STA 291
Summer 2008

Lecture 13

Which Method to Choose?

- Questions to ask yourself:
  - Are there one or two populations?
  - Is this a test for the mean (quantitative data) or the proportion (qualitative data, something with a yes/no answer, might have "out of" in the wording)?
  - Is this a "large sample" or "small sample" case? If $n < 30$ and we are testing for the mean, then it is the small sample case. To make sure it is "large sample" for the proportion, check $np_0 > 5$ and $n(1-p_0) > 5$.
  - If there are two populations: Are the samples independent (coming from unrelated populations) or dependent (related populations, e.g. multiple biological readings from the same individual, testing performance before and after some treatment, …)
  - If the populations are dependent: Is this a matched pairs experiment? That is, can observations be naturally associated? Are we interested in the differences of the observations?
  - What is the alternative hypothesis? That is, is it one-sided or two-sided? If it is two-sided, it is possible to use a (1-a) CI to test at significance level a. If it is one-sided, care must be taken when computing the *p*-value.

How to identify matched samples

- If the 2 samples have different sample sizes, it cannot be matched.
- For equal sample sizes, is there a natural pair identifier? It is reasonable to have different sample sizes?
- Ask if it make sense to compare within the pair or across the pair.

Which Method to Choose?

- Describing One Population
  - Quantitative Response (Analyzing Means)
    - Large Sample (>=30): c.i. and test for the mean using z-scores
    - Small Sample(<30): c.i. and test for the mean using t-scores
  - Qualitative Response (Analyzing Proportions)
    - Large Sample: c.i. and test for proportion using z-scores
- Describing Two Populations
  - Quantitative Responses (Comparing Means)
    - Two Independent Samples
      - Large Samples: c.i. and test for the means using z-scores
      - Small Samples
        - » Normal populations (sample histograms appear symmetric): t-test
        - » Not normal populations: Wilcoxon-Mann-Whitney test
    - Two Dependent Samples (Matched Pairs)
      - Large Number of Pairs: c.i. and test for the means using z-scores
      - Small Number of Pairs: c.i. and test for the means using t-scores
  - Qualitative Response (Comparing Proportions)
    - Two Independent Samples
      - Large Sample: c.i. and test for proportions using z-scores

Large Sample or Small Sample

- In general, focus on how to do the calculations in the large sample case

  (using z-scores)
- In practice, usually the small sample calculations have to be done by computer

Which Method to Choose?

- The mean hourly wage in 2005 was $24.67 for a college graduate and $14.14 for a high school graduate.
- A sample of students has mean age equal to 22.1 years and the standard deviation is 4.3 years.

Which Method to Choose?

- A sample of college students is asked “How many times do you brush your teeth today”? Out of 60 students surveyed, 48 answered that they brushed their teeth more than once a day.
- A study compares the mean level of contributions to political campaigns in Kentucky by registered Democrats, and registered Republicans.

Which Method to Choose?

- Example: Compare new drug to placebo in a double-blind clinical trial
  - 24 patients
  - 12 randomly assigned to placebo
  - The other 12 receive the new drug
  - Research question: Is there a different effect of placebo and new drug on a “response” on, for example, number of seizures, blood parameter, health status, weight,...

Which Method to Choose?

- Example: Which of two suntan lotions (labeled X and Y) provides better protection against sunburn
  - 8 subjects expose their backs to the sun for a certain time, protected by suntan lotion
  - Possible design:
    - 4 subjects use lotion X
    - the other 4 subjects use lotion Y
  - Different design:
    - Each of the 8 subjects uses ***both*** suntan lotions at the same time
    - one lotion on the left side of the back, the other on the right side

Which Method to Choose?

- A sociologist took a random sample of 100 mall shoppers and asked a variety of questions.
- This survey was first conducted 3 years ago with another sample of 100 shoppers.
- In both surveys respondents were asked whether they spent at least 3 hours in malls during an average week (yes/no).

