Plan for Today

• I’m Back!
• Issues from Pset 5
• Lecture 6 Review
• Discuss Readings
• Problem Set 6
Issues from Problem Set 4 and 5

• Just picked these up today

• Scores seem quite high, particularly with Pset 5

• Let me know if you have questions, want to review
Lecture 6 Review: Estimation

- Confidence Intervals
  - Two-sided
  - One-sided
  - Totals
  - Proportions
  - Differences

- Normal distribution vs. t-distribution
Confidence Intervals

• What is a confidence interval? Why do we use them?
  • A confidence interval is an interval that has a given probability of containing the true population mean or proportion
  • Confidence intervals allow us to better describe the accuracy of a point estimate of the sample mean.
  • With a confidence interval, we can say how likely it is that a particular range of values contains the actual population mean.

• What is the first step in constructing a confidence interval?
  • Specify a confidence level

• What are some standard confidence levels?
  • Usually 95%, sometimes 90% or 99%
Confidence Intervals

- Confidence intervals for two-sided intervals have the form:
  \[(\text{pop. mean}) = (\text{sample mean}) \pm (\text{multiple})(\text{SE})\]
  \[= (\text{lower CL, upper CL})\]

- How do we determine the multiple? Give an answer in words and in an excel formula.
  - The multiple is based on the confidence level that we choose
  - The probability outside the interval is alpha; \(\text{Alpha} = 1 - \text{CL}\)
  - If the population standard error is known, the multiple is \(Z\)
    \[Z = \text{NORMSINV}(\alpha/2)\]
Confidence Intervals

• How would you find the multiple (Z value) for a 95% confidence interval (assuming the population standard error is known)?
  • NORM.S.INV(0.025) = 1.960
  • NORM.S.INV(0.975) = -1.960
Sample Standard Deviation

- We usually don’t know the population standard deviation (sigma). Why is this a problem?
  - Standard deviation is not a robust measure of variability
  - Standard deviation is sensitive to outliers

- What happens if SE is calculated using a low value of s? Why would this outcome be bad?
  - SE can be under-estimated
  - If you under-estimate the SE, your confidence interval will be overly optimistic
Sample Standard Deviation

- What can we do when we know the sample standard deviation (s), but not the population standard deviation (sigma)?
  - Use the t distribution (t) instead of the normal distribution (Z)

\[
\begin{align*}
\text{Instead of:} & \quad Z &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} & \text{normal distribution} \\
\text{Use:} & \quad t &= \frac{\bar{x} - \mu}{s/\sqrt{n}} & \text{t distribution with (n - 1) degrees of freedom}
\end{align*}
\]
t Distribution

• How does the t distribution differ from the normal distribution?
  • The degrees of freedom parameter (n-1) defines the shape of the t distribution
  • The t distribution is more spread out than the normal distribution (lower peak and thicker tails)

• What is the excel formula for finding the multiple using the t distribution?
  • T.INV(alpha/2, df)
t Distribution

• Rank the following from largest to smallest:
  • 95% confidence level Z statistic
  • 95% confidence level t statistic (n=10)
  • 95% confidence level t statistic (n=80)
  • 99% confidence level t statistic (n=10)

• Answer:
  • 99% confidence level t statistic (n=10)
    • Being 99% sure requires a large multiple (wide confidence interval)
  • 95% confidence level t statistic (n=10)
  • 95% confidence level t statistic (n=80)
  • 95% confidence level Z statistic
    • Larger n gives you more information, lets you have a smaller multiple for you confidence interval, given the same confidence level (think of Z as a t-statistic with n=infinity)
Two-sided vs. One-sided Confidence Intervals

- When do you use two-sided confidence intervals?
  - When you’re concerned with both the lower and upper limits for the mean
- When do you use one-sided confidence intervals? Provide some examples.
  - If you’re only interested in the upper or lower limit
  - Upper only: Regulations limited radiation exposure (e.g. for astronauts)
  - Lower only: Minimum reliability of a machine, particularly for health and safety uses
One-sided Confidence Intervals

- Confidence intervals for one-sided intervals have the form:
  \[(\text{pop. mean}) < (\text{sample mean}) + (\text{multiple}) \times (\text{SE})\]
  \[(\text{pop. mean}) > (\text{sample mean}) - (\text{multiple}) \times (\text{SE})\]

