Unit 10.
Two Sample Inference

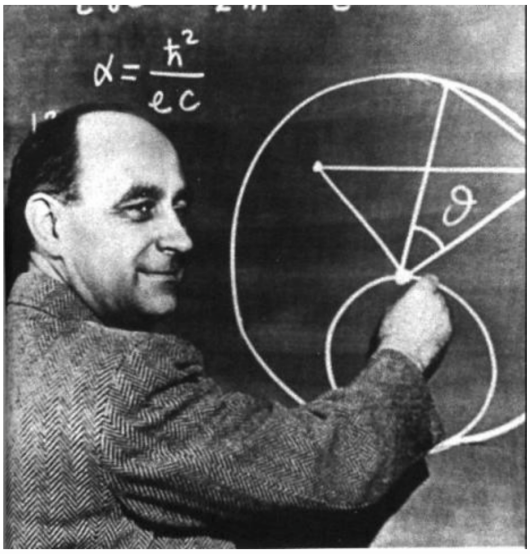
"Who would not say that the glosses (commentaries on the law) increase doubt and ignorance? It is more of a business to interpret the interpretations than to interpret the things"

- Michel De Montaigne (1533-1592)

Cheers!

There are two possible outcomes: If the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.

Enrico Fermi (1901 - 1954)



Enrico Fermi

Source: With permission, download from CAUSEweb.org

Table of Contents

1. Learning Objectives

When you have finished this unit, you should be able to:

- Calculate point and confidence interval estimates of the difference of two independent means, both of Normal distributions.
- Calculate point and confidence interval estimates of the ratio of two independent variances, both of Normal distributions.
- Calculate point and confidence interval estimates of the difference of two independent event probabilities, both of Binomial distributions.
- Perform and interpret the statistical hypothesis tests described in the two sample settings described in these notes.

2. Two Independent Groups – Continuous Outcome. Normal Distribution

In this setting, our focus is on [$\mu_{Group\ 1}$ - $\mu_{Group\ 2}$], the difference between the means of the two independent groups.

Good to know
If the two means are equal, then their difference will be zero.

Example - A randomized controlled trial (RCT) of a new drug.

- Some consenting participants are randomized to the **control** group and are treated with standard care;
- Others are randomized to the **intervention** group and are treated with the trial drug.
- Are intervention group responses (the population mean of these is $\mu_{Intervention}$) different from control group responses (the population mean of these is $\mu_{Control}$)? Typically, the RCT compares the average response of the intervention group with the average response of the control group.
- Because the two groups, control and intervention, are independent, the tools for paired data are not appropriate.
- Our null hypothesis equality of population means, H_O: $\mu_{Intervention} = \mu_{Control}$
- Operationally, we will work with the equivalent null hypothesis that says that the difference between these two means is zero, H_O: $[\mu_{Intervention} - \mu_{Control}] = 0$

Example – An observational study that compares two independent groups.

Is mean blood pressure the same for males and females?

Is body mass index (BMI) similar for breast cancer cases versus non-cancer patients?

Is length of stay (LOS) for patients in hospital "A" the same as that for similar patients in hospital "B"?

- Our null hypothesis is H_O: $\mu_{Hospital\ A} = \mu_{Hospital\ B}$
- Operationally, we will work with the equivalent null hypothesis H_O: $[\mu_{Hospital\ A} - \mu_{Hospital\ B}] = 0$

2a. How to Solve for the Correct Standard Error

In doing a test of equality of independent means (which says that their difference is zero), the test statistic is a standardization of the difference between two sample means $(\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}})$. Thus, we need the standard error of $(\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}})$.

The correct standard error, $SE_{(\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}})}$ depends on whether the variances are known and, if they are not known, whether the unknown variances can be assumed to be equal or must be assumed to be unequal.

Scenario # 1 - Population σ_1^2 and σ_2^2 are both known

$$SE\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Scenario #2 - Population σ_1^2 and σ_2^2 are UNknown but are assumed EQUAL

$$\hat{SE}\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{S_{\text{pool}}^2}{n_1} + \frac{S_{\text{pool}}^2}{n_2}}$$

S_{pool}^2 **is a weighted average of the two separate sample variances, with weights equal to the associated degrees of freedom contributions.**

$$S_{\text{pool}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Scenario #3 - Population σ_1^2 and σ_2^2 are Unknown and are assumed to be NOT EQUAL

$$\hat{SE}\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

HOMEWORK DUE Monday December 5, 2022
Question #1 of 3

See again the Unit 7 notes. This exercise is "connecting the dots" between the ideas of normal random variables and z-scores. So you really should look back at course notes for Unit 7 (Normal Distribution). A good place to re-visit is page 22.

See again the supplementary notes for Unit 8. More to the point, however are the supplementary notes (sums and differences of normals, student-t, chi square, and F) accompanying the unit 8 course materials. **For this exercise, see again pages 21-22!**

The National Health and Nutrition Examination Survey of 1975-1980 give the following data on serum cholesterol levels in US males.

Group	Age, years	Population Mean, μ	Population Standard Deviation, σ
1	20-24	180	43
2	25-34	199	49

Suppose the distribution of serum cholesterol is normal in each age group. If you draw simple random samples of size 50 from each of the two groups, what is the probability that the difference between the two sample means (Group 2 mean – Group 1 mean) will be more than 25?

Source: National Center or Health Statistics, R. Fulwood, W. Kalsbeek, B. Rifkind, et al. "Total serum cholesterol levels of adults 20-74 years of age: United States, 1976-1980". Vital and Health Statistics Series 11, No. 236. DHHS Pub. No (PHD) 86-1686, Public Health Service, Washington, DC, US Government Printing Office, May 1986. Cited in Daniel (p 140, 5.4.1 Copyright 1999 by John Wiley & Sons, Inc. By permission of John Wiley.

2b. Hypothesis Test for $[\mu_1 - \mu_2] = 0$

Tips -
There are 3 t-tests here. The correct one depends on what is known or assumed about the population variances, which in turn tells you what to use as the standard error.

Scenario #1 - If the population variances are KNOWN,
THEN a test of equality of means utilizes p-value calculations using the standard normal distribution.

Scenario #2 - If the population variances are NOT KNOWN but assumed EQUAL,
THEN a test of equality of means utilizes a pooled estimate of the common variance and p-value calculations using the student-t distribution. The degrees of freedom are $(n_1 - 1) + (n_2 - 1)$

Scenario #3 - If the population variances are NOT KNOWN and assumed UNEQUAL,
THEN a test of equality of means utilizes the separate sample variances and p-value calculations using the student-t distribution. The degrees of freedom are given by Satterthwaite's formula (it's nasty; see page 15)

Note – A test of equality of variances is given in Section 2d.

