Introductory Statistics Lectures

Review for Final Exam

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:51:54 2009)

Contents

1 Review for Final Exam

1.1 Cautions

A researcher developing a new chemotherapy agent conducts an experiment where 10 treatment mice receive the new agent during a month period. The tumor mass in each mouse is measured before and after the treatment. The researcher observes that the average tumor mass in the treatment group has decreased by 12 grams. The best existing chemotherapy agent in use can only reduce tumor mass by 10 grams on average in one month. The researcher is elated and writes a paper submitting it to the New England Journal of Medicine.

Question 1. The paper only gives the above data and states that the new chemotherapy agent is more effective in decreasing tumor mass in comparison to the standard agent.

You are reviewing the paper for the journal. What do you think about the researcher's conclusions?

Question 2. What hypothesis test would you suggest the researcher uses?

Question 3. You are also unhappy that the researcher simply reported a difference of 12 grams. What should the researcher report to give more information on their experiment precision?

A drug company has developed a new agent that is supposed to reduce cholesterol. The drug company conducts a study of 1,000 individuals who have their cholesterol measured before and after receiving the drug. The results are not statistically significant. They cannot support the alternative hypothesis that the drug reduces cholesterol.

Since the survival of the company is dependent on the success of the drug, they get more investor money and repeat the study on another 1,000 patients. Again, they don't get statistically significant results. After repeating the study 10 times, the 10th study has a sufficiently small p-value to support that the drug reduces cholesterol!

The company only reports the good results from the 10th study to the FDA. The drug is approved since they have shown it is effective.[1]

Question 4. What's wrong with drawing conclusions from only the 10th study?

[1] This is a fictitious example that exaggerates what has actually happened in recent times.

1.2 Key concepts to take away from this class.

Sampling error

- If you don't measure the whole population, you cannot directly draw conclusions (or prove hypothesis) from study / experiment results.
- You must statistically analyze your results to show that they are significant — indicating that sampling error is not the likely cause of the observed results.

Whether you are designing a study or reading experimental results, all good studies and experiments should be composed of the following steps.

Design of Experiments and Studies

1. Determine your hypothesis and the margin of error you desire. (Better: determine desired power.)
2. Determine what to measure. Best to use ratio level of measurement. Worst is categorical.
3. Determine the necessary sample size to attain your desired margin of error.
4. Carefully plan your study/experiment and how you will randomize to avoid bias.
5. Collect your data, plot it, look at it, check for outliers, and ensure the test's requirements are satisfied.
6. Analyze the data, make the formal decision.
7. Form the final conclusion and report the results. Make sure to:
   a) Report the conclusion in clear understandable works.
   b) Report the p-value.
   c) Give a confidence interval for the parameter you are estimating.
   d) Discuss the meaning of results.

- Only draw conclusions about populations which your sample represents.
- Correlation does not imply causation.

We use **statistics** based on samples to estimate population **parameters**. Since a statistic is an estimate of a parameter, knowing the distribution of the statistic — **the sampling distribution** — tells us how "good" the statistic is.

- If the mean of the sampling distribution equals the parameter, the statistic is unbiased.
- The narrower the sampling distribution (smaller standard deviation of the statistic), the smaller the confidence interval for the parameter yielding a more precise estimate.

Question 5. What does the CLT state and why is it useful?

1.3 Hypothesis Tests

Know the hypothesis testing steps.

What you should know for each test

- Know the hypothesis H_0 and H_a.
- Know the requirements.

Formal decision

- Reject H_0 if p-value $\leq \alpha$
- Otherwise **fail** to reject H_0

Remember,

- The p-value is the probability of observing the sample data assuming H_0 is true. **We always assume H_0 is true unless the sample data supports rejecting it**.
- We can't make statements about supporting H_0 unless we calculate β.

Conclusion

Reject H_0 "The sample data supports the claim that [state H_a in words]."

