UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: Ivo Dinov,
Asst. Prof. of Statistics and Neurology

Teaching Assistants:
Brandi Shanata & Tiffany Head

University of California, Los Angeles, Fall 2007
http://www.stat.ucla.edu/~dinov/courses_students.html

Sample Size Calculations & Confidence Intervals for Proportions

Planning a Study to Estimate μ

- It is important before you begin collecting data to consider whether the estimates will be sufficiently precise.
- Two factors to consider:
  - the population variability of Y
  - sample size

Planning a Study to Estimate μ

- First: In certain situations the variability of Y should not be controlled for (response in a medical study to treatment). However, in most studies it is important to reduce the variability of Y, by holding extraneous conditions as constant as possible.
  - For example: study of breast cancer might want to examine only women

Planning a Study to Estimate μ

- Second: Once the experiment is planned to reduce the variability of Y as much as possible, we consider the sample size.
  - For example: how many women should we sample to achieve the desired precision for our estimate?
- RECALL: $SE = \frac{s}{\sqrt{n}}$

Planning a Study to Estimate μ

- To decide on a proper value of n, we must specify what value of SE is desirable and have a guess of s.
  - For SE we need to ask what value would we tolerate?
  - For s we could use information from a pilot study or previous research

$$\text{Desired } SE = \frac{\text{Guessed } s}{\sqrt{n}}$$

Planning a Study to Estimate μ

Example: Reindeer (Cont')

$\bar{y}$ = 54.78

s = 8.83

SE = 0.874

Suppose we would like to estimate the sample size necessary for next year's round-up to keep SE $\leq$ 0.6

$$\frac{8.83}{\sqrt{n}} \leq 0.60$$

$$14.72 \leq \sqrt{n}$$

$$216.58 \leq n \approx 217 \textit{ reindeer}$$

Can't have 0.6 of a reindeer, so we round (**ALWAYS** round up on sample size calculations) to n = 217 reindeer.

Planning a Study to Estimate μ

- What happens to n as the desired precision gets smaller?

Example: Reindeer (cont') Suppose we would like to estimate the sample size necessary for next year's round-up to keep SE $\leq$ 0.3

$$0.30 \geq \frac{8.83}{\sqrt{n}}$$

$$n \geq 866.32 \approx 867 \textit{ reindeer}$$

- When we double the precision (ie. cut SE in half) it requires 4 times as many reindeer.
  - This is the result of the $\sqrt{\ }$

Decisions About SE

- How do we make the decision of what SE we will tolerate is the estimation of μ
  - RECALL: $\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
    - the $\pm$ part is called the margin of error and is equivalent to $t(df)_{0.025}$ * SE for a 95% confidence interval

$$\underbrace{\qquad}_{-t(df)_{0.025}SE} \quad \bar{y} \quad \underbrace{\qquad}_{+t(df)_{0.025}SE}$$

    - If we scan the 0.025 (or 95%) column of the t table the t multipliers are roughly equal to 2.

$$t(df)_{0.025}SE \approx 2SE$$

Decisions About SE

- So then for example, maybe we reason that we want our estimate to be within $\mu \pm$ 1.2 with 95% confidence
  - Using the logic from the previous slide thinking of the span of the CI, suppose a total span of 2.4 or $\pm$ 1.2 is desired,

$$\underbrace{\qquad}_{-1.2} \quad \bar{y} \quad \underbrace{\qquad}_{+1.2}$$

then SE would need to be $\leq$ 0.60

$$t(df)_{0.025}SE \approx 2SE$$

$$2SE = 1.2$$

$$SE = 0.6$$

Conditions for Validity of Estimation Methods

- We have to be careful when making estimations
  - computers make it easy
  - interpretations are valid only under certain conditions

Conditions of validity of the SE formula

- For $\bar{y}$ to be an estimate of μ, we must have sampled randomly from the population
  - If not the inference is questionable/biased
- The validity of SE also requires:
  - The population is large when compared to the sample size
    - rare that this is a problem
    - sample size can be as much as 5% of the population without seriously inflating SE.
  - Observations must be independent of each other
    - we want the n observations to give n independent pieces of information about the population.