Before you begin - See again course notes *8. Statistical Literacy – Estimation and Hypothesis Testing* pp 21-22 for a review of sums and differences of independent Normal random variables

The following three scenarios are considered:

- i. The two population variances are KNOWN
- ii. The two population variances are UNKNOWN, but assumed equal
- iii. The two population variances are UKNOWN and assumed UNEQUAL

Example Scenario i. The Population Variances are KNOWN:

(*Note: These data are hypothetical.*)

Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those not receiving zidovudine to see if zidovudine is *beneficial*. We may assume that the scores are normally distributed with distributions with KNOWN variances $\sigma_1^2 = 40^2$ and $\sigma_2^2 = 35^2$. The data summaries are the following:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\bar{X}_1 = 120$	$\bar{X}_2 = 96$

Research Question.

Do patients receiving zidovudine have higher functional status scores?

Null Hypothesis Assumptions.

$\bar{X}_1$ is distributed Normal (μ_1 , $\sigma^2/15$) and $\bar{X}_2$ is distributed Normal (μ_2 , $\sigma^2/22$)

H_O and H_A.

$H_O : \mu_1 = \mu_2$
$H_A : \mu_1 > \mu_2$ one sided. The investigator is researching treatment *benefit*

The test statistic is a z-score.

$$z_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2)|H_O \text{true}]}{SE[(\bar{X}_1 - \bar{X}_2)|H_O \text{true}]} \right]$$

Solution (yes – we can solve for it!) the standard error of $\left(\bar{X}_1 - \bar{X}_2\right)$?

$$SE\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{40^2}{15} + \frac{35^2}{22}}$$

$$= 12.7416$$

Proof by contradiction reasoning and the definition of the p-value calculation.

Under the null hypothesis model, the chances of this extremeness, or more so, is

$$\text{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq \left(120 - 96\right) | \, H_O \text{ is true}\right].$$

Calculations.

$$\text{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq \left(120-96\right) \mid \text{H}_\text{O}\text{is true}\right]$$

$$= \Pr\left[\frac{\left(\bar{X}_1 - \bar{X}_2\right)-\left(0\right)}{SE\left(\bar{X}_1 - \bar{X}_2\right)} \geq \frac{\left(120-96\right)-\left(0\right)}{12.7416}\right]$$

$$= \Pr\left[z_{score} \geq 1.8836\right]$$

$$= .0298$$

Note: z_{score}=1.88 says "the observed difference in average functional status scores equal to (120-96) = 24 is 1.88 standard error units away from (greater than) the null hypothesis expected difference of 0."

"Evaluate".

Application of the null hypothesis assumption has led to an unlikely result. Under the null hypothesis H_O, the chances that the 15 patients in the zidovudine treated group would have a mean score that is (120-96)=24 points higher than the average of the 22 scores among the control group is approximately 3 in 100. This is a small chance, suggesting that we abandon the null hypothesis. → Reject the null hypothesis.

Interpret.

These data provide statistically significant evidence that treatment with zidovudine improves functional status.

Example Scenario ii. Population Variances UNKNOWN but Assumed Equal ($\sigma_1^2 = \sigma_2^2$):
(*Note: These data are hypothetical.*)

Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those not receiving zidovudine to see if zidovudine is *beneficial*. We may assume that the scores are normally distributed with distributions that have the ***same*** variance σ^2. However, the common σ^2 is unknown.

Now suppose that the data summaries are the following:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\bar{X}_1 = 120$	$\bar{X}_2 = 96$
$S_1 = 40$	$S_2 = 35$

Research Question.

Do patients receiving zidovudine have higher functional status scores?

Null Hypothesis Assumptions.

$\bar{X}_1$ is distributed Normal (μ_1 , $\sigma^2/15$) and $\bar{X}_2$ is distributed Normal (μ_2 , $\sigma^2/22$)

H_O and H_A.

$H_O : \mu_1 = \mu_2$
$H_A : \mu_1 > \mu_2$ one sided. The investigator is researching treatment benefit

The test statistic is a t-score when the variance is UNknown.

$$t_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) \; - \; E[(\bar{X}_1 - \bar{X}_2)|H_O true]}{\hat{SE}[(\bar{X}_1 - \bar{X}_2)|H_O true]} \right]$$

Estimate of the standard error of $\left(\bar{X}_1 - \bar{X}_2\right)$?

Reminder: We use the caret notation (^) as a reminder that what we are using is an estimate.
Since we are assuming that the unknown variances are the same, there is just one thing to estimate. We use the following estimate.

$$\hat{SE}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} \quad \textbf{where}$$

$$S^2_{pool} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

For these data:

$$\hat{\sigma}^2 = S^2_{pool} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)} = \frac{(15-1)40^2 + (22-1)35^2}{(15-1)+(22-1)} = 1375$$

$$S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} = \sqrt{\frac{1375}{15} + \frac{1375}{22}} = 12.42;$$

Degrees of freedom = $(n_1\text{-}1) + (n_2\text{-}1) = (15\text{-}1) + (22\text{-}1) = 35$.

Proof by contradiction reasoning and the definition of the p-value calculation.

Under the null hypothesis model, the chances of this extremeness, or more so, *in the direction of the alternative hypothesis*, is

$$\text{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq (120-96) \mid H_o true\right].$$

Calculations.

$$\textbf{p-value} = \Pr\left[\left(\bar{X}_1 - \bar{X}_2\right) \geq (120-96) \mid \text{H}_\text{O} \text{ is true}\right]$$

$$= \Pr\left[\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (0)}{S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right)} \geq \frac{(120-96)-(0)}{12.42}\right]$$

$$= \Pr\left[t_{score} \geq 1.93\right] \quad \textbf{where degrees of freedom = 35}$$

$$= .03 \quad \textbf{which is pretty close to the p-value = .0298 we got using the Z-test!}$$

Note. t_{score}=1.93 says "the observed difference in average functional status scores equal to (120-96) = 24 is 1.93 standard error units away from (greater than) the null hypothesis expected difference of 0."

"Evaluate".

Application of the null hypothesis assumption has led to an unlikely result. Under the null hypothesis H_O, the chances that the 15 patients in the zidovudine treated group would have a mean score that is (120-96)=24 points higher than the average of the 22 scores among the control group is 3 in 100. This is a small chance, suggesting that we abandon the null hypothesis. → Reject the null hypothesis.

