Inference for Numerical Data I

Dajiang Liu

@PHS 525

Feb-18-2016

How to Select Significance Threshold Levels

- Is it okay to change hypothesis after seeing the data?
  - The answer is NO
- What is the right threshold to use:
  - $\alpha = 0.05$ is an arbitrary choice
- The actual choice depends on the consequence of type I and II errors?
  - Examples:
    - Adverse drug effects
      - More important to have sufficient power for detecting negative drug effects
      - Minimize type II errors are more important
    - Differential expression analyses:
      - More important to control for false positives
        - Following up false positives in wet lab experiment can be costly
      - Minimize type I errors are more important

Exercise from Ch. 4 - Problem 4.17

- Online communication. A study suggests that the average college student spends 2 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and nd that on average they spent 3.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0: \bar{x} < 2 \text{ v.s. } H_A: \bar{x} \geq 3.5$$

Exercise from Ch. 4 - Problem 4.17

- Online communication. A study suggests that the average college student spends 2 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and nd that on average they spent 3.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0: \bar{x} < 2 \text{ v.s. } H_A: \bar{x} \geq 3.5$$

  - Answer: the hypothesis should be about the parameters. $\bar{x}$ is a random variable and the sample mean, which cannot be used to formulate a hypothesis. The right hypothesis should be

$$H_0: \mu < 2 \text{ v.s. } H_A: \mu \geq 3.5$$

Exercise from Ch. 4 – Problem 4.28

- A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.
- (a) Write the hypotheses in words.
- (b) What is a Type 1 error in this context?
- (c) What is a Type 2 error in this context?
- (d) Which error is more problematic for the restaurant owner? Why?
- (e) Which error is more problematic for the diners? Why?
- (f) As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.

Exercise from Ch. 4 – Problem 4.28

- A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.
- (a) Write the hypotheses in words:
- Answer: the null hypothesis is that the restaurant does not violate the sanitation regulation; the alternative hypothesis is the restaurant violates the regulation.
- (b) What is a Type 1 error in this context?
- Answer: the restaurant does not violate regulation, yet has the license revoked.
- (c) What is a Type 2 error in this context?
- Answer: the restaurant violates the regulation, yet does not have its license revoked.
- (d) Which error is more problematic for the restaurant owner? Why?
- Answer: Type I error
- (e) Which error is more problematic for the diners? Why?
- Answer: type II error
- (f) As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.
- Answer: Diner would be better off with a less stringent type I error threshold.

Exercise from Ch. 4 – Problem 4.30

- A car insurance company advertises that customers switching to their insurance save, on average, $432 on their yearly premiums. A market researcher at a competing insurance discounter is interested in showing that this value is an overestimate so he can provide evidence to government regulators that the company is falsely advertising their prices. He randomly samples 82 customers who recently switched to this insurance and finds an average savings of $395, with a standard deviation of $102.
- (a) Perform a hypothesis test and state your conclusion.
- (b) Do you agree with the market researcher that the amount of savings advertised is an overestimate? Explain your reasoning.
- (c) Calculate a 90% confidence interval for the average amount of savings of all customers who switch their insurance.
- (d) Do your results from the hypothesis test and the confidence interval agree? Explain.

What Have We Learned in Ch. 4

- Confidence interval for a point estimate
- Hypothesis testing using confidence intervals
- P-values for testing hypothesis on sample mean
- Hypothesis testing with p-values

What Will We Learn in Ch. 5

- What is the right sample (mean) estimate to use?
- How are these sample (mean) estimates distributed
- How to perform hypothesis testing using these distributions with methods in Ch. 4

Inference on Paired Data

- Paired data
  - For data point in each group, a corresponding data point will be available for the other group

- Example:
  - Book prices
    - Amazon and UCLA bookstore price
  - Gene expression levels in different tissues

Inference on the Paired Data

- Compare the mean values for different groups
  - Data point:
    - X_{diff}~UCLA book price – Amazon book price
    - X_{diff} is a new random variable
    - We can calculate its mean and variances

- Hypothesis on the average book price between two sellers:
  - $H_0: \mu_{diff} = 0$
  - Question:
    - What is the correct hypothesis testing procedure?

Algorithm for Testing the Mean Difference for Paired Data

- For each pair of values, obtain their difference
  - We define the random variable that measures the difference as X_{diff}
- Calculate the sample mean estimator, i.e. $\bar{X}_{diff}$
- Calculate the sample standard deviation $\sigma_{X_{diff}}$ for X_{diff}
- The sample mean standard error is given by

$$SE\left(\bar{X}_{diff}\right) = \frac{\sigma_{X_{diff}}}{\sqrt{N}}$$

- Compute the Z-score:

$$Z = \frac{\left(\bar{X}_{diff} - \mu_{diff}\right)}{se\left(\bar{X}_{diff}\right)}$$

- Compute p-values from the Z-scores.

Testing for Sample Mean Difference Between Two Samples

- In many scenarios, the data points may not be matched

- For example,
  - The income levels for Penn and Maryland
  - The running time for males and females

- How to compare these differences?
  - A natural choice is to use the sample mean differences
  - $\bar{X}_{Group1} - \bar{X}_{Group2}$

Testing for Sample Mean Difference Between Two Samples

- Point estimate

$$\bar{X}_{Group_1} - \bar{X}_{Group_2}$$

- Standard error for the point estimate

$$se\left(\bar{X}_{Group_1} - \bar{X}_{Group_2}\right) = \sqrt{\frac{\sigma^2_{Group_1}}{N_1} + \frac{\sigma^2_{Group_2}}{N_2}}$$

For testing the hypothesis

$$H_0: \mu_{diff} = 0$$
$$H_A: \mu_{diff} \neq 0$$

Example: Baby_smoke dataset

- Test whether smoker/non-smoker mothers give babies with different birth weight? (Table 5.9)

	Smoker	Non-smoker
Mean	6.78	7.18
SD	1.43	1.60
Sample Size	50	100

Example: Cherry Blossom Run Time

- Test whether males and female run times are different (Table 5.5)?

	Men	Women
$\bar{X}$	86.75	102.13
SD	12.5	15.2
n	45	55

Homework Problem

- Exercise 4.7, 4.8, 4.14, 4.21, 4.25
  - Due March 1nd