Lecture 9

Hypothesis Testing: The General Concept, and For Comparing Means Between Two Populations

Section A: Hypothesis Testing for Population Comparisons: An Overview

Learning Objectives

- Upon completion of this lecture section, you will begin to understand
  - A conceptual framework for the process of statistical hypothesis tests
  - How confidence intervals and hypothesis testing are related

Motivation

- Frequently, in public health/medicine/science etc.., researchers/practitioners are interested in comparing two (or more) outcomes between two (or more) populations using data collected on samples from these populations

Motivation

- It is not only important to estimate the magnitude of the difference in the outcome of interest between the two groups being compared, but also to recognize the uncertainty in the estimate
- One approach to recognizing the uncertainty in these estimates is to create confidence intervals: a complimentary approach is called *hypothesis testing*

Hypothesis Testing

- Types of two group comparisons, continuous outcomes
  - Paired, Unpaired
- Types of two group comparisons, binary outcomes
  - Unpaired
- Types of two group comparisons, time-to-event outcomes
  - Unpaired
- These approaches can also be extended to allow for more than two population comparisons in one test for unpaired studies

Application of CLT

- It turns out that differences of two quantities whose distributions is normal, have a normal distribution
- As such, we can extend the basic principles of the CLT to understand and quantify the sampling variability
  - Mean differences between two independent populations
  - Difference in proportions between two independent populations
  - The natural logs of ratios (which are actually differences)

7

Confidence Interval Logic

- The creation of a confidence interval for a difference (means, risks, ln(ratios) uses the results from the CLT coupled with the properties of the normal curve to create an interval the (likely) includes the unknown truth for the population comparison measure
- Confidence interval idea: letting "data take you to the truth"

8

Another Approach to Handling Uncertainty

- Another possible approach for linking the sample results to the unknown truth is the start with some competing, exhaustive possibilities for the unknown truth about the population comparison measure, and use data from samples to choose between these possibilities
- One possibility for the competing truths:

  Truth 1: NO DIFFERENCE between populations
  Truth 2: A DIFFERENCE between the two populations

9

Application/Extension of CLT: Ratios

- These two possibilities can be phrased in terms of the null values we discussed in lecture 8
- For example, when comparing means between two populations, if
  - The means are the same ($\mu_1 = \mu_2$) then $\mu_1 - \mu_2 = 0$
  - The means are not the same ($\mu_1 \neq \mu_2$) then $\mu_1 - \mu_2 \neq 0$

10

Another Approach to Handling Uncertainty

- These two competing truths are called the *null hypothesis* (H_o) and the *alternative hypothesis* (H_A

  H_o: NO DIFFERENCE between populations
  H_A: A DIFFERENCE between the two populations

- These can be expressed in several, equivalent ways for the types of data outcomes we have considered (continuous, binary, time-to-event)

11

Another Approach to Handling Uncertainty

- Continuous

12

Another Approach to Handling Uncertainty

- Binary
- Time to Event

13

Another Approach to Handling Uncertainty

- How can the study data be used to choose between one of these two truths, while accounting for the uncertainty in the study data
- The theoretical sampling distribution will again be utilized in this process

14

Utilizing the Sampling Distribution: 95% CIs

- Confidence intervals
  - IDEA: for most studies , sample estimated difference will be "close" to unknown truth

15

Utilizing the Sampling Distribution: Hypothesis Testing

- Hypothesis Testing: Starts by assuming H_o is truth
  IDEA: if H_o is truth, then the sample estimated difference will be "close" to H_o

16

Another Approach to Handling Uncertainty

- Hypothesis Testing Approach: getting a p-value

17

Another Approach to Handling Uncertainty

- What does a p-value measure?
  - The p-value is the probability of getting a study result as extreme or more extreme (as far or farther from the null value) by chance alone, if the null hypothesis is the underlying population truth

