Exam 2 - Study Guide and Practice

Vojtech Kejzlar

Content

Chapter 4.

- Based on a research scenario be able to identify the following:
  - explanatory and response variables and their types
  - confounding variables and how they affect/relate to both explanatory and response variables
  - the type of the study: observational study or an experiment
- Understand the difference between association and causation
- Understand the role of blinding and randomization in experiments

Chapter 5.

- Given an inference scenario comparing two groups with a categorical response variable, be able to
  - express the research question in the form of null and alternative hypotheses about the difference between population proportions $\pi_1 - \pi_2$
  - express H_0 and H_a in terms of both population parameters and as a full sentence
  - asses the strength of evidence against H_0 using the p-value
  - state the conclusion of your inference based on the p-value in the context of the problem
  - judge the possibility of a cause-and-effect relationship if H_0 is rejected
  - construct a theory-based confidence interval for the difference between population proportions $\pi_1 - \pi_2$ and interpret it in the context of the problem
- Numerical (contingency table) and graphical summaries (segmented bar-chart) of categorical variables
- Simulation-based hypothesis test about the difference between population proportions $\pi_1 - \pi_2$
- Central limit theorem (CLT) for the difference between sample proportions $\hat{p}_1 - \hat{p}_2$ and its role in hypothesis testing
- Be able to carry out theory-based hypothesis test about the difference between population proportions $\pi_1 - \pi_2$ and judge its validity

Chapter 6.

- Given an inference scenario comparing two independent groups with a quantitative response variable, be able to
  - express the research question in the form of null and alternative hypotheses about the difference between population means $\mu_1 - \mu_2$
  - express H_0 and H_a in terms of both population parameters and as a full sentence
  - asses the strength of evidence against H_0 using the p-value
  - state the conclusion of your inference based on the p-value in the context of the problem
  - judge the possibility of a cause-and-effect relationship if H_0 is rejected
  - construct a theory-based confidence interval for the difference between population means $\mu_1 - \mu_2$ and interpret it in the context of the problem
- Numerical (5-number summary) and graphical summaries (boxplot) of quantitative variables
- Simulation-based hypothesis test about the difference between population means $\mu_1 - \mu_2$
- Central limit theorem (CLT) for the difference between sample means $\bar{x}_1 - \bar{x}_2$ and its role in hypothesis testing
- Be able to carry out theory-based hypothesis test about the difference between population means $\mu_1 - \mu_2$ and judge its validity

Chapter 7.

- Given an inference scenario comparing two groups with a quantitative response variable, be able to identify independent groups design or paired data design
- Given an inference scenario with paired data design, be able to

  - express the research question in the form of null and alternative hypotheses about the population mean difference μ_{diff}
  - express H_0 and H_a in terms of both population parameters and as a full sentence
  - asses the strength of evidence against H_0 using the p-value
  - state the conclusion of your inference based on the p-value in the context of the problem
  - judge the possibility of a cause-and-effect relationship if H_0 is rejected
  - construct a theory-based confidence interval for the the population mean difference μ_{diff} and interpret it in the context of the problem
- Simulation-based hypothesis test about the population mean difference μ_{diff}
- Central limit theorem (CLT) for the sample mean difference $\bar{x}_d$ and its role in hypothesis testing
- Be able to carry out theory-based hypothesis test about the population mean difference μ_{diff} and judge its validity

Practice problems

(1) An article in a 2006 issue of the Journal of Behavioral Decision Making reports on a study involving 47 undergraduate students at Harvard. All of the participants were given \$50, but some (chosen at random) were told that this was a "tuition rebate" while others were told that this was "bonus income." After one week, the students were contacted again and asked how much of the \$50 they had left, meaning how much they had saved. Researchers wanted to know whether the students at Harvard receiving the "rebate" would tend to save more money than those receiving the "bonus."

(a) Is this am observational study or an experiment? Was blinding and/or randomization used? Justify your answers.

(2) Here is the beginning of an article from the Yahoo! Health News website on January 7, 1998:

Walking two miles or more per day can cut the overall risk of dying in half, according to a new study. It also reduces the risk of dying from cancer – and appears to cut the risk of death due to cardiovascular diseases, US researchers report. Between 1980 and 1982, multicenter researchers in the Honolulu Heart Program studied 707 nonsmoking, retired men, aged 61 to 81 years, and collected mortality data on these men over the following 12 years. During the study, 208 of the men died. The study results show that while 43.1% of men who walked less than one mile per day died, only half this figure – 21.5% – of the men who walked more than two miles per day died.

(a) [1 point] What is the research question?

(b) [2 points] Clearly explain why a person's general health is a confounding factor in the relationship between walking and the risk of dying.

(c) [2 points] Does this study imply that walking 2 miles each day causes reduction in risk of dying?

