

Comparing Measures of Center

MGMT 662: Integrative Research Project

July 31, 2008.

1

1.1 Goals

Goals of this class meeting

- Learn how to test for significant differences between means from two populations.
- Learn how to test for significant differences between proportions from two populations.
- Learn how to test for significant differences between medians from two populations.
- Learn the assumptions behind these tests.
- Learn how to decide what tests are appropriate for a given set of data.

2 Proportions

2.1 Single Sample Inferences

Testing a single sample proportion

- **Proportion:** Percentage of times some characteristic occurs.
 - Example: percentage of voters who support Barack Obama for president.
- Notation:
 - π : population proportion.
 - p : sample proportion.
- Sample proportion:

$$p = \frac{X}{n} = \frac{\text{Number of items in sample having characteristic}}{\text{sample size}}$$

Sampling Distribution

- Sample size must be sufficiently large.
- Population is *not normally distributed*.
- Central Limit Theorem:
 - Mean of the sampling distribution of p will equal π .
 - Standard deviation of the sampling distribution will be,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- Sampling distribution of p will be normal.

Practical Uses

- The standard deviation requires knowledge of π .
- Use p instead. Use t instead of z , $df = n - 1$.
- Some suggest: Use p instead and go ahead and use z as long as $np > 5$ and $n(1-p) > 5$.
- Conservative estimate: suppose $\pi = 0.5$. Use z .

Example

From July 14-22, 2008, Quinnipiac University surveyed 1094 likely Wisconsin voters and asked if the November election was today, which candidate they would choose. These were the results:

- 547 said Obama.
- 427 said McCain.
- 109 said unsure.
- 11 said other.

Run a statistical test that answers the following questions:

- Is there statistical evidence McCain will get less than 50% of the vote?
- Is there statistical evidence that Obama will get more than 50% of the vote?

2.2 Differences in Proportions

Differences in Proportions

- Assumptions:
 - Sample sizes for both groups are large (population is *not* normally distributed).
 - Samples are independent from one another.
- Central Limit Theorem applies:
 1. Mean of the sampling distribution = $\pi_1 - \pi_2$
 2. Standard deviation of the sampling distribution is:

$$\sqrt{\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

3. Sampling distribution is normally distributed.
- Notation: \bar{p} is weighted average of your sample proportions, $\bar{p} = (X_1 + X_2)/(n_1 + n_2)$.
 - Use t with degrees of freedom $n_1 + n_2 - 2$.

Example

From January 20-21, 2008, Quinnipiac University surveyed 533 registered voters and asked if the November election was today and you could only vote for McCain or Obama, which candidate they would choose. These were the results:

- 235 said Obama.
- 245 said McCain.
- 53 said unsure.

Run a statistical test that answers the following questions:

- Has Obama gained support since January?
- Has McCain gained support since January?
- Critique this analysis. Is there are reason not to believe your results? Is there a way to improve?

2.3 Related Samples

Related Samples

- McNemar Test - Z-test for differences in proportions with related samples.

Group 1	Group 2		Totals
	Yes	No	
Yes	A	B	A+B
No	C	D	C+D
Totals	A+C	B+D	n

- Make the following 2 x 2 table:

- Sample proportions: $p_1 = (A + B)/n$, $p_2 = (A + C)/n$.

- Population proportions:

- π_1 = proportion of population that would answer yes before some treatment.
- π_2 = proportion of population that would answer yes after some treatment.

McNemar Test

- Null hypothesis:

- $H_0 : \pi_1 - \pi_2 = 0$ (i.e. there is no difference in the population proportions).

- Alternative hypotheses:

- $H_a : \pi_1 - \pi_2 > 0$ (i.e. the population proportion is greater before the treatment).
- $H_a : \pi_1 - \pi_2 < 0$ (i.e. the population proportion is greater after the treatment).
- $H_a : \pi_1 - \pi_2 \neq 0$ (i.e. the population proportions differ due to the treatment).

- Given a large sample size, the following test statistic is approximately $N(0, 1)$ under H_0 :

$$Z = \frac{B - C}{\sqrt{B + C}}$$

Example

- Suppose a sample of 600 potential Verizon mobile phone customers are asked whether they would switch to Verizon once their existing contractual obligations expired. The potential customers then subject to a marketing campaign for Verizon, then asked again if they would switch.
- These are the fake results:

Before Marketing Campaign	After Marketing Campaign		Totals
	Yes	No	
Yes	306	12	
No	36	246	
Totals			600

Example - continued

Answer the following questions:

1. What is the proportion of customers that would switch before the marketing campaign?
2. What is the proportion of customers that would switch after the marketing campaign?
3. Is there statistical evidence the marketing campaign would be effective?
4. Why did you choose the test you did?
5. Would your conclusions be different if you incorrectly chose a different statistical procedure?

3 Single Sample

One Sample Inferences

- Suppose you are interested in whether a population mean is different than some specified value.
- Examples:
 - Is the average birth weight from mothers who had gestational diabetes greater than 7 pounds?
 - Is the average fill of beer at La Crosse Lager games different from 12 ounces?
 - Is the average undergraduate GPA of MBA students above the national average 2.9 (I totally made up that number)?