- How do we determine the multiple? Give an answer in words, and give the excel formula.
  - The multiple is based on the confidence level that we choose
  - The probability outside the interval is alpha; Alpha = 1 – CL
  - If the population standard deviation is known, the multiple is Z
    \[Z = \text{NORMSINV}(alpha)\]
  - If sigma is not known, the multiple is t
    \[t = \text{T.INV}(alpha, df)\]
Confidence Interval for a Total

• How do you calculate the confidence interval for a total given the mean and standard error? Answer both in words and give the equations.
  • Multiply the average, x-bar, by the total number of observations, N
  • Multiply the SE by the total number of observations, N
  • Confidence Interval = N*x-bar +/- N*SE

• How do you calculate the confidence interval for a total given the proportion?
  • Multiply the prop, p-hat, by the total number of observations, N
  • Multiply the SE of p-hat, by the total number of observation, N
  • Confidence Interval = N*p-hat +/- N*SE
Confidence Intervals with Proportions

• Confidence Intervals for proportions have the form:
  \[(\text{pop. Proportion}) = (\text{sample prop.}) \pm (\text{multiple}) \times (\text{SE})\]

• Where the standard error is:
  \[\text{SE}(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\]

• Should we use \(t\) or \(Z\) when using proportions? Why?
  • Use \(Z\) for proportions
  • This is possible because \(SE\) is insensitive to differences between the population proportion and the sample proportion.
Difference of Means

• Assuming that the samples are drawn randomly (i.e. $x_1$ and $x_2$ are independent random variables), then:

\[
\mu_Y = \mu_{x_1} - \mu_{x_2}
\]

\[
\sigma_Y = \sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}
\]

\[
SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_{x_1}^2}{n_1} + \frac{\sigma_{x_2}^2}{n_2}} \approx \sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}
\]

• Use t distribution with $df = n_1 + n_2 - 2$
Difference of Means – Pooled Std. Dev.

• IF:
  • $s_1$ and $s_2$ are significantly different, AND
  • Either $n_1$ OR $n_2$ is small (<30), AND
  • There is no reason to believe $\sigma_1$ does not equal $\sigma_2$
• Then use the pooled standard deviation ($s_p$)

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

\[
SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

• Use t distribution with $df = n_1 + n_2 - 2$
Difference of Proportions

- Assuming that the samples are drawn randomly, then:
  - Difference in sample proportions = \( \hat{p}_1 - \hat{p}_2 \)
  - Difference in standard deviations =

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2} \approx \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}
\]
Normal Distribution or t-Distribution?

For each of the following, should you use the normal distribution (Z) or the t-distribution (t)?

- Sample size is 1000, population standard deviation (sigma) is known
  - Normal distribution (Z)
- Sample size is 100, population standard deviation is unknown
  - t-distribution (t)
- Sample size is 80, sample proportion is known
  - Normal distribution (Z)
- We’re looking at the difference between means, and the population standard deviation is unknown
  - t-distribution (t)
- We’re looking at the difference between proportions
  - Normal distribution (Z)
Equations…
Confidence Level Equations

• Confidence Interval for two-sided intervals:
  (pop. mean) = (sample mean) +/- (multiple)*(SE)
  = (lower CL, upper CL)
  (pop. Proportion) = (sample prop.) +/- (multiple)*(SE)

• Confidence Interval for one-sided intervals
  (pop. mean) < (sample mean) + (multiple)*(SE)
  (pop. mean) > (sample mean) – (multiple)*(SE)

• Confidence Interval for a Total
  - If T=Nx, then T-hat = N*x-bar  <and> SE(T-hat)=N*SE(x-bar)
  - If T=N(p-hat), then T-hat = N*p-hat <and> SE(T-hat)=N*SE(t-hat)
  - Multiply the average by the total number of observations in the sample
Sample Mean and Standard Error

- Standard error for mean
  \[ SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \]

- Standard error for proportion
  \[ SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

- Difference in means and proportion
  \[ \mu_Y = \mu_{X_1} - \mu_{X_2} \]
  \[ \sigma_Y = \sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2} \]

- Standard deviation for difference in proportion
  \[ \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2}{n_1} + \frac{\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2}{n_2}} \]