Conditions of validity of the SE formula

- *Definition:* A hierarchical structure exists when observations are nested within the sampling units
  - this is a common problem in the sciences

Example: Measure the pulse of 10 patients 3 times each.

- We don't have 30 pieces of independent information.
  - One possible naïve solution: we could use each persons average

Conditions of validity of a CI for μ

- Data must be from a random sample and observations must be independent of each other
  - If the data is biased, the sampling distribution concepts on which the CI method is based do not hold
  - knowing the average of a biased sample does not provide information about μ

Conditions of validity of a CI for μ

- We also need to consider the shape of the data for Student's T distribution:
  - If Y is normally distributed then Student's T is exactly valid
  - If Y is approximately normal then Student's T is approximately valid
  - If Y is not normal then Student's T is approximately valid only if n is large (CLT)
    - How large? Really depends on severity of non-normality, however our rule of thumb is $n \geq 30$
  - Page 202 has a nice summary of these conditions
  - **NOTE:** If sampling distribution cannot be considered normal Student's T will not hold.

Verifications of Conditions

- In practice these conditions are often assumptions, but it is important to check to make sure they are reasonable
  - Scrutinize study design for:
    - random sampling
    - possible bias
    - non-independent observations
  - Population Normal?
    - previous experience with other similar data
    - histogram/normal probability plot
    - increase sample size
    - try a transformation and analyze on the transformed scale

CI for a Population Proportion

- So far we have discussed a confidence interval using quantitative data
- There is also a CI for a dichotomous categorical variable when the parameter of interest is a population proportion

$\hat{p}$ is the sample proportion
p is the population proportion

CI for a Population Proportion

- When the sample size is large, the sampling distribution of $\hat{p}$ is approximately normal
  - Related to the CLT
- When the sample size is small, the normal approximation may be inadequate
  - To accommodate this we will modify $\hat{p}$ slightly

CI for a Population Proportion

- The adjustment we are going to make to $\hat{p}$ is to use $\tilde{p}$ instead

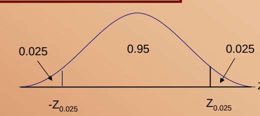
$$\hat{p} = \frac{y}{n} \longrightarrow \tilde{p} = \frac{y + 0.5\left(z^2_{\alpha/2}\right)}{n + \left(z^2_{\alpha/2}\right)}$$

- Relax and remember that the formula for $\hat{p}$ was:

$$\hat{p} = \frac{y}{n}$$

CI for a Population Proportion

- So what is the $z^2_{\alpha/2}$ bit?



- RECALL: In chapter 4, z_α was the cut point of the upper part of the standard normal distribution for a given α
- Now we want $z_{\alpha/2}$ because we are calculating a confidence interval and need to account for both sides of the distribution
  - So in the distribution above α would be 0.05, which corresponds to a 95% confidence interval

CI for a Population Proportion

- The standard error of $\tilde{p}$ also needs a slight modification

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \longrightarrow SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z^2_{\alpha/2}}}$$

- A sample value $\tilde{p}$ is typically within $\pm 2SE_{\tilde{p}}$

CI for a Population Proportion

- Before we define the formula for a CI for p let's remember the formula for a ***CI(μ)***

RECALL: $\bar{y} \pm t(df)_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

Where 100(1 - α) is the desired confidence

- If we pick this apart we are really saying that a CI(μ) is:

the estimate of μ ± (an appropriate multiplier) x (SE)

CI for a Population Proportion

- Incorporate that logic and we get:

$$\tilde{p} \pm z_{\alpha/2}\left(SE_{\tilde{p}}\right)$$

Where 100(1 - α) is the desired confidence

This time we will use a z multiplier instead of a t multiplier

Application to Data

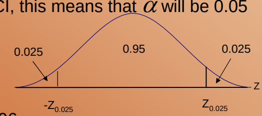
Example: Suppose a researcher is interested in studying the effect of aspirin in **reducing heart attacks**. He randomly recruits **500** subjects with evidence of early heart disease and has them take one aspirin daily for two years. At the end of the two years he finds that during the study only **17** subjects had a heart attack.