18

Another Approach to Handling Uncertainty

- How is the p-value used to make a decision about the two competing hypotheses?
  - The p-value gives a probability: there needs to be a rule about whether the p-value means the study results are "likely" or "unlikely" if the null is true
  - General cutoff: 0.05
    This is called the "rejection level" or "alpha (α) level"
    It does not have to be 0.05, but as we'll see 0.05 corresponds to a 95% confidence interval

19

Another Approach to Handling Uncertainty

- How is the p-value used to make a decision about the two competing hypotheses?
  - If $p<0.05$: the decision is made to "reject the null hypothesis" in favor of the alternative: the result is called "statistically significant" (at the 0.05 level)
  - If $p\geq 0.05$, the decision is "fail to reject the null hypothesis" : not a very strong conclusion, but we will see why this language is used shortly

20

Another Approach to Handling Uncertainty

- What is the relationship between the 95% CI, the appropriate null value and the p-value?

  If $p<0.05$, then the 95% CI for the measure of interest (mean difference, difference in proportions, RR etc..) will not include the null value

  If $p\geq 0.05$, then the 95% CI for the measure of interest will include the null value

21

Summary

- Confidence intervals and hypothesis testing are two complimentary ways of addressing uncertainty in sample based comparisons to making statements about the unknown population comparisons
- Both methods operate on the principle that for most random sample based studies, the sample results show be "close" to the truth

22

Summary

- The confidence interval approach starts with the study results, and creates an interval around the study results to create a range of possibilities for the unknown truth
- The hypothesis testing approach starts with an assumption about the unknown truth, and then measures how far the study results are from this assumed truth

23

Summary

- The end result of hypothesis testing is a p-value. The p-value quantifies how likely the study results are (or results even less likely) if the samples being compared came from populations with equal parameters of interest (means, proportions, incidence rates …)

24

Summary

- In general, the mechanics of the test (and the name of the test employed) depends on the type of data being compared
- However, the conceptual foundation of all hypothesis tests is the same

25

Section B: (Hypothesis Testing) Comparing Means Between Two Populations: The Paired Approach

26

Learning Objectives

- In this lecture section you will learn how to estimate and interpret a p-value for a hypothesis test of a mean difference between two populations for the paired study design
- The method for getting the p-value is called the paired t-test
  - "paired" because of the study design
  - "t-test" because (sometimes) the sampling distribution is a t-distribution

27

Example 1: Paired Comparison

- Two different physicians assessed the number of palpable lymph nodes in 65 randomly selected male sexual contacts of men with AIDS or AID-related condition[1]

	Doctor 1	Doctor 2	Difference
Mean ($\bar{x}$)	7.91	5.16	-2.75
sd (s)	4.35	3.93	2.83

[1] example based on data taken from Rosner B Fundamentals of Biostatistics, 6th ed. (2005) Duxbury Press. (based on research by Coates, et al. (1988) Assessment of generalized….*Journal of Clinical Epidemiology*, 41(2).

28

95% Confidence Interval

- 95% CI for difference in mean number of lymph nodes, Doctor 2 compared to Doctor 1: -3.45 to -2.05
- Had all such men been examined by these two physicians, the average difference in number of lymph nodes discovered by the two physicians would be between -3.45 and -2.05
- Notice, all possibilities for the true mean difference are negative, and 0 is not included in the interval

29

Hypothesis Testing Approach

- Set up the two competing hypotheses

$$H_o: \mu_{doctor2} = \mu_{doctor1} \qquad H_o: \mu_{doctor2} - \mu_{doctor1} = 0$$
$$H_A: \mu_{doctor2} \neq \mu_{doctor1} \qquad H_A: \mu_{doctor2} - \mu_{doctor1} \neq 0$$

- Assume H_o is true: Figure out "how far" observed mean difference ($\bar{x}_{doctor2} - \bar{x}_{doctor1}$) is from expected difference under H_o (0) in terms of standard errors. The distance measure is:

30

Hypothesis Testing Approach

- Translate distance into p-value by comparing it to the distribution of such differences because of sampling variability when the true, population level difference is 0
- Compare the p-value to the preset rejection level (alpha level): for our purposes, and most of the research world, this is 0.05
- We have a result that is 7.83 standard errors below the expected mean difference of 0 (under the null hypotheses): How likely is this to occur just by chance (because of random sampling error)?