(3) Circle true or false for each statement below. If false, say why.

(a) True False: Association always implies cause-and-effect relationship between two variables.

(b) True False: P-value for two-sided test is roughly twice as large as for one-sided

(c) True False: A p-value can be negative

(d) True False: A p-value can be larger than 1

(e) True False: If we fail to reject the null hypothesis, we prove that it is true.

(f) True False: There is no difference in how we analyze data based on independent groups design and paired data design.

(g) True False: If we conduct an observational study, we can cautiously conclude a causal relationship.

(h) True False: Any confidence interval always covers the population parameter of interest.

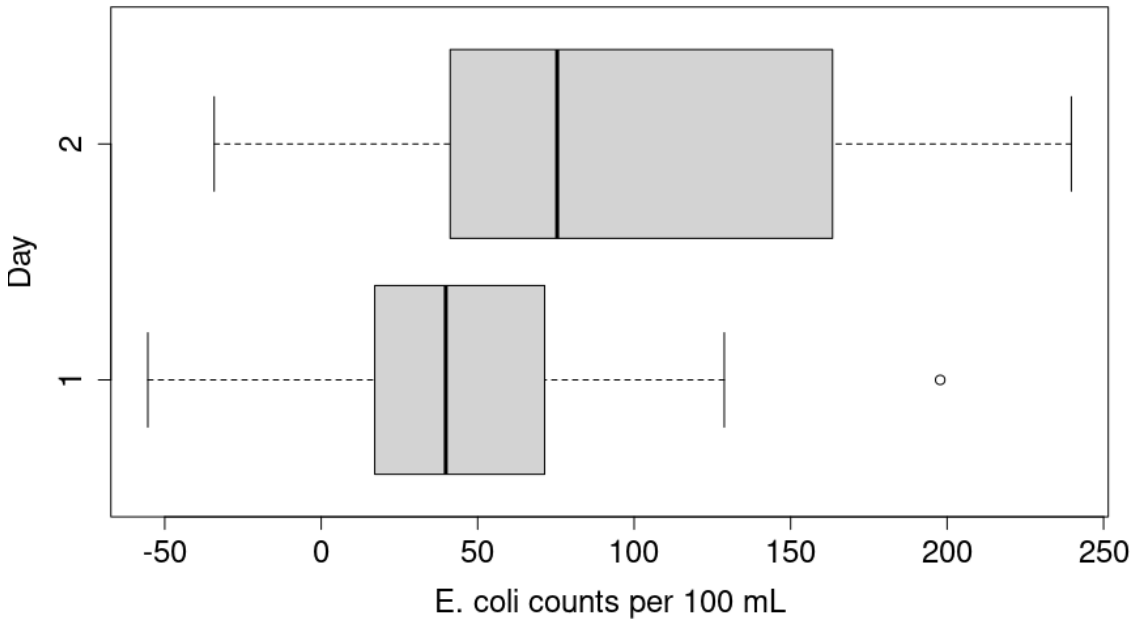
(i) True False: The parameter of interest for paired data design is the difference between population means.

(j) True False: The upper quartile separates the bottom 25% of the data from the top 75%.

(4) When stating null and alternative hypotheses, the hypotheses are

(a) about the statistic only
(b) about both the parameter and the statistic
(c) about the parameter only
(d) sometimes about the parameter and sometimes about the statistic

(5) The boxplots below corresponds to observed *E. coli* counts (per 100 mL) in a small village in central Africa during the first and second day after filtration.



Select the largest value
(a) Maximum *E. coli* count during the first day
(b) Upper quartile *E. coli* count during the second day
(c) Median *E. coli* count during the second day

Compute the 5-number summary for *E. coli* counts during both days.

(6) The National Weather Service maintains a weather station at the Albany International Airport. Data for the 700 days between January 1, 2015 and November 30, 2016 show the following results:

Day type	Precipitation	No Precipitation	Total
Weekday	265	235	500
Weekend	119	81	200
Total	384	316	700

Research question: Is there a significant difference in the proportion of rainy weekends and weekdays in Albany?

(a) What is the proportion of weekdays with no precipitation?

(b) What is the proportion of all days during which rained.

(c) What is the population of interest?

(d) What is the parameter of interest in this study? What symbol do we use for it? Be as specific as possible!

(e) What is the value of observed statistic. Use appropriate symbol!

(f) State the appropriate null hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

(g) State the appropriate alternative hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

(h) Use the the plots below that represent the area under the normal distribution to determine if we have enough evidence to reject the null hypothesis. Make sure that you calculate the value of standardized statistic and that you state your conclusion in the context of the problem. Use significance level $\alpha = 0.05$.