Interpret.

These data provide statistically significant evidence that treatment with zidovudine improves functional status.

Example Scenario iii. Population Variances UNKNOWN and Assumed Unequal ($\sigma_1^2 \neq \sigma_2^2$):

The analysis is slightly different when the variances are unequal for two reasons.

- The estimated SE should reflect the dissimilarity of the variances.
- With a larger # of unknowns, our degrees of freedom should be smaller.

Suppose that the data summaries are the same (so we can appreciate the comparison):

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\bar{X}_1 = 120$	$\bar{X}_2 = 96$
$S_1 = 40$	$S_2 = 35$

Our test statistic is still a t-score and has the same structure:

$$t_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2) \mid H_O true]}{\hat{SE}[(\bar{X}_1 - \bar{X}_2) \mid H_O true]} \right]$$

The estimate of the SE is now the following.
Since the unknown variances are NOT assumed to be equal, we use the following estimate.

$$\hat{SE}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$ **For these data,**

$$= \sqrt{\frac{40^2}{15} + \frac{35^2}{22}}$$

$$= 12.7416$$

Eeew…. We have to use the Satterthwaite formula for the degrees of freedom:

Degrees of freedom $= \dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{(n_1-1)}+\dfrac{\left(S_2^2/n_2\right)^2}{(n_2-1)}}$. In this example we get

$$= \frac{\left(\dfrac{40^2}{15}+\dfrac{35^2}{22}\right)^2}{\dfrac{\left(40^2/15\right)^2}{(14)}+\dfrac{\left(35^2/22\right)^2}{(21)}} = 27.44 \approx 27 \quad \textit{psst - round DOWN}$$

Thus,

$$\text{p-value} = \Pr\left[\left(\bar{X}_1-\bar{X}_2\right)\geq(120-96)\,|\,\text{H}_\text{O}\text{ is true}\right]$$

$$= \Pr\left[\frac{\left(\bar{X}_1-\bar{X}_2\right)-(0)}{\hat{SE}\left(\bar{X}_1-\bar{X}_2\right)}\geq\frac{(120-96)-(0)}{12.74}\right]$$

$$= \Pr\left[t_{score}\geq 1.88\right] \quad \textbf{where degrees of freedom = 27}$$

$= .035$ which is larger than both the Z-test p-value=.0298 and the previous t-test p-value=.03

Interpret.

The conclusion is the same – this is statistically significant evidence of a benefit of zidovudine on functional status.

2c. Confidence Interval for $[\mu_1 - \mu_2]$

Tips -
Again, there are 3 solutions, depending on what is known or assumed about the population variances, which in turn tells you what to use as the standard error.

Scenario #1 - If the population variances are KNOWN,
THEN a confidence interval estimate of the difference in means utilizes percentiles from the standard normal distribution.

Scenario #2 - If the population variances are NOT KNOWN but assumed EQUAL,
THEN a confidence interval estimate of the difference in means utilizes a pooled estimate of the common variance and percentiles from the student-t distribution. The degrees of freedom are $(n_1 - 1) + (n_2 - 1)$

Scenario #3 - If the population variances are NOT KNOWN and assumed UNEQUAL,
THEN a confidence interval estimate of the difference in means utilizes the separate sample variances and percentiles from the student-t distribution. The degrees of freedom are given by Satterthwaite's formula (see again, page 11)

Note – A test of equality of variances is given in Section 2d.

Again, the following three scenarios are considered:

- i. The two population variances are KNOWN
- ii. The two population variances are UNKNOWN, but assumed equal
- iii. The two population variances are UKNOWN and assumed UNEQUAL

Example Scenario i. The Population Variances are KNOWN:

Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those not receiving zidovudine. We may assume that the scores are normally distributed with distributions with KNOWN variances $\sigma_1^2 = 40^2$ and $\sigma_2^2 = 35^2$. The data summaries are the following:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\overline{X}_1 = 120$	$\overline{X}_2 = 96$

Point Estimate: We want a point estimate of the difference [$\mu_{Group\ 1}$ - $\mu_{Group\ 2}$]

- Our point estimator will be the difference between sample means, $[\bar{X}_{Group\ 1} - \bar{X}_{Group\ 2}]$

Solution:

$$\left[\hat{\mu}_{Group1} - \hat{\mu}_{Group2}\right] = [\bar{X}_{Group\ 1} - \bar{X}_{Group\ 2}] = [120 - 96] = 24$$

Standard Error of the Point Estimate:
As before (see again, page 9)

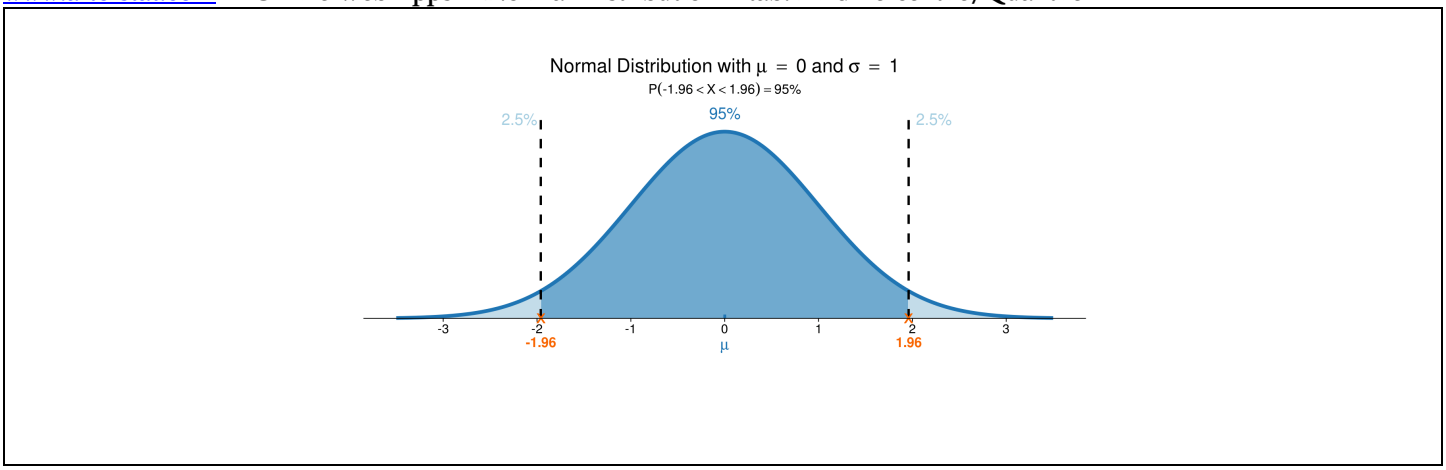
$$SE\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{40^2}{15} + \frac{35^2}{22}}$$

$$= 12.7416$$

Confidence Coefficient ("Multiplier) –
For a 95% confidence interval, 95% coverage means 5% NON-coverage. Splitting the non-coverage evenly in the two tails tells us that we want the 2.5th and 97.5th percentiles of the Normal(0,1):

www.artofstat.com > Online Web Apps > Normal Distribution > tab: Find Percentile/Quantile



https://istats.shinyapps.io/NormalDist/

And, thus, the required multiplier = 1.96 Note: 1.96 is the value of the 97.5th percentile.