31

Hypothesis Testing Approach

- Getting the p-value
- The resulting p-value is very small (<0.0001): Interpretation?

32

Hypothesis Testing Approach

- Making a decision
- How does this decision compared to the decision we would make from the 95% CI for the difference in population means?

33

Hypothesis Testing Approach

- Note: p-value invariant to direction of comparison. Suppose we had instead presented the estimate, and the hypotheses in terms of doctor 1 minus doctor 2?

$$H_o: \mu_{doctor1} = \mu_{doctor2} \qquad H_o: \mu_{doctor1} - \mu_{doctor2} = 0$$
$$H_A: \mu_{doctor1} \neq \mu_{doctor2} \qquad H_A: \mu_{doctor1} - \mu_{doctor2} \neq 0$$

- The distance measure:

$$t = \frac{(\bar{x}_{doctor1} - \bar{x}_{doctor2}) - 0}{SE(\bar{x}_{doctor1} - \bar{x}_{doctor2})} = \frac{(\bar{x}_{doctor1} - \bar{x}_{doctor2})}{\left(\frac{s_{diff}}{\sqrt{n}}\right)} = 7.83$$

34

Hypothesis Testing Approach

- Getting the p-value

35

Example 2: Paired Comparison

- Cereal and cholesterol: 14 males with high cholesterol given oat bran cereal as part of diet for two weeks, and corn flakes cereal as part of diet for two weeks[2]

	Corn Flakes	Oat Bran	Difference
Mean ($\bar{x}$)	171.2 mg/dL	157.8	13.4
sd (s)	38.7	42.5	15.5

[2] example based on data taken from Pagano M. Prinicples of Biostatistics, 2nd ed. (2000) Duxbury Press. (based on research by Anderson J, et al. (1990) Oat Bran Cereal Lowers……*American Journal of Clinical Nutrition*, 52.

36

Example 2: Paired Comparison

- The resulting 95% mean difference and 95% confidence interval in cholesterol levels for the corn flake group compared to the oat bran group was :

  13.4 mg/dL (4.5 mg/dL, 22.3 mg/dL)

37

Example 2: Hypothesis Testing Approach

- Set up the two competing hypotheses

$$H_o: \mu_{CF} = \mu_{OB} \qquad H_o: \mu_{CF} - \mu_{OB} = 0$$
$$H_A: \mu_{CF} \neq \mu_{OB} \qquad H_A: \mu_{CF} - \mu_{OB} \neq 0$$

- Assume H_o is true: Figure out "how far" observed mean difference ($\bar{x}_{CF} - \bar{x}_{OB}$) is from expected difference under H_o (0) in terms of standard errors. The distance measure is:

38

Hypothesis Testing Approach

- Getting the p-value

39

Example 3: Paired Design

- Before versus After Study :Data

	BP Before OC	BP After OC	After-Before
1.	115	128	13
2.	112	115	3
3.	107	106	-1
4.	119	128	9
5.	115	122	7
6.	138	145	7
7.	126	132	6
8.	105	109	4
9.	104	102	-2
10.	115	117	2

$$\bar{x}_{diff} = 4.8 \text{ mmHg}$$
$$S_{diff} = 4.6 \text{ mmHg}$$

40

Example 3: Paired Design

- The resulting 95% mean difference and 95% confidence interval for the difference in blood pressure after oral contraceptive use compared to before oral contraceptive use was :

  4.5 mmHg (1.5 mmHg, 8.1 mmHg)

- The resulting p-value from the paired t-test is 0.016
  - If the there was no difference in population mean SBPs after and before OC use the chances of getting a sample of 10 women from the population with a sample mean difference of 4.5 (or a difference more extreme) is 16 in 1,000