3.1 Sampling Distribution

Sampling Distribution of the Mean

- We already learned the Central Limit Theorem tells us that if:
 - Sample size is sufficiently large *or*...
 - Population has a normal distribution.
- Then,
 - The sampling distribution of \bar{x} will be normal.
 - The mean of the sampling distribution will equal the mean of the population (consistent):

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution will decrease with larger sample sizes, and is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Practical Uses

- The standard deviation of the sampling distribution has an unknown population parameter.
- Estimate σ with our sample standard deviation, s .
- This additional uncertainty causes the sampling distribution to widen, depending on the sample size.
- New distribution: Student t distribution.
 - Probabilities in distribution now depend on degrees of freedom.
 - $df = n - 1$.
 - As sample size increases, degrees of freedom increase, uncertainty about estimate for σ decreases, t-distribution looks more and more like the normal distribution.

Example

- Dataset: average pay for public school teachers and average public school spending per pupil for each state and the District of Columbia in 1985.
- Download dataset `eduspending.xls`.
- Conduct the following exercises:

- Show some descriptive statistics for teacher pay and expenditure per pupil.
 - Is there statistical evidence that teachers make less than \$25,000 per year?
 - Is there statistical evidence that expenditure per pupil is more than \$3,500?
- Do you see any weaknesses in our statistical analysis? Anything we did wrong?

Using SPSS

Opening the data:

1. Save *eduspending.xls* somewhere.
2. Open SPSS.
3. Click radio button **Open an existing data source**.
4. Double-click **More files...**
5. Change **Files of type:** to **Excel**.
6. Go find and double click *eduspending.xls*.
7. Click **continue**.

Descriptive Statistics

1. Click **Analyze** menu, select **Descriptive Statistics**, then select **Descriptives**.
2. Click on **Pay** and click right arrow button.
3. Click on **Spending** and click right arrow button.
4. Click **Options**.
 - (a) Check any options you find interesting.
 - (b) Click **OK**
5. Click **OK**

Test Hypotheses

1. Click **Analyze** menu, select **Compare Means**, then select **One-Sample T test**.
2. Select **Pay**, then click right arrow.
3. Enter in **Test Value** text box 30000.
4. Output tables show descriptive statistics for pay, and hypothesis test results.

3.2 Nonparametric Tests

Nonparametric Tests

- Why?
- Sign test: can use tests for proportions for testing the median.
 - For a null hypothesized population median θ .
 - Count how many observations are above the median.
 - Test whether that proportion is greater, less than, or not equal to 0.5.
 - For small sample sizes, use binomial distribution instead of normal distribution.

Example

- Dataset: 438 students in grades 4 through 6 were sampled from three school districts in Michigan. Students ranked from 1 (most important) to 5 (least important) how important grades, sports, being good looking, and having lots of money were to each of them.
- Open dataset `gradschools.xls`. Choose second worksheet, titled `Data`.
- Answer some of these questions:
 - Is the median importance for grades is greater than 3?
 - Is the median importance for money less than 3?

Using SPSS to conduct nonparametric tests for medians

1. Click **Analyze** menu, select **Nonparametric Tests**, then select **Binomial...**
2. Click on **Grades** (or a different variable of interest), then click on right arrow.
3. Click radio button for **Cut point** and enter 3 into text box.
4. Do you want the exact (binomial distribution) p-value or asymptotic distribution (normal distribution)?
 - (a) Exact: click on **Exact...**
 - (b) Click **Exact** radio button.
 - (c) Click **Continue**.
5. Click **OK**
 - Output table shows exact p-value and normal distribution p-value for a two-tailed test.
 - What is your conclusion?

4 Two Samples

Differences Between Two Means

- Hypothesis tests about $\mu_1 - \mu_2$.
- Need to make assumption about independence of samples.
- Need to make assumption about equality of variances.

4.1 Independent Samples

Independent Samples and Unequal Variances

- Assumptions:
 1. Both samples are sufficiently large *or*...
 2. Small sample comes from a normally distributed population.
 3. Samples are independent.
- Sampling Distribution of $\bar{x}_1 - \bar{x}_2$:
 - Centered around $\mu_1 - \mu_2$.
 - Complicated looking standard deviation of sampling distribution:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Sampling distribution is normally distributed.

Practical Uses

- Standard deviation of sampling distribution involves unknown population parameters.
- Estimate these variances with sample variances:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Due to this uncertainty \rightarrow use Student t distribution, instead of normal distribution.
- Degrees of freedom = smaller of $n_1 - 1$ and $n_2 - 1$.

Hypothesis Tests

- Set up null and alternative hypotheses:
 - $H_0 : \mu_1 - \mu_2 = 0$
 - $H_1 : \mu_1 - \mu_2 \neq 0$ (can also do one-tailed hypotheses)
- Test Statistic:
$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- Find p-value:
 - For a right tailed test: $P(t_{df=n_1+n_2-2} > t^*)$.
 - For a left tailed test: $P(t_{df=n_1+n_2-2} < t^*)$.
 - For a two tailed test: $2P(t_{df=n_1+n_2-2} > |t^*|)$.
- Reject if $\alpha < p - value$, otherwise fail to reject.

