ST 517 Note Outline 7: Inference for Two Groups

Lecture 7.1: Introduction and Foundations

Previously: Statistical inference a single population/group

- Confidence intervals to estimate value of parameter
- Hypothesis tests to look for evidence that supports a specific hypothesis about parameter

Now: Statistical inference to compare two populations/groups

- Same two tools (intervals and tests) as before
  - Same basic construction of confidence intervals and same basic steps to hypothesis testing as before
  - New details for parameters, estimates, standard errors, and conditions
- Use both tools together!

Notation:

- Same basic notation as before, but now use subscripts to denote different groups:

n_1 = sample 1 size	n_2 = sample 2 size
p_1 = population 1 proprotion	p_2 = population 2 proportion
$\hat{p}_1$ = sample 1 proportion	$\hat{p}_2$ = sample 2 proportion
μ_1 = population 1 mean	μ_2 = population 2 mean
$\bar{y}_1$ = sample 1 mean	$\bar{y}_2$ = sample 2 mean
σ_1 = population 1 standard deviation	σ_2 = population 2 standard deviation
s_1 = sample 1 standard deviation	s_2 = sample 2 standard deviation

Recall:

- For any random variables:
$$E(a_1Y_1 + a_2Y_2 + \cdots + a_nY_n) = a_1E(Y_1) + a_2E(Y_2) + \cdots + a_nE(Y_n)$$
- For independent random variables:
$$V(a_1Y_1 + a_2Y_2 + \cdots + a_nY_n) = a_1^2V(Y_1) + a_2^2V(Y_2) + \cdots + a_n^2V(Y_n)$$
- If $Y_1, \ldots, Y_n$ are normally distributed rvs (possibly with different means and/or variances), then any linear combination of the Y_i's also has a normal distribution
- Sampling distribution of the sample mean and sampling distribution of the sample proportion are normally distributed when certain conditions hold

Comparing Two Groups

- Easy way to compare two values is to take their difference
- Compare means for two populations:
    - Parameter:
    - Statistic:
    - Sampling distribution of this statistic:
        - Shape:
        - Mean:
        - Standard deviation:
        - Conditions:
- Compare proportions for two populations:
    - Parameter:
    - Statistic:
    - Sampling distribution of this statistic:
        - Shape:
        - Mean:
        - Standard deviation:
        - Conditions:

Example: Show $E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2$ and $\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Example: Show $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$ and $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

Lecture 7.2: Inference for Means of Two Independent Groups

- When to use:

- In general, parameter is:

Confidence Interval for Difference in Two Population Means

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Multiplier found using t-distribution with degrees of freedom $\nu = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$

- Conditions:

Hypothesis Test for Difference in Two Population Means

1. Identify research question, population, and parameter of interest

2. Establish null and alternative hypotheses
    - Null hypothesis:
        - Notation: μ_0 – null value

    - Choice of alternatives:

3. Identify type of hypothesis test and check conditions

    - Type of test:

    - Conditions:

4. Calculate test statistic

    - If conditions are met, appropriate test statistic is:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

5. Identify null distribution and calculate p-value

    - Null dist: Under above conditions, test stat follows t-distribution with degrees of freedom ν

    - P-value found as probability under H_0, but in the direction of H_A

6. Make decision about null hypothesis

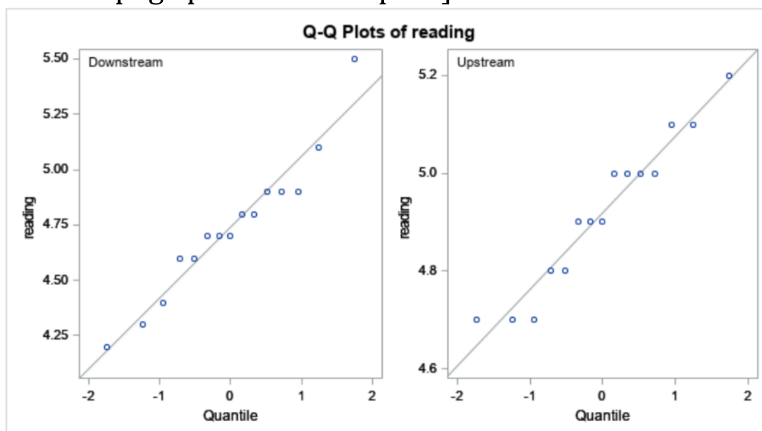
    - Decision Rule of Thumb: If p-value $\leq \alpha$, reject H_0

7. State conclusion in the context of alternative hypothesis

    - If you rejected H_0: Conclude that there is enough evidence to support the alternative hypothesis.
    - If you did <u>not</u> reject H_0: Conclude that there is <u>not</u> enough evidence to support the alternative hypothesis.
    - Describe alternative hypothesis in context!