Putting it all together –

The required confidence interval is of the following form

$$\text{lower limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] - (\text{Z-score}_{.975})\, SE\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

$$\text{upper limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] + (\text{Z-score}_{.975})\, SE\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

Solution:
Lower limit = 24 - (1.96) [12.7416] = **-0.9735**
Upper limit = 24 + (1.96) [12.7416] = **+ 48.9735**

Interpretation –

With 95% confidence, we estimate that the difference in means, $[\mu_{\text{Zidovudine}} - \mu_{\text{Control}}]$ is between -0.97 and +48.97.
Note – This is a two sided confidence interval. Also note that it includes the null hypothesis value of 0!.

Example Scenario ii. Variances UNKNOWN but Assumed Equal ($\sigma_1^2 = \sigma_2^2$):
(*Note: These data are hypothetical.*)

Same example (**Sorry! Are you bored yet?**). Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those not receiving zidovudine. We are assuming that the scores are normally distributed with distributions that have the same variance σ^2, unknown. Our data summaries are the following:

Zidovudine	Control
$n_1 = 15$	$n_2 = 22$
$\bar{X}_1 = 120$	$\bar{X}_2 = 96$
$S_1 = 40$	$S_2 = 35$

Point Estimate: We want a point estimate of the difference [$\mu_{Group\ 1}$ - $\mu_{Group\ 2}$]

- Our point estimator is the same as before: $[\bar{X}_{Group\ 1} - \bar{X}_{Group\ 2}] = 24$

Standard Error of the Point Estimate:

As before (see again, page 12)

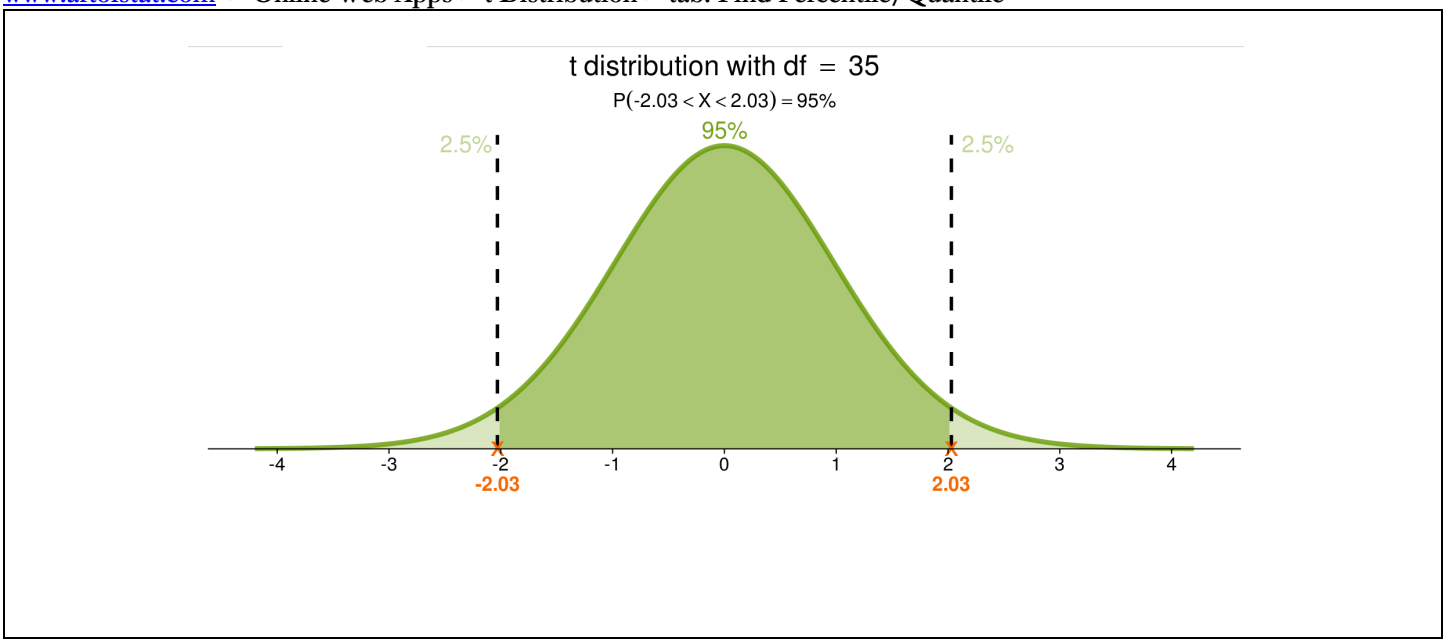
$$\hat{SE}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S_{pool}^2}{n_1} + \frac{S_{pool}^2}{n_2}} = 12.42$$

Degrees of freedom = **35**.

Confidence Coefficient ("Multiplier) –

The correct Student-t distribution for this confidence interval is the same as for the hypothesis test described on pp 10-12. For a 95% confidence interval, 95% coverage means 5% NON-coverage. Splitting the non-coverage evenly in the two tails tells us that we want the 2.5th and 97.5th percentiles of the Student T-distribution:

www.artofstat.com > Online Web Apps > t Distribution > tab: Find Percentile/Quantile



https://istats.shinyapps.io/tdist/

Thus, the required multiplier = 2.03 Note – 2.03 is the value of the 97.5th percentile.

Putting it all together –

The required confidence interval is of the following form

$$\text{lower limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] - (\text{Student } t_{DF=35;.975})\, \hat{SE}\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

$$\text{upper limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] + (\text{Student } t_{DF=35;.975})\, \hat{SE}\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

Solution:
Lower limit = 24 - (2.03) [12.42] = **-1.21**
Upper limit = 24 + (2.03) [12.42] = **+ 49.21**

Interpretation –

With 95% confidence, we estimate that the difference in means, $[\mu_{\text{Zidovudine}} - \mu_{\text{Control}}]$ is between -1.21 and +49.21. Note – This is a two sided confidence interval.