41

Summary

- The paired t-test is a method for getting a p-value for testing the computing hypotheses

$$H_o: \mu_1 = \mu_2$$
$$H_A: \mu_1 \neq \mu_2$$

  using data from paired samples from the paired populations

- The resulting decision will concur with the results from the 95% confidence interval for the difference in means (with a rejection level of 0.05)

42

Paired t-test: Approach

- Set up the two competing hypotheses about the unknown population means

$$H_o : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- Assume H_o true: Compute how far the observed result (sample mean difference) is from the expected difference of 0

43

Summary

- Translate the distance into a p-value and make a decision
- The p-value measure the chance of getting the study results (or something even less likely, ie: more extreme) when the samples are assumed to have come from populations with the same means
- The p-value (this a called a two-sided p-value, more to come) is invariant to the direction of comparison

44

Section C: (Hypothesis Testing) Comparing Means Between Two Populations: The Unpaired Approach

45

Learning Objectives

- In this lecture set you will learn how to estimate and interpret a p-value for a hypothesis test of mean difference between two populations for the unpaired (two independent groups) study design
- The method for getting the p-value is called the unpaired t-test (or the two-sample t-test)
  - "unpaired" because of the study design
  - "t-test" because (sometimes) the sampling distribution is a t-distribution

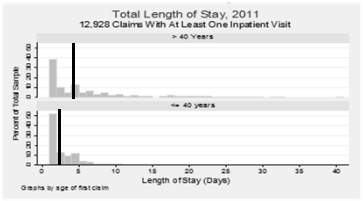
46

Example 1:Unpaired (Two Independent Groups)

- Hospital length of stay, by age of first claim (Heritage Health[3])

$\bar{x}_{>40\ \text{years}} = 4.9$ days
$s_{>40\ \text{years}} = 3.1$ days
$n_{>40\ \text{years}} = 3,770$

$\bar{x}_{\leq 40\ \text{years}} = 2.7$ days
$s_{\leq 40\ \text{years}} = 4.9$ days
$n_{\leq 40\ \text{years}} = 9,158$



47

Example 1:Unpaired (Two Independent Groups)

- Mean difference in hospital length of stay and 95% CI (older – younger age of first diagnosis)

  2.2 days (2.05 days, 2.35 days)

48

Hypothesis Testing Approach

- Set up the two competing hypotheses

$$H_o: \mu_{>40} = \mu_{\leq 40} \qquad H_o: \mu_{>40} - \mu_{\leq 40} = 0$$
$$H_A: \mu_{>40} \neq \mu_{\leq 40} \qquad H_A: \mu_{>40} - \mu_{\leq 40} \neq 0$$

- Assume H_o is true: Figure out "how far" observed mean difference ($\bar{x}_{>40} - \bar{x}_{\leq 40}$) is from expected difference under H_o (0) in terms of standard errors. The distance measure is:

$$t = \frac{(\bar{x}_{>40} - \bar{x}_{\leq 40}) - 0}{\hat{SE}(\bar{x}_{>40} - \bar{x}_{\leq 40})} = \frac{(\bar{x}_{>40} - \bar{x}_{\leq 40})}{\left(\sqrt{\frac{s_{>40}^2}{n_{>40}} + \frac{s_{\leq 40}^2}{n_{\leq 40}}}\right)} = \frac{2.2}{0.075}$$

49

Hypothesis Testing Approach

- Translate distance into p-value by comparing it to the distribution of such differences because of sampling variability when the true, population level difference is 0
- Compare the p-value to the preset rejection level (alpha level): for our purposes, and most of the research world, this is 0.05
- We have a result that is 29.3 standard errors above the expected mean difference of 0 (under the null hypotheses): How likely is this to occur just by chance (because of random sampling error)?

50

Hypothesis Testing Approach

- Getting the p-value
- The resulting p-value is very small (<0.0001): Interpretation?