Example: An environmental agency has concerns about a community releasing sewage into the local river, which could lower the level of dissolved oxygen in the river and cause damage to fish. They collect 30 water samples from the river, 15 from areas upstream of the release point and 15 from areas downstream of the release point, and measure the dissolved oxygen readings (in parts per million, ppm) for each sample. The resulting data is used to create the summary table and qq-plots below. Does this provide evidence that the mean dissolved oxygen readings are lower downstream from the release point? Use a 10% significance level. Consider both the results of the hypothesis test and the confidence interval to address this question. [Note: 2nd blank page provided for space]

	n	Mean	Std. Dev.
Downstream	15	4.74	0.32
Upstream	15	4.92	0.16



Lecture 7.3: Pooled Inference for Difference in Means

Pooled Inference

- Special case of independent samples t-test

- When to use:

- In general, parameter is:

Pooled Hypothesis Test

- Steps 1, 6 and 7 are same as before (these are the same for every hypothesis test!)
- Conditions:

- Test statistic:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where:

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

- Null distribution:

- P-value still found under H_0 but in direction of H_a

Pooled Confidence Interval

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\alpha/2,v}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

- Multiplier found using

- Conditions same as for pooled t-test

Example: Does the color of paper make a difference in exam scores? A history professor printed his exam on two colors of paper. He administered them to his class by randomly assigning them to his students. Thirty-one students took the exam version that was on pink paper. Forty-eight students were assigned to the version on gold paper. The resulting scores are summarized below. Does this indicate that there is a significant difference between the two colors of the exam? Consider both the results of the hypothesis test and the confidence interval to address this question. [Note: 2nd blank page provided for space]

Color	n	Mean	StdDev
Pink	31	72	8.1
Gold	48	64	9.2

Determining if Population Variances are Equal

- Can only use pooled inference when it is reasonable to assume population variances are equal: $\sigma_1^2 = \sigma_2^2$

- Ways to decide if this assumption is plausible:

   1.

   2.

   3.

      - **Sewage Ex:**

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	14	4.17	0.0115

Lecture 7.4: Inference for Means of Paired Data

Example: Potential insurance fraud?
Insurance adjusters are concerned about the high estimates they are receiving from Jocko's Garage. To see if the estimates are unreasonably high, each of 10 damaged cars was taken to Jocko's and to another garage and the amount for the repair estimates were recorded. Would it be appropriate to conduct an independent samples t-test to compare estimates between the garages?

Recall: A matched-pairs experiment

- Each unit receives two treatments. The units could be:
  1) A single subject (each subject serves as their own control)
  2) Two subjects that have been matched together (one receives the treatment and the other receives the control)

Example: For each of the following scenarios, determine if the data is paired or not.

a. *Does a full moon affect behavior?* A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as 'moon days' if they were in a 3-day period centered at the day of the full moon. For each patient, the average number of disruptive behaviors was recorded for moon days and for all other (non-moon) days.

b. *Do Americans listen to more music?* We have data on the amount of time (per week) that people with smart phones listed to music on their phone, based on random samples from the United States and the United Kingdom.

c. *Regression to the mean?* The phrase *regression to the mean* refers to the idea that if the value of random variable is extreme, a future observation of that variable is more likely to be closer to the mean than the extremes. The phenomenon was noticed in the late 19th century by a researcher comparing the heights of fathers to that of their biological sons.

Paired Inference

- Key idea: Want to see how the measurements on each unit compare to each other

- When to use:

- Inference basically the same as for a single variable

- In general, parameter is:

Confidence Interval for Mean Difference

$$\bar{d} \pm t^*_{\alpha/2, v} \frac{s_D}{\sqrt{n}}$$

- Multiplier found using

- Conditions:

Hypothesis Test for Mean Difference

- Steps 1, 6, and 7 are the same as always
- Same general options for hypotheses
  - μ_0 – null value

- Conditions same as for confidence interval
- Test statistic:

$$t = \frac{\bar{d} - \mu_0}{s_D / \sqrt{n}}$$

- Null distribution:

- P-value still found under H_0 in direction of H_a

Example: Does a full moon affect behavior? A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a 3-day period centered at the day of the full moon. For each patient, the average number of disruptive behaviors was recorded for moon days and for all other days. The following summary statistics for a sample of 15 patients are:

	Mean	**Standard Deviation**
Moon Days	3.02	1.50
Other Days	0.59	0.44
Difference	2.43	1.46

Does this data provide evidence that dementia patients are more disruptive during full moons? Use a 1% significance level. Consider both the results of the hypothesis test and the confidence interval to address this question. [Note: 2nd blank page provided for space]