Example Scenario iii. Variances UNKNOWN and Unequal ($\sigma_1^2 \neq \sigma_2^2$):

Point Estimate:
Same.

$$[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = [120 - 96] = 24$$

Standard Error of the Point Estimate:

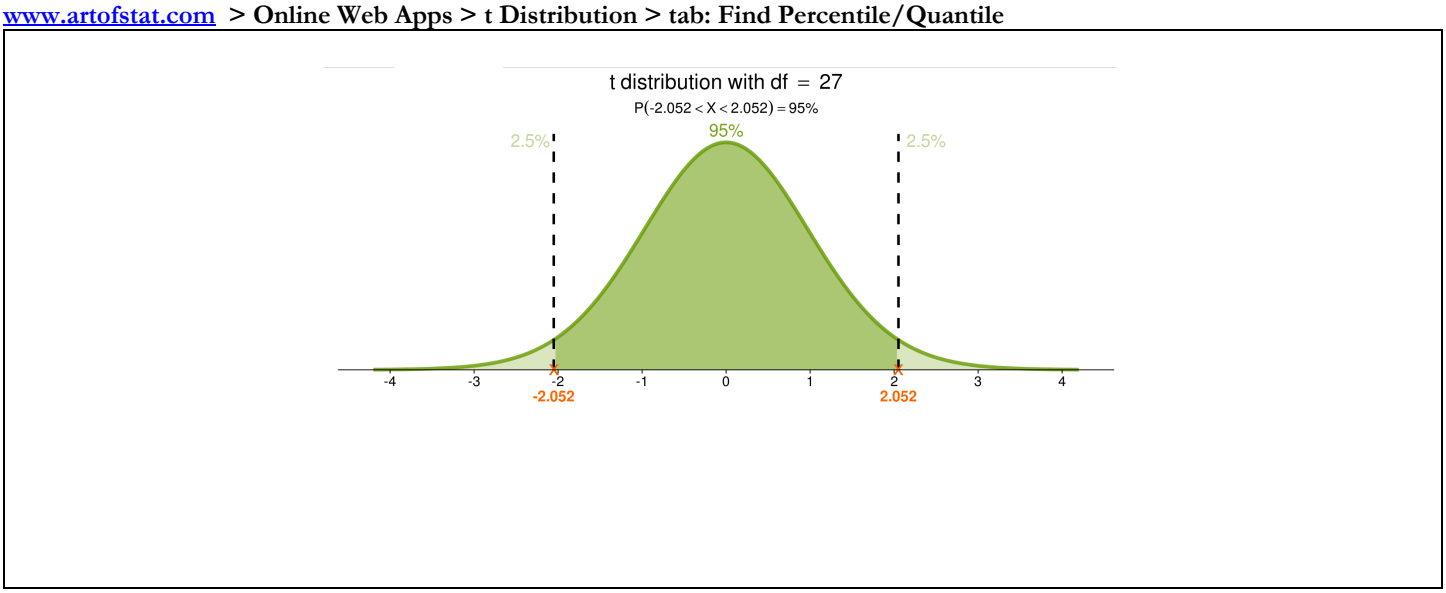
See again page 13. We have

$$S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 12.7416 .$$

Degrees of freedom calculation using Satterthwaite formula shown on page 15 yields df = **27**

Confidence Coefficient ("Multiplier) –
Thus, here, we use the Student-t distribution with degrees of freedom = 27.

www.artofstat.com **> Online Web Apps > t Distribution > tab: Find Percentile/Quantile**



https://istats.shinyapps.io/tdist/

And, thus, the required multiplier = **2.052**

Putting it all together –

The required confidence interval is of the following form

$$\text{lower limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] - (\text{Student } t_{DF=27;.975})\, \hat{SE}\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

$$\text{upper limit} = \left[\bar{X}_{Group1} - \bar{X}_{Group2}\right] + (\text{Student } t_{DF=27;.975})\, \hat{SE}\left[\bar{X}_{Group1} - \bar{X}_{Group2}\right]$$

Solution:

Lower limit = 24 - (2.052) [12.74] = **-2.14**

Upper limit = 24 + (2.052) [12.74] = **+ 50.14**

Interpretation –

With 95% confidence, we estimate that the difference in means, $[\mu_{\text{Zidovudine}} - \mu_{\text{Control}}]$ is between -2.14 and +50.14.

Note - It makes sense that this interval is wider (less precise) than that obtained previously. We had to estimate two unknown variances instead of just one.

HOMEWORK DUE Monday December 5, 2022
Question #2 of 3

A possible environmental determinant of lung function in children is the amount of cigarette smoking in the home. To study this question, two groups of children were studied. Group 1 consisted of 23 nonsmoking children aged 5-9 both of whose parents smoke in the home. Group 2 consisted of 20 nonsmoking children aged 5-9 neither of whose parents smoke. The sample mean (sample SD) of FEB1 for group 1 is 2.1 L (0.7) and for the Group 2 children, the sample mean (sample SD) of FEV1 is 2.3 L (0.4).

Under the assumption of normality, compute a 95% confidence interval for the true mean difference in FEV1 between 5-9 year old children whose parents smoke and comparable children whose parents do not smoke.

In developing your answer assume that the unknown population variances are NOT equal.

2d. Test of Equality of Two Variances – Null: $(\sigma_1^2/\sigma_2^2) = 1$

Tip -

A test of equality of variances utilizes p-value calculations using the F-distribution.

Before you begin - See again course notes *8. SUPPLEMENT: Normal, t, Chi Square, F and Sums of Normals* pp 17-20 for a review of the F distribution.

In this setting, our focus is on σ_1^2/σ_2^2 , the ratio of two independent variances.

- If the two variances are equal, then their ratio will be 1.

Examples

- Are the reproducibilities of two laboratory assays similar?
  Tip! "reproducibility" is all about noise and variability

- We might want to assess the similarity of two independent normal population distributions. This would include a comparison of the two variance parameters.

- We might want to assess the similarity of two variance parameters before doing an analysis that requires assuming them to be equal.

- Our null hypothesis is $H_O: \sigma_1^2 = \sigma_2^2$

- Operationally, we will work with the equivalent null hypothesis $H_O: [\sigma_1^2/\sigma_2^2] = 1$

Example

Health services researchers are interested in patterns of length of stay (LOS) among patients entering the hospital through the emergency room as compared to those among elective hospitalizations.