51

Hypothesis Testing Approach

- Making a decision
- How does this decision compared to the decision we would make from the 95% CI for the difference in population means?

52

Hypothesis Testing Approach

- Note: p-value invariant to direction of comparison. Suppose we had instead presented the estimate, and the hypotheses in terms of the 40 and under group compared to the over 40 group?

$$H_o: \mu_{>40} = \mu_{\leq 40} \qquad H_o: \mu_{\leq 40} - \mu_{>40} = 0$$
$$H_A: \mu_{>40} \neq \mu_{\leq 40} \qquad H_A: \mu_{\leq 40} - \mu_{>40} \neq 0$$

- The distance measure:

$$t = \frac{(\bar{x}_{\leq 40} - \bar{x}_{>40}) - 0}{\hat{SE}(\bar{x}_{\leq 40} - \bar{x}_{>40})} = \frac{(\bar{x}_{\leq 40} - \bar{x}_{>40})}{\left(\sqrt{\frac{s_{>40}^2}{n_{>40}} + \frac{s_{\leq 40}^2}{n_{\leq 40}}}\right)} = \frac{-2.2}{0.075}$$

53

Hypothesis Testing Approach

- Getting the p-value

54

Example 2

- "A Low Carbohydrate as Compared with a Low Fat Diet in Severe Obesity"[1]
  - 132 severely obese subjects randomized to one of two diet groups
  - Subjects followed for six month period
- At the End of Study Period
  - "Subjects on the low-carbohydrate diet lost more weight than those on a low fat diet (95% confidence interval for the difference in weight loss between groups, -1.6 to - 6.2 kg; p< .01)"

[1] Samaha, F., et al. A Low-Carbohydrate as Compared with a Low-Fat Diet in Severe Obesity, *New England Journal of Medicine*, 348: 21

55

Example 2

- Scientific Question—Is Weight Change Associated with Diet Type?

	Diet Group	
	Low-Carb	Low-Fat
Number of subjects (n)	64	68
Mean weight change (kg) Post-diet less pre-diet	-5.7	-1.8
Standard deviation of weight changes (kg)	8.6	3.9

56

Example 2

- Using the data from the weight change/diet type study

$$\bar{x}_{LC} - \bar{x}_{LF} = -5.7 - (-1.8) = -3.9 \text{ kg}$$

$$\hat{SE}(\bar{x}_{LC} - \bar{x}_{LF}) = \sqrt{\frac{8.6^2}{64} + \frac{3.9^2}{68}} \approx 1.17 \text{ kg}$$

57

Example 2: 95% CI

- 95% CI for mean difference in weight change

$$(\bar{x}_{LC} - \bar{x}_{LF}) \pm 2\,\hat{SE}(\bar{x}_{LC} - \bar{x}_{LF})$$

$$-3.9 \pm 2 \times 1.17$$

$$-6.24 \text{ kg to } -1.56 \text{ kg} \approx (-6.2 \text{ kg}, -1.6 \text{ kg})$$

58

Example 2:Hypothesis Testing Approach

- Set up the two competing hypotheses

$H_o: \mu_{LC} = \mu_{LF}$ $\qquad$ $H_o: \mu_{LC} - \mu_{LF} = 0$

$H_A: \mu_{LC} \neq \mu_{LF}$ $\qquad$ $H_A: \mu_{LC} - \mu_{LF} \neq 0$

- Assume H_o is true: Figure out "how far" observed mean difference $(\bar{x}_{doctor2} - \bar{x}_{doctor1})$ is from expected difference under H_o (0) in terms of standard errors. The distance measure is:

$$t = \frac{(\bar{x}_{LC} - \bar{x}_{LF}) - 0}{\hat{SE}(\bar{x}_{LC} - \bar{x}_{LF})} = \frac{-3.9}{1.17} \approx -3.3$$

59

Hypothesis Testing Approach

- Translate distance into p-value by comparing it to the distribution of such differences because of sampling variability when the true, population level difference is 0
- Compare the p-value to the preset rejection level (alpha level): for our purposes, and most of the research world, this is 0.05
- We have a result that is 3.3 standard errors below the expected mean difference of 0 (under the null hypotheses): How likely is this to occur just by chance (because of random sampling error)?