Fail to reject H_0 "The sample data does not provide sufficient evidence to support the claim that [state H_a in words]."
Alternatively you could say: "The sample data does not contradict the the claim that [state H_0 in words].")

Types of errors

Type I error :

- Occurs when you reject H_0 but H_0 is true.
- α is the maximum Type I error you will accept.
- The p-value is **actual** probability of a Type I error for the specific sample data.

Type II error :

- Occurs when you fail to reject H_0 but H_0 is false.
- β is the probability of a Type II error.

Power $= 1 - \beta$

Power is the probability you will reject H_0 if it is false and resultantly support a true alternative hypothesis. (This is what we generally hope to do in a study or experiment.)

- An experiment with a higher power will be more likely to support a true H_a.
- Good experiments should have a power of at least 0.8-0.9.
- Power increases if n increases.

If a study/experiment has a very small sample size, the power will be small (large β) and even if H_a is true you won't be likely to have statistical evidence to support it. You will likely fail to reject H_0 when it is false and make a Type II error. Conducting a study with low power will likely waste your time and money.

Organizing hypothesis tests
We can organize the hypothesis test we have discussed based on the number of samples and their purpose.

One sample hypothesis tests
H_a : includes $\neq, \leq, \geq$
proportion test $H_0 : p = p_0$
mean test **(**z**)** σ **known** $H_0 : \mu = \mu_0$
mean test (t**-test)** σ **unknown** $H_0 : \mu = \mu_0$

Two sample hypothesis tests
H_a : includes $\neq, \leq, \geq$.
proportion test $H_0 : p_1 = p_2$
mean test **(**z**)** σ **known** $H_0 : \mu_1 = \mu_2$
mean test (t**-test)** σ **unknown** $H_0 : \mu_1 = \mu_2$
Confidence intervals for two sample tests are for the **difference** in the two parameters (eg. $p_1 - p_2 = \Delta p$).

Many sample hypothesis tests
H_a : **at least one** . . . is different.
test of homogeneity $H_0 : p_1 = p_2 = p_3 = \cdots$
1-Way ANOVA $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots$

Other hypothesis tests
correlation test $H_0 : \rho = 0$, $H_a : \rho \neq 0$
test of independence H_0 : variable X and Y are independent. H_a : variable X and Y are dependent.

1.4 Linear regression and correlation

Linear correlation coefficient r

Deals with **paired quantitative** data.
r measures the strength of linear correlation.

1. $-1 \leq r \leq +1$
2. r is scale invariant.
3. r is invariant if x and y are interchanged.

4. r **only** measures the **strength** of **linear** relationships.

r^2 proportion of linear variation in y that is explained by x.

Finding r check for linear relationship, then find r and make sure it is significant (p-value $\leq \alpha$).

Linear Regression and Predictions

Requirements: (1) linear relationship (2) residuals are random (independent), have constant variance across x and are normally distributed.

1. Determine **predictor variable** (x) and **response variable** (y). (**paired quantitative** data)
2. Check for linear relationship: `plot(x,y)` (otherwise stop!)
3. Check for influential points.
4. Check for **statistically significant correlation**: `cor.test(x,y)`
   If a significant relation **does not exist** (p-val $\nleq \alpha$), the best predictor for **any** y is $\bar{y}$ (the mean of y). (STOP: linear regression is not useful.)
5. Find the regression equation.
6. Plot the line on the data.
7. Check the residuals. Should look random (no patterns) and have equal variance.
8. Make a prediction using the regression equation.
   Don't predict outside of sample data x values!

1.5 Selecting the statistical method

The following is a partial list of statistical methods that we have discussed:

1. mean
2. median
3. mode
4. standard deviation
5. z-score
6. percentile
7. coefficient of variation
8. scatter plot
9. histogram
10. boxplot
11. normal-quantile plot
12. confidence interval for mean
13. confidence interval for difference in means
14. confidence interval for proportion
15. confidence interval for difference in proportions
16. one sample mean test
17. two independent sample mean test
18. match pair test
19. one sample proportion test
20. two sample proportion test
21. test of homogeneity

22. test of independence
23. linear correlation coefficient & test
24. regression
25. 1-way ANOVA

For each situation below, which method is most applicable?