Here are the summaries we need:

Group 1: Elective	Group 2: Emergency
$n_1 = 14$	$n_2 = 11$
$S_1 = 10.9$ days	$S_2 = 4.2$ days

Research Question.

Is the variability in LOS *different* depending on the patient type, emergency versus elective?

Assumptions.

Two independent samples, each a simple random sample from a Normal distribution,
$X_1 \ldots X_{n1}$ distributed Normal (μ_1, σ_1^2) and $Y_1 \ldots Y_{n2}$ distributed Normal (μ_2, σ_2^2)

H_O and H_A.

$$H_O: \sigma_1^2 = \sigma_2^2 \text{ (equivalently: } \sigma_1^2/\sigma_2^2 = 1)$$

$$H_A: \sigma_1^2 \neq \sigma_2^2 \text{ (equivalently: } \sigma_1^2/\sigma_2^2 \neq 1) \text{ two sided.}$$

Proof by contradiction reasoning and the definition of the p-value calculation.

When the null hypothesis assumption is true, our F statistic is distributed F; that is

$$F = \left[\frac{S_1^2}{S_2^2}\right] \quad \textbf{is distributed F with numerator df = } (n_1\textbf{-1}) \textbf{ and denominator df = } (n_2\textbf{-1})$$

"Evaluation" rule.

Our p-value calculation answers the following question:
Under the null hypothesis (H_O) model, what are the chances of obtaining an F-statistic value as extreme or more extreme (in the direction of the alternative) than the value we obtained from our sample?

$$\text{p-value} = (2)\ \Pr\left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2}\right) \middle| H_O \text{ is true}\right].$$

Calculations.

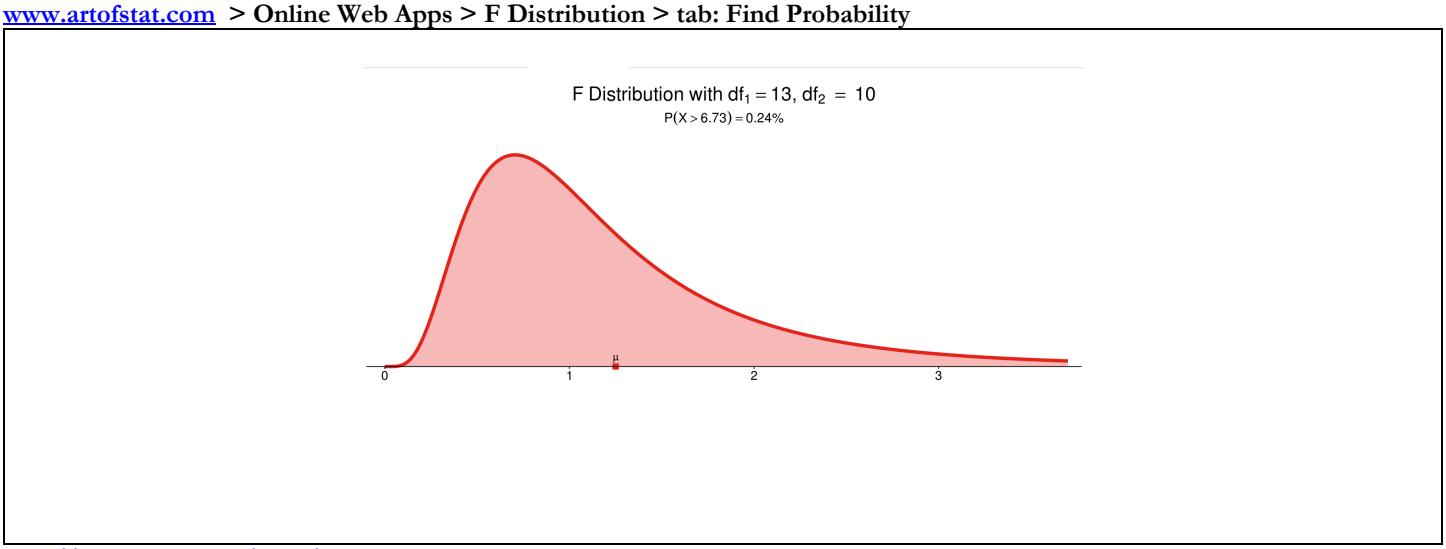
p-value

$$= (2)\Pr\left[F_{df=13,10} \geq \left(\frac{S_1^2}{S_2^2} \right) \middle| \, H_O \text{ is true} \right] = (2)\Pr\left[F_{df=13,10} \geq \left(\frac{10.9^2}{4.2^2} \right) \right] = (2)\Pr\left[F_{df=13,10} \geq 6.73 \right]$$

= (2) (0.0024)

= 0.0048

www.artofstat.com > Online Web Apps > F Distribution > tab: Find Probability



https://istats.shinyapps.io/FDist/

"Evaluate".

Application of the null hypothesis (H_O) model to the data has led to a very unlikely event, one that occurs, roughly, only 5 times in 1000. Reject the null hypothesis.

Interpret.

These data provide statistically significant evidence that the variability in length of stay is not the same in the two groups, elective patients versus for emergency patients.

2e. Confidence Interval for the Ratio of Two Variances (σ_1^2/σ_2^2)

Tip -

A confidence interval estimate of the ratio of two variances utilizes percentiles of the F-distribution.

As before - See again course notes *8. SUPPLEMENT: Normal, t, Chi Square, F and Sums of Normals* pp 17-20 for a review of the F distribution.

(1-α)100% Confidence Interval for σ_1^2/σ_2^2 Setting – Two Independent Normal Distributions	
Lower limit =	$\left(\frac{1}{F_{n_1-1;n_2-1;(1-\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right]$
Upper limit =	$\left(\frac{1}{F_{n_1-1;n_2-1;(\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right]$

Example

(Source: Daniel WW. Biostatistics: A Foundation for Analysis in the Health Sciences, Fourth Edition. 1987. Page 163)

Reaction time to a stimulus was examined in two independent groups, each a simple random sample from a Normal population distribution. One group ($X_1 \ldots X_{n1}$) is comprised of n_1=21 healthy adults. The other group ($Y_1 \ldots Y_{n2}$) includes n_2 = 16 Parkinson's disease patients. Calculate a 95% confidence interval estimate of σ_1^2 / σ_2^2.