60

Hypothesis Testing Approach

- Getting the p-value
- The resulting p-value is very small (<0.01): Interpretation?

61

Hypothesis Testing Approach

- Making a decision
- How does this decision compared to the decision we would make from the 95% CI for the difference in population means?

62

Example 3: Unpaired (Two Independent Groups)

- Menu Labeling and Calorie Intake[4]

> *Objectives.* We assessed the impact of restaurant menu calorie labels on food choices and intake.
> *Methods.* Participants in a study dinner (n=303) were randomly assigned to either (1) a menu without calorie labels (no calorie labels), (2) a menu with calorie labels (calorie labels), or (3) a menu with calorie labels and a label stating the recommended daily caloric intake for an average adult (calorie labels plus information). Food choices and intake during and after the study dinner were measured.
> *Results.* Participants in both calorie label conditions ordered fewer calories than those in the no calorie labels condition. When calorie label conditions were combined, that group consumed 14% fewer calories than the no calorie labels group. Individuals in the calorie labels condition consumed more calories after the study dinner than those in both other conditions. When calories consumed during and after the study dinner were combined, participants in the calorie labels plus information group consumed an average of 250 fewer calories than those in the other groups.
> *Conclusions.* Calorie labels on restaurant menus impacted food choices and intake; adding a recommended daily caloric requirement label increased this effect, suggesting menu label legislation should require such a label. Future research should evaluate menu labeling's impact on children's food choices and consumption. (*Am J Public Health.* 2010;100:312–318. doi:10.2105/AJPH.2009.160226)

[4] Roberto C, et al. Evaluating the Impact of Menu Labeling on Food Choices and Intake. *American Journal of Public Health* (2010); 100(2); 313-318.

63

Example 3: Unpaired (Two Independent Groups)

- Figure from article (plus information from a separate table)

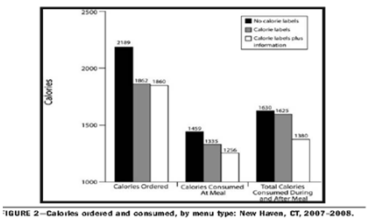


FIGURE 2—Calories ordered and consumed, by menu type: New Haven, CT, 2007–2008.

$\bar{x}_{\text{no calorie labels}} = 1{,}630$ calories
$s_{\text{no calorie labels}} = 810$ calories
$n_{\text{no calorie labels}} = 95$

$\bar{x}_{\text{calorie labels}} = 1{,}625$ calories
$s_{\text{calorie labels}} = 741$ calories
$n_{\text{calorie labels}} = 96$

$\bar{x}_{\text{calorie labels+info}} = 1{,}380$ calories
$s_{\text{calorie labels+info}} = 639$ calories
$n_{\text{calorie labels+info}} = 106$

64

Example 3: Unpaired (Two Independent Groups)

- Resulting mean differences and 95% CIs

$\bar{x}_{\text{no calorie labels}} - \bar{x}_{\text{calorie labels}} = 1{,}630 - 1{,}625 = 5$ calories

  - 95% CI (-216.7, 226.7)

$\bar{x}_{\text{no calorie labels}} - \bar{x}_{\text{calorie labels+info}} = 1{,}630 - 1380 = 250$ calories

  - 95% CI (45.3 , 454.7)

$\bar{x}_{\text{calorie labels}} - \bar{x}_{\text{calorie labels+info}} = 1{,}625 - 1380 = 245$ calories

  - 95% CI (62.0, 448.0)

65

Example 3: Unpaired (Two Independent Groups)

