$$

U_1 is the number of times an individual from pop 1 has a lower rank than an individual from pop 2 out of all pairwise comparisons

Mann-Whitney U-Test

Method:

1. Declare hypotheses
2. Rank all individuals from both groups together in order (smallest to largest): R_1, R_2, \dots, R_n
3. Sum ranks for all individuals in each group
4. Calculate test statistic, U
5. U is the larger of U_1 or U_2
6. Determine the P-value by comparing the observed U with the critical value of the null distribution for U (table E)
7. If both samples have $n > 10$, use Z transformation

$$Z = \frac{2U - n_1n_2}{\sqrt{n_1n_2(n_1 + n_2 + 1)/3}}$$

How to deal with tied ranks:

- Determine the ranks that the values would have been assigned if they were slightly different
- Average these ranks, and assign **that** average to each tied individual
- Count all those individuals when deciding the rank for the next largest individual

Intuitive way to understanding MWU test

- View the MWU test as the ‘extremeness’ of the distributions of the two populations

Example: two populations each have three data points:

1 1 1 0 0 0

There are twenty orderings of rank for this example:

{111 000, 110 100, 110 010, 110 001, 101 100, 101 010, 101 001, 100 110, 100 101 etc.}

Note: you can calculate this using $(6 \text{ choose } 3) = 6*5*4*3*2*1 = 20$

Mann-Whitney U-Test

Assumptions:

- Both samples are random samples
- Both populations have the same shape of distribution
 - **They can both be skewed but it must be in the same direction**
 - **Same variance**

Mann-Whitney U-Test

PP: When intruding lions take over a pride of females, they often kill most or all of the infants in the group in order to reduce the time until females are again sexually receptive. A long term study measured the time to reproduction of female lions after losing cubs to infanticide and compared this to the time to reproduction of females who had lost their cubs to accidents. The data are not normally distributed within the two groups and we have been unable to find a transformation that makes them normal.

Accidental: 110, 117, 133, 135, 140, 168, 171, 238, 255

Infanticide: 211, 232, 246, 251, 275

Mann-Whitney U-Test

PP: When intruding lions take over a pride of females, they often kill most or all of the infants in the group in order to reduce the time until females are again sexually receptive. A long term study measured the time to reproduction of female lions after losing cubs to infanticide and compared this to the time to reproduction of females who had lost their cubs to accidents. The data are not normally distributed within the two groups and we have been unable to find a transformation that makes them normal.

Accidental: 110, 117, 133, 135, 140, 168, 171, 238, 255

Infanticide: 211, 232, 246, 251, 275

1. $H_0: \mu_{\text{accidental}} = \mu_{\text{infanticide}}$

$H_A: \mu_{\text{accidental}} \neq \mu_{\text{infanticide}}$

In words: the mean time to reproduction of female lions who have lost cubs due to accident or infanticide is the same.

Mann-Whitney U-Test

When intruding lions take over a pride of females, they often kill most or all of the infants in the group in order to reduce the time until females are again sexually receptive. A long term study measured the time to reproduction of female lions after losing cubs to infanticide and compared this to the time to reproduction of females who had lost their cubs to accidents. The data are not normally distributed within the two groups and we have been unable to find a transformation that makes them normal.

Accidental: 110, 117, 133, 135, 140, 168, 171, 238, 255

Infanticide: 211, 232, 246, 251, 275

1. $H_0: \mu_{\text{accidental}} = \mu_{\text{infanticide}}$

$H_A: \mu_{\text{accidental}} \neq \mu_{\text{infanticide}}$

2. Rank all individuals from both groups together

110, 117, 133, 135, 140, 168, 171, 211, 232, 238, 246, 251, 255, 275

3. Sum ranks for all individuals in each group

Accidental (R_1): $1+2+3+4+5+6+7+10+13 = 51$

Infanticide (R_2): $8+9+11+12+14 = 54$

Mann-Whitney U-Test

When intruding lions take over a pride of females, they often kill most or all of the infants in the group in order to reduce the time until females are again sexually receptive. A long term study measured the time to reproduction of female lions after losing cubs to infanticide and compared this to the time to reproduction of females who had lost their cubs to accidents. The data are not normally distributed within the two groups and we have been unable to find a transformation that makes them normal.

Accidental: 110, 117, 133, 135, 140, 168, 171, 238, 255

Infanticide: 211, 232, 246, 251, 275

1. $H_0: \mu_{\text{accidental}} = \mu_{\text{infanticide}}$

$H_A: \mu_{\text{accidental}} \neq \mu_{\text{infanticide}}$

2. Rank all individuals from both groups together

110, 117, 133, 135, 140, 168, 171, 211, 232, 238, 246, 251, 255, 275

3. Sum ranks for all individuals in each group

Accidental (R_1): $1+2+3+4+5+6+7+10+13 = 51$

Infanticide (R_2): $8+9+11+12+14 = 54$

4. Calculate test statistic, U (NB: U is larger of U_1 or U_2):

$$U_1 = n_1 n_2 + n_1(n_1 + 1)/2 - R_1 = (9)(5) + (9)(10)/2 - 51 \\ = 45 + 45 - 51 = 39$$

$$U_2 = n_1 n_2 - U_1 = 45 - 39 = 6$$

$$U = 39$$

Mann-Whitney U-Test

When intruding lions take over a pride of females, they often kill most or all of the infants in the group in order to reduce the time until females are again sexually receptive. A long term study measured the time to reproduction of female lions after losing cubs to infanticide and compared this to the time to reproduction of females who had lost their cubs to accidents. The data are not normally distributed within the two groups and we have been unable to find a transformation that makes them normal.

Accidental: 110, 117, 133, 135, 140, 168, 171, 238, 255

Infanticide: 211, 232, 246, 251, 275

1. $H_0: \mu_{\text{accidental}} = \mu_{\text{infanticide}}$

$H_A: \mu_{\text{accidental}} \neq \mu_{\text{infanticide}}$

2. Rank all individuals from both groups together

110, 117, 133, 135, 140, 168, 171, 211, 232, 238, 246, 251, 255, 275

3. Sum ranks for all individuals in each group

Accidental: $1+2+3+4+5+6+7+10+13 = 51$

Infanticide: $8+9+11+12+14 = 54$

4. Calculate test statistic, U (NB: U is larger of U_1 or U_2): **U = 39**

5. Determine critical value from Table E

Critical Value is 38 for $n_1 = 9$, $n_2 = 5$ for $\alpha = 0.05$

Our value of **U** is > Critical value so we can reject the null hypothesis that females have equal times to reproduction regardless of how their previous cub died.

For analysis of means, which of the following statistical procedures assume that the distributions of the variables it measures have the same shape:

- a) Sign test
- b) Mann-Whitney U-test
- c) Log Transformations
- d) Arcsine Transformations