Preliminary calculations yield the following:

$$\hat{\sigma}_1^2 = S_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1} = 1600 \quad \textbf{and} \quad \hat{\sigma}_2^2 = S_2^2 = \frac{\sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_2 - 1} = 1225$$

Numerator degrees of freedom = n_1-1 = 20
Denominator degrees of freedom = n_2-1 = 15

Point Estimate is S_1^2 / S_2^2

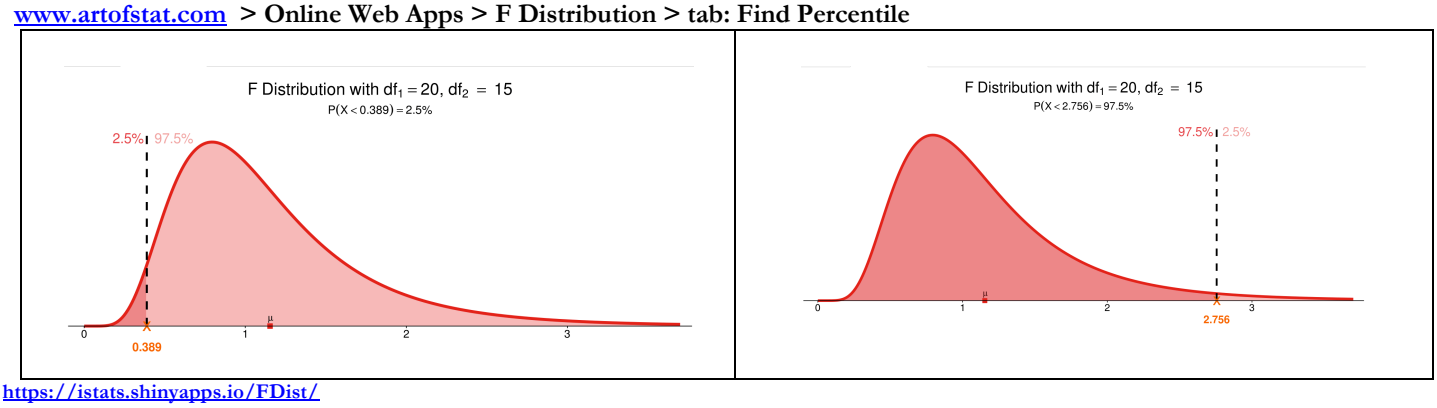
$$S_1^2 / S_2^2 = 1600/1225 = 1.306$$

Confidence Coefficient Multipliers

$$\left(\frac{1}{F_{n_1-1;n_2-1;(1-\alpha/2)}}\right) = \left(\frac{1}{F_{20;15;.975}}\right) = \left(\frac{1}{2.76}\right)$$

$$\left(\frac{1}{F_{n_1-1;n_2-1;(\alpha/2)}}\right) = \left(\frac{1}{F_{20;15;.025}}\right) = \left(\frac{1}{0.3886}\right)$$

www.artofstat.com > Online Web Apps > F Distribution > tab: Find Percentile



https://istats.shinyapps.io/FDist/

Solution for Lower and Upper Confidence Interval Limit Values

Lower Limit of confidence interval =

$$\left(\frac{1}{F_{n_1-1;n_2-1;(1-\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right]=\left(\frac{1}{2.76}\right)[1.306]=0.47$$

Upper Limit of confidence interval =

$$\left(\frac{1}{F_{n_1-1;n_2-1;(\alpha/2)}}\right)\left[\frac{S_1^2}{S_2^2}\right]=\left(\frac{1}{0.3886}\right)[1.306]=3.36$$

Interpretation –

With 95% confidence, we estimate that the ratio of the two variances is between 0.47 and 3.36. As this interval includes the null hypothesis value of 1, we will make the assumption of equal variances later when we want to do a Student T-test of equality of independent means.

HOMEWORK DUE Monday December 5, 2022
Question #3 of 3

The setting is the same as that of question #2.

A possible environmental determinant of lung function in children is the amount of cigarette smoking in the home. To study this question, two groups of children were studied. Group 1 consisted of 23 nonsmoking children aged 5-9 both of whose parents smoke in the home. Group 2 consisted of 20 nonsmoking children aged 5-9 neither of whose parents smoke. The sample mean (sample SD) of FEB1 for group 1 is 2.1 L (0.7) and for the Group 2 children, the sample mean (sample SD) of FEV1 is 2.3 L (0.4).

Under the assumption of normality, construct a 95% confidence interval for the ratio of the variances of the two groups. What is your conclusion regarding the reasonableness of the assumption of equality of population variances?

3. Two Independent Groups – 0/1 Discrete Outcome: Binomial

In this setting, our focus is on [π_1 - π_2], the difference between the two binomial event probabilities.

- **Tip/Hack!** If the two event probabilities are equal, then their difference will be zero.

Examples -

- Are the probabilities of **disease occurrence** the same in two independent populations (e.g. – Iceland versus Hawaii)?
- In a randomized controlled trial, patients are randomized to either intervention (drug) or control (standard care) Is the probability of the **event of "improvement"** the same in both groups?

Our data are the outcomes x_1 and x_2 of two independent Binomial random variables:

- X_1 distributed Binomial(n_1, π_1)
- X_2 distributed Binomial(n_2, π_2)

Our null hypothesis is H_O: $\pi_{Intervention} = \pi_{Control}$

Operationally, we will work with the equivalent null hypothesis H_O: $[\pi_{Intervention} - \pi_{Control}] = 0$

3a. Hypothesis Test for $[\pi_1 - \pi_2] = 0$

Tips -
There is more than one test and they are equivalent!

Approach #1. One test of equality of binomial event probabilities yields a Z-score and utilizes p-value calculations using the standard normal distribution.

Approach #2. Another test of equality of binomial event probabilities utilizes p-value calculations using the chi square distribution. This will be introduced in Unit 11, Chi Square Tests.

Approach #3. A third test of equality of binomial event probabilities is the Fisher Exact Test and utilizes p-value calculations using the Central Hypergeometric distribution. This will be introduced in Biostats 640 Unit 3, Discrete Distributions.

Here, approach #1 (normal theory Z-score approximate test) is described.

Example

Consider a needle exchange trial that compares access to clean injection paraphernalia in two groups. The control group has access through pharmacy sales. The intervention group has access to pharmacy sales and a needle exchange. Investigators wanted to know if the additional provision of a needle exchange program would be effective in reducing risky injection behavior and, ideally, HIV sero-conversion.

Among the preliminary aims was an analysis to identify variables that are associated with both randomization assignment and outcome. Such variables are potential confounders of response to intervention. The literature suggests that women might respond differently to intervention than men. Therefore, an interim analysis sought to determine if there are gender differences in randomization assignment.