- If it's a hypothesis test, what are the null and alternative (state in words and mathematically).
- If it's a graphical method, describe what you would be looking for.

Question 6. A researcher wants to determine if the type of vehicle a person drives (car, truck, SUV, or motorcycle) has an effect on their mean level of testosterone.

Question 7. A drug company want's to determine if a new drug they have developed can increase IQ. They enroll 10 subjects in a study and measure their IQ before and after giving the drug.

Question 8. A friend asks you what the average house price in Tucson is.

Question 9. An economist conducts a random sample of people to estimate the difference in incomes between men and women.

Question 10. A global warming researcher wants to determine if there is a relationship between CO_2 and temperature.

1.6 Problems

Question 11. What does standard deviation represent?

Four students taking statistics decided to party until 4 am the night before the final. They ended up sleeping in and missing the final the next morning. They went to the professor's office later that afternoon and said that they had gotten a flat tire on the way to the exam and didn't have a spare. Being unable to quickly get help, they missed the final. The professor said he understood and told them to come back tomorrow to take the final.

When the students came the next day, the professor put them in separate rooms and gave them the following 100 point 2 question exam:
5 pts What does the CLT state about $\bar{x}$?
95 pts Which tire?

Question 12. What is the probability that they all guess the same tire? (CHECK: 0.0156)

Question 13. Would it be unusual for them to pass?

A class consists of 30 students with 10 females. If 3 students are randomly selected **without** replacement:

Question 14. Does the binomial distribution apply for this situation?

Question 15. What is the probability that all 3 are female? (Check 0.0296)

Question 16. What is the probability that at least 1 is female? (Check 0.719)

Men's weights are normally distributed with a mean of 172 lb and a standard deviation of 29 lb.

Question 17. If 1 man is randomly selected, find the probability that his weight is between 171 and 174 lbs. (Check: 0.0412). Would it be unusual for this to occur?

Question 18. If 64 men are randomly selected, find the probability that their mean weight is between 171 and 174 lbs. (Check: 0.318). Would it be unusual for this to occur?

Question 19. College officials wish to estimate the percentage of students who support a tuition increase to add more computer labs. How many randomly selected students must be surveyed in order to be 95% confident that the sample percentage has a margin of error of 3%? (Check $n = 1068$)

Given:

$$G = \frac{\sum (x_i - 1)^2 + 10}{\sum x_i}$$

Question 20. Find G if $x = \{-2, 3, 4\}$ (Check: 6.4)

Question 21. In a study of perception, 80 men are tested and 7 are found to have red/green color blindness. Construct a 90% confidence interval estimate of the proportion of all men with this type of color blindness. (Check: $\hat{p} = 0.0875$, $z_{\alpha/2} = 1.64$, $E = 0.052$)

A newly married couple does not want to get pregnant during their first year of marriage and decides to only use condoms as a contraceptive device. Studies[2] have shown that the probability of pregnancy per use of a condom is 0.15% (individual probability of pregnancy). During the first year the couple use a condom every time for a total of 162 times.

Question 22. Does the binomial distribution apply in this problem?

Question 23. Find the probability of 0 pregnancies over 1 year. (Check: 0.784)

[2] The data presented in this problem are approximate values derived from actual published data based on typical use effectiveness.

Question 24. Find the probability of 1 or more pregnancies over 1 year. (Check: 0.216)

Question 25. Would it be unusual to become pregnant during 1 year of use when only using this form of contraceptive?

Question 26. What would be the mean number of pregnancies that the couple should expect to have given the above information? (Check: 0.243)

Question 27. How many uses would be required for an average of 1 pregnancy? (Check: 667)

Question 28. Based on your above results, if you were the primary care physician for this couple, what would you tell them about their plan to prevent pregnancy? Would it be effective?