TWO-SIDED

$Z = 1.56$

$\alpha = 0.1$

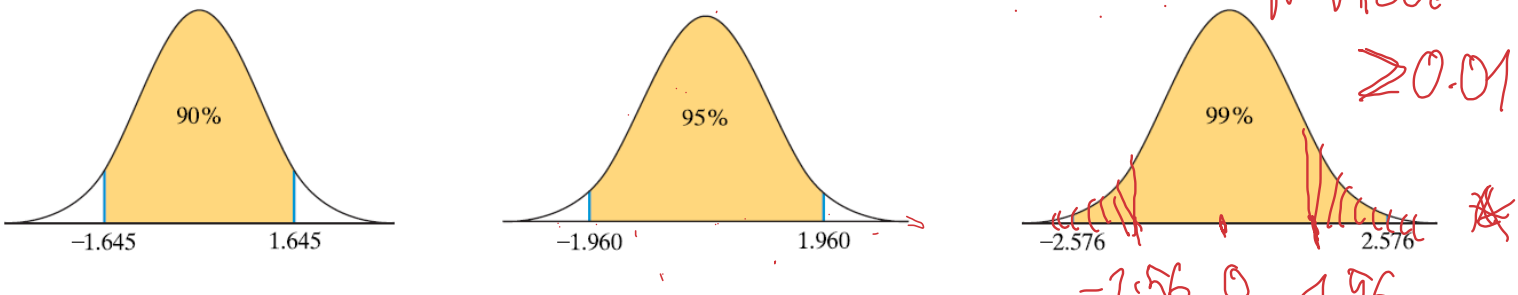


FIGURE 1. Z scores and area under the normal curve.

(i) Comment on validity of using the theory-based inference for this problem.

(j) Explain in one sentence what a Type I error would mean in this context.

(k) Explain in one sentence what a Type II error would mean in this context.

(l) Use the following table to construct a 95% confidence interval for the difference between population proportions. Calculate the margin of error. Interpret the interval.

Confidence level	Multiplier
90	1.645
95	1.960
99	2.576

(m) Does your confidence interval contain the value 0? Are you surprised about the result? Explain why we could expect that it does or does not contain the value 0.

(7) The U.S. Department of Transportation provides the number of miles that residents of the 75 largest metropolitan areas travel per day in a car. Suppose that for a random sample of 50 Buffalo residents the mean is 22.5 miles a day and the standard deviation is 8.4 miles a day, and for an independent random sample of 50 Boston residents the mean is 18.6 miles a day and the standard deviation is 7.4 miles a day. A researcher wants to test to see if there is a significant difference in mean travel distances between the two cities.

(a) Is this an independent groups scenario or paired data scenario? Explain.

(b) What is the population of interest?

(c) What is the parameter of interest in this study? What symbol do we use for it? Be as specific as possible!

(d) What is the value of observed statistic. Use appropriate symbol!

(e) State the appropriate null hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

$H_0: \mu_{MOC} - \mu_{MOS} = 0$

(f) State the appropriate alternative hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

$H_a: \mu_{MOC} - \mu_{MOS} \neq 0$

(g) Use the the plots below that represent the area under the T-distribution to determine if we have enough evidence to reject the null hypothesis. Make sure that you calculate the value of standardized statistic t and that you state your conclusion in the context of the problem. Use significance level $\alpha = 0.05$.

$t = 2.46$

IS P-VALUE SMALLER THAN α?

H_a:

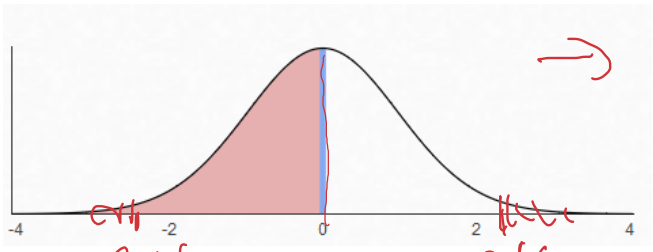


FIGURE 2. The shaded area under the curve is 0.5

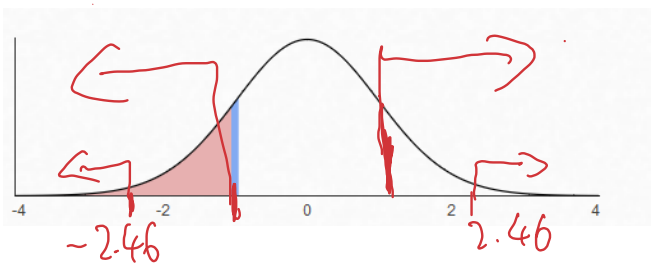


FIGURE 3. The shaded area under the curve is 0.16

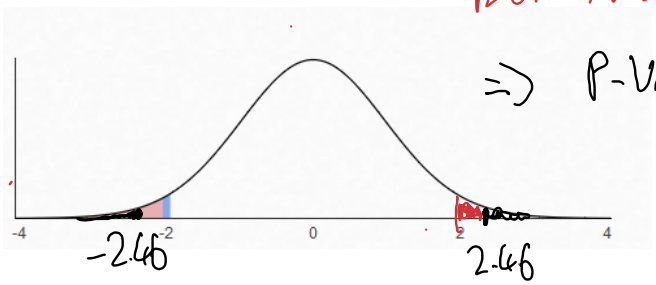


FIGURE 4. The shaded area under the curve is 0.025

(h) Comment on validity of using the theory-based inference for this problem.