Multiple Choice Question

- *Which of the following statements are true?*
- **“95% confidence” means that**
  - 95% of the true population parameters are in the confidence interval
  - If we were to repeat the procedure of sampling and calculating confidence intervals from the same population, then 95% of the population parameters are going to be in every calculated interval
  - If we were to repeat the procedure of sampling and calculating confidence intervals from the same population, then 95% of the times our confidence interval will contain the true population parameter

Multiple Choice Question

In testing the hypothesis that the population mean is 100 against the alternative that it is larger than 100, the p-value is found to be 0.074, and the sample mean is 108. Which of the following statements regarding the p-value is true?

(i) The probability of observing a sample mean at least as large as 108 from a population whose mean is 100 is 0.074.

(ii) The probability of observing a sample mean smaller than 100 from a population whose mean is 108 is 0.074.

(iii) The probability that the population mean is larger than 100 is 0.074.

(iv) The probability that the population mean is 100 is 0.074.

Test vs. Confidence Interval

Assume that the *p*-value is equal to 0.043 for a test of the null hypothesis mu=2, with two-sided alternative.

What conclusion can we make about a 95% confidence interval for mu?

Questions from Old Exam

- Given that Z is a standard normal random variable, determine $P(-1.52 = Z = 0.85)$.

- For some value of z, the probability that a standard normal variable is above z is 0.8810. Determine the value of z.

Questions from Old Exam

- Assume that you want to test the hypothesis H0:mu=1000 against a two-sided alternative. For a sample of size *n*=100, the sample mean is 980 and the sample standard deviation is 200.
- Calculate the value of the test statistic and determine the *p*-value.
- Set up the rejection region for a test at significance level (alpha level) 0.01.

Questions from Old Exam

- When testing H0:mu=0 vs. H1:mu<0, the observed value of the z-score was found to be -2.15. The *p*-value for this test would be. *(Multiple Choice, Circle the correct answer)*

(i) 0.0158 (ii) 0.0316 (iii) 0.9684 (iv) 0.9842

- When testing H0:mu=0 vs. H1:mu<0, the observed value of the z-score was found to be 2.15. The *p*-value for this test would be. *(Multiple Choice, Circle the correct answer)*

(i) 0.0158 (ii) 0.0316 (iii) 0.9684 (iv) 0.9842

Questions from Old Exam

- Which of the following is true regarding the sampling distribution of the mean for a large sample size?
  (Multiple Choice, Circle the correct answer)

- It has the same shape, mean, and standard deviation as the population distribution.
- It has the same shape and mean as the population distribution, but has a smaller standard deviation.
- It has a normal distribution with the same mean as the population distribution but with a smaller standard deviation.
- It has a normal distribution with the same mean and standard deviation as the population.

Questions from Old Exam

- On UK's campus, students are debating whether Chipotle's or Qdoba's offers the better burritos. In order to get some statistical information related to this controversy, researchers calculate a 95% confidence interval for the difference in the mean protein content of a standard burrito. The confidence interval for the difference Burrito 2 – Burrito 1 (the names of the franchises are not disclosed here) was from -1.3 to 3.0. Which of the following statements can the researcher make?

(Multiple Choice, Circle the correct answer)

(i) Burrito 1 has less protein content than Burrito 2.

(ii) Burrito 1 has a greater protein content than Burrito 2.

(iii) No comparison can be made. We need to perform a hypothesis test.

(iv) Burrito 1 and Burrito 2 do not differ significantly in protein content.

Some Stuff To Know for Final:

- Testing Hypothesis
- Relationship between two-sided test and CI
- Type I and Type II errors
- You can count on a question on interpreting CI's
- Computing probabilities from normal distributions
- Finding mean, median, mode, IQR, outliers, etc.
- Creating a box plot or stem-and-leaf plot
- Terminology (population, sample, parameter, statistic, skewed, symmetric, …)
- Some sort of probability problem
- For older material, focus on items that were on the previous exams.

- The remainder of slides are various examples pasted from some various instructors' slides

Example: KY Kernel Jan 17,2007

- UK researcher developed a blood substitute. A total of 712 trauma patients in the study. 349 receive PolyHeme, 363 receive regular blood.
- 46 died in the PolyHeme group
- 35 died in the regular group.
- Is there any difference in the two rates of death?