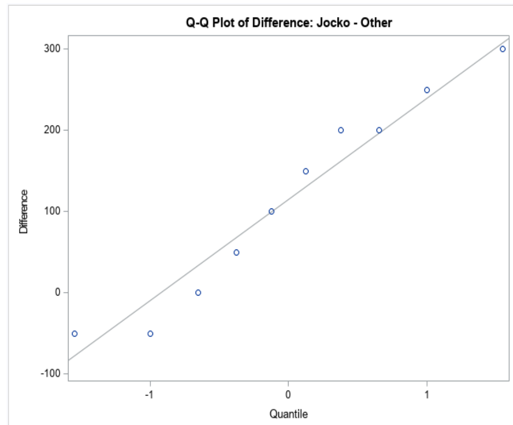
Example (Potential insurance fraud?): Insurance adjusters are concerned about the high estimates they are receiving from Jocko's Garage. To see if the estimates are unreasonably high, each of 10 damaged cars was taken to Jocko's and to another garage and the estimates recorded. SAS output of the resulting data is provided. Based on this output, is there evidence that Jocko's is overcharging for repairs? Consider both the results of the hypothesis test and the confidence interval to address this question. [Note: 2nd blank page provided for space]

Difference: Jocko - Other

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	115.0	124.8	39.4757	-50.0000	300.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
115.0	25.6997	204.3	124.8	85.8647	227.9

DF	t Value	Pr > |t|
9	2.91	0.0172



Paired vs. Independent Samples Inference

- Distinction due to data structure based on how data was collected
- Can be illustrated with data table
- Insurance Fraud Example:

Car	Jocko	Other	Difference
1	500	400	100
2	1550	1500	50
3	1250	1300	-50
4	1300	1300	0
5	750	800	-50
6	1000	800	200
7	1250	1000	250
8	1300	1100	200
9	800	650	150
10	2500	2200	300

Sample	Response
Jocko	500
Jocko	1550
Jocko	1250
Jocko	1300
Jocko	750
Jocko	1000
Jocko	1250
Jocko	1300
Jocko	800
Jocko	2500
Other	400
Other	1500
Other	1300
Other	1300
Other	800
Other	800
Other	1000
Other	1100
Other	650
Other	2200

Lecture 7.5: Inference for Two Proportions

- When to use:

- In general, parameter is:

Confidence Interval for Difference in Two Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Multiplier found using
- Conditions:

Hypothesis Test for Difference in Two Proportions

- Steps 1, 6, and 7 are the same as always
- Same general options for hypotheses using appropriate parameter
    - p_0 – null value

- Type of test:
- Conditions same as for confidence interval
- Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

- Null distribution:
- P-value still found under H_0 in direction of H_a

Example: A random sample of adults was asked if they would return the money if they found a wallet on the street. Of the 93 women, 83 said they would, and of the 75 men, 53 said they would. Assume these people represent all adults. Does this provide evidence that equal proportions of men and women say they would return the money? Consider both the results of the hypothesis test and the confidence interval to address this question. [Note: 2nd blank page provided for space]

Lecture 7.6: Additional Examples

Additional Example 1: In a randomized clinical trial, the effectiveness of two treatments—chemotherapy versus a combination of chemotherapy and radiation—was explored in a sample of breast cancer patients. Of 154 patients who had chemotherapy only, 76 survived at least 15 years. Of 164 patients who had chemotherapy and radiation, 98 survived at least 15 years.

a. Does this data provide evidence that the treatment of chemotherapy and radiation was significantly better than the treatment of chemotherapy only? Use a 5% level. Consider both the results of the hypothesis test and the confidence interval to address this question.
b. Write a sentence interpreting the confidence level in the context of this problem.

Additional Example 2: A company institutes an exercise break for its workers to see if this will improve job satisfaction, as measured by a questionnaire that was given to a random sample of workers before the break was instituted and also after. The resulting data was used to create the summary statistics below.

	n	Mean	Std.Dev
After	10	35.5	9.1
Before	10	27	8.7
Difference	10	8.5	7.5

a. Use the appropriate summaries to calculate and interpret a 90% confidence interval for the mean difference in job satisfaction for all employees at this company.
b. Conduct the hypothesis test at the 10% level.
c. Based on the test and the interval, is there evidence of an improvement in job satisfaction?

Additional Example 3: An accounting firm is deciding between IT training conducted in-hour and the use of third party consultants. To get some preliminary cost data, each type of training was implemented at two of the firm's offices (located in different cities). The table below shows the average annual training cost per employee at each location.

	n	Mean	Std.Dev
In-House	195	490	32
Consultant	195	500	48
Difference	195	-10	40

a. Use the appropriate summaries to calculate and interpret a 95% confidence interval for the true difference in mean cost.
b. Conduct the hypothesis test of a difference in average cost at the 5% level.
c. Based on the test and the interval, is there evidence that the mean costs are *different*?