Example

- Dataset: average pay for public school teachers and average public school spending per pupil for each state and the District of Columbia in 1985.
- Test the following hypotheses:
 - Does spending per pupil differ in the North (region 1) and the South (region 2)?
 - Does teacher salary differ in the North and the West (region 3)?
- Do you see any weaknesses in our statistical analysis? Anything we did wrong?

Using SPSS to test differences in means

1. Click **Analyze** menu, select **Compare Means**, then select **Independent-Samples T test**.
2. Select **Pay** or **Spending**, depending on which you are currently interested in.
3. Click the right arrow that is just to the left of **Test Variables**.
4. Select **Area** and click on right arrow to the left of **Grouping Variable**.
5. You need to tell SPSS what your grouping variable means and what groups you are interested in:

- (a) Click on **Define Groups**
 - (b) Click radio button **Use specified values**.
 - (c) Enter in the appropriate numbers for Group 1 and Group 2 (i.e. if you want the North to be group 1, type a 1 in Group 1 text box, and if you want the West to be group 2, type a 3 in the Group 2 text box).
 - (d) Click **Continue**.
6. Click **OK!**
- The first output table shows some descriptive statistics for each group.
 - The next output table shows:
 - Statistical evidence about whether the variances are different.
 - Statistical evidence about whether the means are different.
 - Descriptive statistics about the difference in the means.
 - Confidence intervals for the difference in the means.

Independent Samples and Equal Variances

Assumptions:

1. Both samples are sufficiently large *or*...
2. Small sample comes from a normally distributed population.
3. Samples are independent.
4. Each population has the same variance, only the mean may be different.

Independent Samples and Equal Variances

- Sampling Distribution of $\bar{x}_1 - \bar{x}_2$:
 - Centered around $\mu_1 - \mu_2$.
 - Complicated looking standard deviation of sampling distribution:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}},$$

- Sampling distribution is normally distributed.
- σ^2 is the one and only variance (each population has the same variance).
- Estimate σ_2 with pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}$$

- What's the benefit? More degrees of freedom = $df = n_1 + n_2 - 2$.

4.2 Paired Samples

Dependent Samples - Paired Samples

- Use a **paired samples t-test** if instead the two samples have the same individuals before and after some treatment.
- Really simple: for each individual subtract the before treatment measure from the after treatment measure.
- Treat your new series as a single series.
- Conduct one-sample tests.

4.3 Nonparametric Tests

Nonparametric Tests for Differences in Medians

- Assumptions behind Mann-Whitney U test:
 - Samples are independent of one another.
 - The underlying distributions have the same shape (i.e. only the location of the distribution is different).
 - It has been argued that violating the second assumption does not severely change the sampling distribution of the Mann-Whitney U test.
- Null hypothesis: $\theta_1 = \theta_2$
- Alternative hypotheses:
 - $\theta_1 > \theta_2$ (right tailed)
 - $\theta_1 < \theta_2$ (left tailed)
 - $\theta_1 \neq \theta_2$ (two tailed)
- Statistical procedure involves ranking all the data points (regardless of group), then comparing the sums of the ranks for each group.

Example

- Dataset: 438 students in grades 4 through 6 were sampled from three school districts in Michigan. Students ranked from 1 (most important) to 5 (least important) how important grades, sports, being good looking, and having lots of money were to each of them.
- Open dataset `gradschools.xls`. Choose second worksheet, titled `Data`.
- Answer some of these questions:

- Is the median importance for grades different for grades 4 and 6?
- Is the median importance for money different for grades than 4 and 6?

Using SPSS to conduct nonparametric tests for medians

1. Click **Analyze** menu, select **Nonparametric Tests**, then select **2 Independent Samples...**
 2. Move **Money** (or whatever you are interested in) into **Test Variable List**
 3. Move **Grade** into **Grouping Variable**
 4. Define Groups: Group 1 is 4, Group 2 is 6.
 5. You can get exact p-values if absolutely necessary (takes more time).
 6. Click **OK**.
- Mann-Whitney test statistic can be huge; it's equal the smaller of the two sums of ranks.
 - The Significance is the p-value.
 - What is your conclusion?

5

Wrap-up

- We've talked about statistical significance.
- Looked into mathematical detail for computing the more simple statistical tests.
- We learned how to test for differences in:
 - proportions.
 - means.
 - medians.
- Right now: right up a small description of how you might apply some of today's techniques to your work.
 - Describe what the variables are. What are you measuring? How is it measured? What classification is the data? What is the domain for the variables?

- Very specifically, what question could you answer with today’s techniques?
- Describe and defend what statistical procedure you would use.
- If you really think, none of today’s techniques can possibly be applied, describe why, then answer the above questions about one of your classmate’s project.
- Be ready in a few minutes to possibly present to the class.