65 randomly chosen subjects are given two bottles of shampoos: A and B. After a week, each state which one they prefer

	Subject 1	Subject 2	Subject 3
A	prefer		prefer
B		prefer	

...

- ...
- ...
- ...

125 randomly chosen subjects are given two bottles of pills: A and B. After a month on each pill, report LDL cholesterol level

	Subject 1	Subject 2	Subject 3
A	167	155	233
B	188	159	214

...

- ...
- ...
- ...

- After we take the difference for each subject, the problem becomes a one sample problem:

- If the preference has 50-50% chance?
- If the difference has mean zero?

Focus on the difference

- Prefer A; not prefer A; prefer A; ……
- For example 30 out of 65 prefer A

- In example two
- \- 21; - 4; 19; ……
- For example

$$\bar{X} = 11.4, \qquad s = 7.8$$

- For the first problem

$$z = \frac{0.4615 - 0.5}{\frac{\sqrt{0.5(1-0.5)}}{\sqrt{65}}} = -0.6202$$

- P-value is 2P(Z >|-0.6202 |) =2P(Z>0.6202)=0.535

- If you use the computer to do the problem, the p-value will be slightly different. Due to the fact that our calculation is only an approximation (use CLT).

- Computer is more accurate.
- For sample size very large the difference goes away. (For example 300 subjects out of 650 prefer A)

For the cholesterol problem

- $$H_0 : m = 0, \qquad H_1 : m \neq 0$$

$$z = \frac{11.4 - 0}{\frac{7.8}{\sqrt{125}}} = 16.34$$

- P-value = 2P(Z>|16.34|) = 0.00000000000......

- Actually I should be looking up the t-table with degrees of freedom 125-1 = 124

(since I used s in place of sigma)

- Using t-table applet I get also a tiny p-value (with more than 30 zero's)

- The formal conclusion: reject (overwhelmingly) the null hypothesis of difference =0, imply the difference is not zero. Apparently the difference is positive – the average difference is 11.4.
- This imply pill B has lower LDL values compared to pill A.

Example

- A large company has two shifts – a day shift and a night shift. Products produced must meet some specifications. The Manager of the company believe the products produced by the two shifts have different quality. To investigate this believe random samples of products were selected from each shift.
- Day shift: 188 of the 200 selected products meet specification.
- Night shift: 180 of the 200 selected products meet specification.

- Say each shift produce a large number of products, 50,000. So the sample of 200 is just a random sample out of this population.

Example -- analysis

- What type of population?
- One or two samples? (paired?)
- Type of alternative hypothesis?

- “proportion that meet the specification”
- Two shifts – two samples
- “believe ...... different”

comments

- Try to pick the day/night shift at a same day……so that machine condition are similar, temperature, humidity are similar, etc.
- Use workers that are similar in quality/training

- Idea: Reduce/matching other possible factors

- I could easily make this with sample sizes 200 and 300…… it is not paired

Testing hypothesis

- $H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$
- Calculating the test statistic

$$z_{obs} = \frac{0.94 - 0.90}{\sqrt{\frac{0.92(1-0.92)}{200} + \frac{0.92(1-0.92)}{200}}} = 1.474$$

- P-value = 2P(Z>1.474) = 0.14
- Not significant.

comments

- The calculations are approx. – rely on central limit theorem, need large samples.
- Here n=200, kind of large.
- If you use computer, you get p-value of 0.197. I will trust computer more, since it uses some adjustment to improve accuracy.

- For sample sizes in the range of 1000 the difference of formula and computer should negligible.

Confidence interval

- Before we do it we know the 95% confidence interval will contain 0.
- Why 0?

Well the null hypothesis is $p1-p2=0$

Confidence interval

- 188/200=0.94 vs. 180/200=0.90
- Sample sizes are large enough to use our formula
- 95% confidence interval of the difference

$$(0.94-0.90)\pm 1.96\sqrt{\frac{0.94(1-0.94)}{200}+\frac{0.9(1-0.9)}{200}}$$

$$0.04\pm 0.053=[-0.013,0.093]$$