Among n=101 randomized participants:

Pharmacy Sales	Pharmacy Sales + Needle Exchange
$n_1 = 53$	$n_2 = 48$
# women = 9 = X_1	# women = 13 = X_2
% women = 17.0 = $\bar{X}_1$	% women = 27.1 = $\bar{X}_2$

Research Question.

Is the proportion of women in the pharmacy sales + needle exchange condition (27.1%) significantly different than the proportion of women in the pharmacy sales condition (17.0%), considering the limitations of sample size (53 and 48, respectively)?

Null Hypothesis Assumptions.

- The event of "female gender" is the outcome of interest. In each group (pharmacy sales versus pharmacy sales + needle exchange), the number (X) who are gender = female is a Binomial random variable.
- We will represent the proportions of women in the two groups as $\bar{X}_1$ and $\bar{X}_2$. These are related to: X_1 is distributed Binomial (n_1=53, π_1) and X_2 is distributed Binomial (n_2=48, π_2) where

π_1 = Proportion "female gender" in group = Pharmacy Sales
π_2 = Proportion "female gender" in group = (Pharmacy Sales + Needle Exchange)

H_O and H_A.

$H_O : \pi_1 = \pi_2$ The proportions of women in both groups are equal
$H_A : \pi_1 \neq \pi_2$ two sided The proportions of women in both groups are NOT equal

The Test statistic is a Z-score *(take care! – it is NOT a Student t)*.

$$z_{score} = \left[\frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2)|H_O true]}{\hat{SE}[(\bar{X}_1 - \bar{X}_2)|H_O true]} \right]$$

Solution for the Correct Estimate of the Standard Error.

Two Independent Binomials – Calculation of $\hat{SE}\left[(\bar{X}_1 - \bar{X}_2)\,|\,H_O true\right]$

$$\hat{SE}\left[(\bar{X}_1 - \bar{X}_2)\,|\,H_O\right] = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}$$ **where**

$\hat{\pi}$ is our best guess of the common π and

$$\hat{\pi} = \left[\frac{X_1 + X_2}{n_1 + n_2} \right].$$ ***Notice that this is just the overall proportion***

For these data:

$$\hat{\pi} = \left[\frac{X_1 + X_2}{n_1 + n_2}\right] = \left[\frac{9+13}{53+48}\right] = .218$$

$$S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\frac{.218(1-.218)}{53} + \frac{.218(1-.218)}{48}} = .0823$$

Proof by contradiction reasoning and the definition of the p-value calculation.

In the needle exchange trial, we want to know the null hypothesis model chances of obtaining a difference in the proportion of women in the two groups as great or greater than |.271 - .170| = .1010

The required p-value calculation is thus

p-value = $2\Pr\left[\left|\left(\bar{X}_2 - \bar{X}_1\right)\right| \geq \left|(.271 - .170)\right|\right]$.

P-Value calculation.

$$\textbf{p-value} = 2\Pr\left[(\bar{X}_2 - \bar{X}_1) \geq (.271 - .170) \mid H_O \text{ is true}\right]$$

$$= 2\Pr\left[\frac{(\bar{X}_2 - \bar{X}_1) - E(\bar{X}_2 - \bar{X}_1)}{S\hat{E}(\bar{X}_2 - \bar{X}_1)} \geq \frac{(.271 - .170) - (0)}{.0823}\right]$$

$$= 2\Pr[z - score \geq 1.23] = 2[.10935]$$

$$= .22$$

Note. z_{score}=1.23 says "the observed difference in % women in the two randomization groups equal to (.271-.170) = .1010 is 1.23 standard error units greater than the expected difference of 0 when the null hypothesis is true."

"Evaluate".
With sample sizes of 53 and 48, there was a reasonable chance, 22% chance, of obtaining a discrepancy in the % women in the two groups equal to 10 percentage points or more.

Interpret.
Application of the null hypothesis has NOT led to an unlikely outcome. Retain the null hypothesis and conclude that there is not a statistically significant difference in the proportion of women in the two study conditions among the 101 available for interim analysis.

3b. Confidence Interval for $[\pi_1 - \pi_2]$

Confidence Interval for a difference between two independent proportions $[\pi_1 - \pi_2]$
Two Independent Binomial Distributions

$$[\hat{\pi}_1 - \hat{\pi}_2] \pm (z_{1-\alpha/2}) \hat{SE}(\hat{\pi}_1 - \hat{\pi}_2)$$

where the required calculations are

(1) $$\bar{X}_1 = \frac{X_1}{n_1} \text{ and } \bar{X}_2 = \frac{X_2}{n_2}$$

(2) $$\hat{\pi}_1 = \bar{X}_1 \text{ and } \hat{\pi}_2 = \bar{X}_2$$

(3) $$\hat{SE}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{X}_2(1-\bar{X}_2)}{n_2}}$$

(4) **For small number of trials ($n \leq 30$ or so) in either group, use**

$$\hat{SE} = \sqrt{\frac{0.5(0.5)}{n_1} + \frac{0.5(0.5)}{n_2}}$$

Example

For the needle exchange trial introduced on page 32, calculate a 95% confidence interval estimate for the difference in the proportions of women, intervention versus control.

Point Estimate of [$\pi_1 - \pi_2$] is difference between the sample means

$$\hat{\pi}_1 = \bar{X}_1 = X_1/n_1 = 9/53 = 0.17$$
$$\hat{\pi}_2 = \bar{X}_2 = X_2/n_2 = 13/48 = 0.271$$

$$[\hat{\pi}_1 - \hat{\pi}_2] = [\bar{X}_1 - \bar{X}_2] = [0.17 - 0.271] = -0.101$$

Estimated standard error of $[\hat{\pi}_1 - \hat{\pi}_2]$ is now obtained using

$$S\hat{E}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} = \sqrt{\frac{.17(1-.17)}{53} + \frac{.271(1-.271)}{48}} = .0823$$

Confidence Coefficient Multiplier is a percentile from the Normal(0,1) Distribution

$$z_{.975} = \mathbf{1.96}$$

Putting it all together.

Lower = (point estimate) - (multiple) (SE of estimate) = -0.101 – (1.96)(0.0823) **= - 0.26**
Upper = (point estimate) + (multiple) (SE of estimate) = -0.101 + (1.96)(0.0823) **= + 0.06**

Interpretation –

With 95% confidence, we estimate that the difference in the proportion of women in the two groups, control – intervention, is between -0.26 and + 0.06. As this interval includes the null hypothesis value of 0, we will assume that randomization has not yielded a worrisome imbalance with respect to gender (albeit the sample sizes are small).