- Standard error for difference in means
  \[ SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_{\bar{x}_1}^2}{n_1} + \frac{\sigma_{\bar{x}_2}^2}{n_2}} \approx \sqrt{\frac{s_{\bar{x}_1}^2}{n_1} + \frac{s_{\bar{x}_2}^2}{n_2}} \]

- Pooled standard error
  \[ SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_{\bar{x}_1}^2}{n_1} + \frac{\sigma_{\bar{x}_2}^2}{n_2}} \approx \sqrt{\frac{s_{\bar{x}_1}^2}{n_1} + \frac{s_{\bar{x}_2}^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]
Excel Formulas

<table>
<thead>
<tr>
<th></th>
<th>Multiple of SE / Critical Values of Z or t</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 tail</td>
<td>2 tail</td>
</tr>
<tr>
<td>Z</td>
<td>(-\text{NORM.S.INV}(\alpha))</td>
<td>(-\text{NORM.S.INV}(\alpha/2))</td>
</tr>
<tr>
<td>t</td>
<td>(-\text{T.INV}(\alpha,df))</td>
<td>(-\text{T.INV}(\alpha/2,df))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P(&lt;Z), P(&lt;t)</td>
</tr>
<tr>
<td>Z</td>
<td>\text{NORM.DIST}(x,\mu,\sigma,1)</td>
</tr>
<tr>
<td></td>
<td>\text{NORM.S.DIST}(Z,1)</td>
</tr>
<tr>
<td>t</td>
<td>\text{T.DIST}(t,df,1)</td>
</tr>
<tr>
<td></td>
<td>\text{T.DIST.RT}(t,df)</td>
</tr>
</tbody>
</table>
Readings

Chapter 7: Confidence Intervals and the Normal Distribution (starts on pg. 180)

“Excel has become the standard platform for quantitative analysis. Carliberg has become a world-class guide for Excel users wanting to do quantitative analysis. The combination makes Statistical Analysis: Microsoft Excel 2010 a must-have addition to the library of those who want to get the job done and done right.”

—Gene V Glass, Regents’ Professor Emeritus, Arizona State University

STATISTICAL ANALYSIS:
Microsoft® Excel 2010

Conrad Carlberg
Problem Set 6

- The spreadsheet UMCPsalaries.xlsx contains the annual salaries of all tenured/tenure-track professors at the University in the fall of 1997. The dataset also contains other information about these professors, such as rank, department, and whether they were awarded a PhD (see "code" tab). Use this dataset to explore the difference between the salaries of male and female professors.
  - Calculate the mean salary of male professors and construct a 95% confidence interval for this mean. (Treat all male professors as a sample. Note that you can quickly calculate the sample mean and sample standard deviation using a pivot table.)
  - Do the same for female professors.
  - Calculate the mean salary difference between male and female professors, and construct a 95% confidence interval for the salary difference.
  - What do you conclude about the salaries of male and female professors at UMCP? Is this evidence of gender discrimination? What other variables might legitimately account for some or all of this difference?
Problem Set 6

• The spreadsheet welfare.xlsx contains data from an experimental program in Arkansas (and similar programs in Baltimore and Virginia) designed to test incentives for welfare recipients to become employed, such as a partial or full loss of benefits unless recipients participated in a job search/work experience program. Use this dataset to answer the following questions about the income of participants in the Arkansas welfare-to-work experiment.
  • The variable "earn10" gives the earnings of welfare recipients in the 10th quarter (2.5 years) after random assignment to the control (resgp=1) and experimental (resgp=2) groups. What proportion of the control group had no income in the 10th quarter? Construct a 95-percent confidence interval for the proportion with no income in the 10th quarter. (Note that you can quickly determine this using a pivot table, with "state" in the report filter and set to 1 (Arkansas); "resgp" in the column area; "earn10" in the row labels; and "count of earn10" in the values area.)
  • Do the same for the experimental group.
  • Construct a 95-percent confidence interval for the difference between the control and experimental groups in the proportion with non-zero income in the 10th quarter. What do you conclude about the effect of the experimental program?
  • Another way to measure the effect of the experimental program is to examine total earnings in the final year of the program. Describe how you would use Excel to construct a 95% confidence interval for the difference between the control and experimental groups in the total earnings in the last year.
  • EXTRA CREDIT: Construct this confidence interval.
Questions?