- Resulting mean differences and 95% CIs with p-values

$\bar{x}_{\text{no calorie labels}} - \bar{x}_{\text{calorie labels}} = 1{,}630 - 1{,}625 = 5$ calories

  - 95% CI (-216.7, 226.7) p=0.96

$\bar{x}_{\text{no calorie labels}} - \bar{x}_{\text{calorie labels+info}} = 1{,}630 - 1380 = 250$ calories

  - 95% CI (45.3 , 454.7) p=0.017

$\bar{x}_{\text{calorie labels}} - \bar{x}_{\text{calorie labels+info}} = 1{,}625 - 1380 = 245$ calories

  - 95% CI (62.0, 448.0) p=0.013

66

A Note About Unpaired Studies and Results

- For "smaller samples" slight corrections need to made to the number of estimated standard errors added and subtracted to get 95% coverage

67

FYI: Equal Variances Assumption

- The test I am showing you is formally called "the two sample t-test assuming unequal population variances"
- The "traditional" t-test ("the two sample t-test assuming equal population variances")assumes equal variances in the two populations being compared via the two samples
  - This can be formally tested with another hypothesis test?!?!
  - But why not just compare observed values of s_1 to s_2?

68

FYI: Equal Variances Assumption

- There is a slight modification to allow for unequal variances—this modification adjusts the degrees of freedom for the test, using slightly different SE computation (the formula I give you)
- If you want to be truly "safe" (desert island choice of t-test)
  - More conservative to use test that allows for unequal variance
- Makes little to no difference in large samples

69

FYI: Equal Variances Assumption

- Actually, the following occurs:
- If underlying population level standard deviations are equal:
  - both approaches give valid confidence intervals, but intervals by approach assuming unequal standard deviations slightly wider (and p-values slightly larger)
- If underlying population level standard deviations are not equal:
  - The approach assuming equal variances does not give valid confidence intervals and can severely under-cover the goal of 95%

70

Summary

- The (two sample) unpaired paired t-test is a method for getting a p-value for testing the computing hypotheses

$$H_o : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

  using data from unpaired samples from two independent populations

- The resulting decision will concur with the results from the 95% confidence interval for the difference in means (with a rejection level of 0.05)

71

Unpaired t-test: Approach

- Set up the two competing hypotheses about the unknown population means

$$H_o : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- Assume H_o true: Compute how far the observed result (sample mean difference) is from the expected difference of 0

72

Summary

- Translate the distance into a p-value and make a decision
- The p-value measure the chance of getting the study results (or something even less likely, ie: more extreme) when the samples are assumed to have come from populations with the same means

73

Section D: Debriefing on the p-value, Part 1

74

Learning Objectives

- In this lecture section , the focus will be on what a p-value can and can't reveal about the study results. Upon completion of this section, you will be able to
  - Define type 1 error and understand it's role in the hypothesis testing process
  - Explain what a p-value is , and what it is not
  - Contrast statistical significance with scientific significance
  - Start to appreciate why a non-statistically significant result yields a decision of "fail to reject the null hypothesis"

75

p-values

- p-values are probabilities (numbers between 0 and 1)
- Small p-values mean that the sample results are unlikely when the null is true
- The p-value is the probability of obtaining a result as/or more extreme than you did by chance alone assuming the null hypothesis H_0 is true
  - how likely your sample result (and other results less likely) are if null is true

76

p-values

- The p-value is NOT :
  - The probability that the null hypothesis is true
  - The probability that the study was well conducted
  - The probability that the study results are important
  - The probability that the alternative hypothesis is true
  - The probability that the study findings are legitimate

77

p-values

- The p-value alone imparts no information about scientific/substantive content in result of a study
  - For example: In the cornflake/oat bran study, the researchers found a statistically significant (p=0.0065!) difference in average LDL cholesterol levels in men who had been on a diet including corn flakes versus the same men on a diet including oat bran cereal
  - Which diet showed lower average LDL levels?
  - How much was the difference, does it mean anything nutritionally?