(i) Use the following table to construct a 99% confidence interval for the population parameter. Calculate the margin of error. Interpret the interval.

Confidence level	Multiplier
90	1.677
95	2.010
99	2.680

(8) 50 experts rated two brands of Colombian coffee in a taste-testing experiment. A rating on 7-point scale was given for each of four characteristics. The table below shows the summary ratings for each expert on each coffee.

Research question: Does the coffee B have, on average, better rating than coffee A?

COFFEE A - COFFEE B → $\bar{x}_D$

Expert	Coffee A	Coffee B	Difference
Mean	24	25.5556	-1.5556
SD	2.6458	1.8782	1.4240

(a) Is this an independent groups scenario or paired data scenario? Explain.

Paired - Each expert is testing both coffees

(b) What is the parameter of interest in this study? What symbol do we use for it? Be as specific as possible!

μ_{DIF} – The mean difference between the scores of coffee A and coffee B

(c) What is the value of observed statistic. Use appropriate symbol!

$\bar{x}_D = -1.5556$

(d) State the appropriate null hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

H_0: $\mu_{DIFF} = 0$

(e) State the appropriate alternative hypothesis for this scenario. Do so both using a full sentence and symbols representing population parameters.

H_a: $\mu_{DIFF} < 0$

(f) Use the the plots below that represent the area under the T-distribution to determine if we have enough evidence to reject the null hypothesis. Make sure that you calculate the value of standardized statistic t and that you state your conclusion in the context of the problem. Use significance level $\alpha = 0.05$.

$t = \frac{\bar{x}_D}{s_D/\sqrt{n}}$

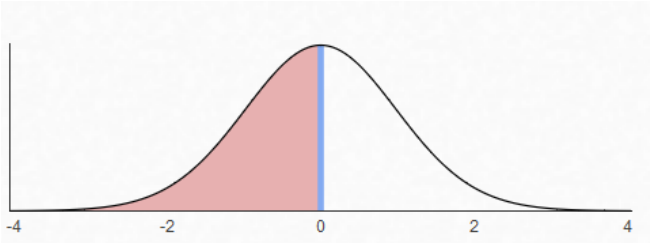


FIGURE 5. The shaded area under the curve is 0.5

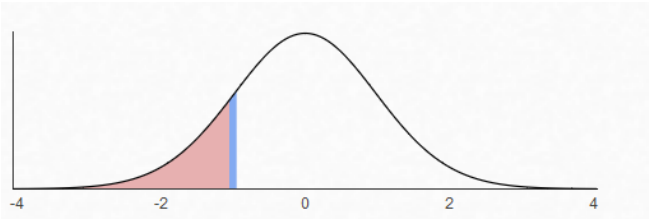


FIGURE 6. The shaded area under the curve is 0.16

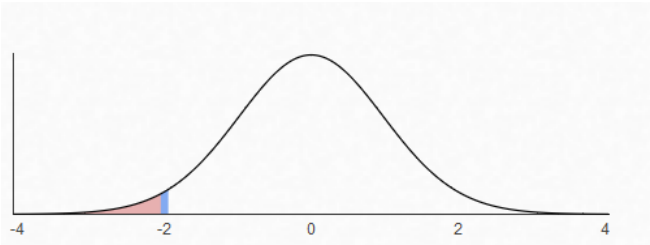


FIGURE 7. The shaded area under the curve is 0.025

(g) Comment on validity of using the theory-based inference for this problem.

(h) Use the following table to construct a 90% confidence interval for the population parameter. Calculate the margin of error. Interpret the interval.

$SE = \frac{s_d}{\sqrt{n}} = \frac{1.424}{\sqrt{50}}$

Confidence level	Multiplier
90	1.677
95	2.010
99	2.680

$\bar{x}_d \pm$ MULTIPLIER $\cdot$ SE

$\bar{x}_d \pm 1.677 \cdot SE$

→ CHAPTER 7

BOOK PRACTICE PROBLEMS

5.CE.4, 5.CE.12, 6.CE.11, 6.CE.12, 7.CE.1, 7.CE.7, 7.CE.15, 7.CE.16