78

p-values

- If the p-value is small either a very rare event occurred and H_0 is true OR H_0 is false
- Type I error
  - Reject H_o in favor of H_A when in fact H_0 is true
  - The probability of making a Type I error is called the *alpha-level* (α-level) or *significance level*
  - This is set in advance of performing the test, and the standard is 0.05

79

Note on the p-value and the alpha-Level

- If the p-value is less then some pre-determined cutoff (e.g. .05), the result is called "statistically significant"
- This cutoff is the α-*level:* The α-*level* is the probability of a type I error

80

Note on the p-value and the alpha-Level

- It is the probability of falsely rejecting H_0 when H_0 true: probabililty of a false positive
- Idea: keep chance of " making a mistake" when H_0 true low and only reject if sample result "unlikely"
  - Unlikeliness threshold determined by α-*level*

81

Note on the p-value and the alpha-Level

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	**H_A**
Reject H_o		
Not Reject H_o		

82

Type 2 Error and Power

- Truth versus decision made by hypothesis testing

	TRUTH	
	H_o	**H_A**
Reject H_o		
Not Reject H_o		

83

Type 2 Error and Power

- If $p \geq 0.05$, why is the decision phrased as "fail to reject the null" as opposed to "accepting the null"?

84

Connection: Hypothesis Testing and CIs

- The confidence interval gives plausible values for the population parameter
  - "data take me to the truth"
- Hypothesis testing postulates two choices for the population parameter
  - "here are two possibilities for the truth, data help me choose one"

85

95% Confidence Interval

- If 0 is not in the 95% CI, then we would reject H_0 that $\mu = 0$ at level $\alpha = .05$ (the p-value < .05)
- Why?

86

95% Confidence Interval

- If 0 is not in the 95% CI, then we would reject H_0 that $\mu_1 - \mu_2 = 0$ at level $\alpha = .05$ (the p-value < .05)
- Why?

87

95% Confidence Interval and p-value

- So, in the BP/OC example, the 95% confidence interval tells us that the p-value is less than .05, but it doesn't tell us that it is p = .009
- The confidence interval and the p-value are complementary
- However, you can't get the exact p-value from just looking at a confidence interval, and you can't get a sense of the scientific/substantive significance of your study results by looking at a p-value

88

More on the p-value

- Statistical Significance Does Not Imply/Prove Causation
- Ex: In blood pressure/oral contraceptives example there could be other factors that could explain the change in blood pressure
- A significant p-value is only ruling out random sampling (chance) as the explanation
- Need a comparison group to better establish causality
  - Self-selected (may be okay)
  - Randomized (better)

89

Statistical Significance ≠ Scientific Significance

- Statistical significance is not the same as scientific significance
- Hypothetical Example: Blood Pressure and Oral Contraceptives: Suppose:
  - n = 100,000; $\bar{x}_{diff}$ = .03 mmHg; s = 4.6 mmHg
  - p-value = .04

90

Statistical Significance ≠ Scientific Significance

- Big *n* can sometimes produce a small p-value even though the magnitude of the effect is very small (not scientifically/substantively significant)
- Very Important
  - Always report a confidence interval
  - 95% CI: 0.002 - 0.058 mmHg

91

Lack of Statistical Significance

- Lack of statistical significance is not the same as lack of scientific significance: must evaluate in context of study, sample size
- Small *n* can sometimes produce a non-significant even though the magnitude of the association at the population level is real and important (our study just can't detect it)
- Sometimes small studies are designed without power in mind just to generate preliminary data

92

Summary

- The p-value alone can only indicate whether the study results were likely due to (random sampling) chance or not if there is no difference in the measure being compared between populations (so far, we have only looked at comparing means, but this idea will hold for proportions and rates as well)
- Not rejecting the null hypothesis is not equivalent to accepting the null hypothesis as the truth: we will dig deeper into this in lectures 12 and 13

93