Calculate a **95% confidence interval** for the true proportion of subjects with early heart disease that have a heart attack while taking aspirin daily.

Application to Data

Example: Heart Attacks (cont')

- First, we need to find $z_{\alpha/2}$
  - because this is a 95% CI, this means that α will be 0.05 and $z_{\alpha/2}$ will be $z_{0.025}$



  - in this case $z_{\alpha/2} = 1.96$

Application to Data

- Next, solve for $\tilde{p}$

The Text rounds this to $\frac{y+2}{n+4}$

$$\tilde{p} = \frac{y + 0.5\left(z^2_{\alpha/2}\right)}{n + z^2_{\alpha/2}} = \frac{y + 0.5\left(z^2_{0.025}\right)}{n + z^2_{0.025}} = \frac{y + 0.5\left(1.96^2\right)}{n + 1.96^2} = \frac{y + 1.92}{n + 3.84}$$

  - that's just the formula for $\tilde{p}$, now we actually have to find $\tilde{p}$

$$\tilde{p} = \frac{17 + 1.92}{500 + 3.84} = 0.038$$

Application to Data

- Next, solve for $SE_{\tilde{p}}$

$$SE_{\tilde{p}} = \sqrt{\frac{(0.038)(0.962)}{500 + 3.84}} = 0.0085$$

- Finally the 95% CI for p

$$\tilde{p} \pm z_{\alpha/2}\left(SE_{\tilde{p}}\right) = 0.038 \pm 1.96(0.0085)$$
$$= 0.038 \pm 0.0167 = (0.0213,\ 0.0547)$$

Application to Data

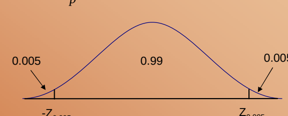
- What is our interpretation of this interval?

CONCLUSION: We are highly confident, at the 0.05 level (95% confidence), that the true proportion of subjects with early heart disease who have a heart attack after taking aspirin daily is between 0.0213 and 0.0547.

  - Is this meaningful?

Practice

- Calculate $\tilde{p}$ and $SE_{\tilde{p}}$ for a 99% confidence interval



So $z_{0.005}$ is 2.58

$$\tilde{p} = \frac{y + 0.5\left(z^2_{\alpha/2}\right)}{n + z^2_{\alpha/2}} = \frac{y + 0.5\left(z^2_{0.005}\right)}{n + z^2_{0.005}} = \frac{y + 0.5\left(2.58^2\right)}{n + 2.58^2} = \frac{y + 3.33}{n + 6.66}$$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z^2_{\alpha/2}}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 2.58^2}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 6.66}}$$

Practice

- This is a lot of work!
- Consider the following shortcuts:
  - The value of $z_{\alpha/2}$ can be carried through for all three formulas

$$\tilde{p} = \frac{y + 0.5\left(z^2_{\alpha/2}\right)}{n + \left(z^2_{\alpha/2}\right)} \qquad SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z^2_{\alpha/2}}} \qquad \tilde{p} \pm z_{\alpha/2}\left(SE_{\tilde{p}}\right)$$

    - just don't forget to square it in $\tilde{p}$ and $SE_{\tilde{p}}$
  - RECALL: The t distribution approaches a z distribution when df $=\infty$
    - this means that at the bottom of the t table there are several t multipliers that can be substituted for z (use the df = row)
    - CAUTION: this will only work for certain levels of α. If not found on the t table you must go back and solve with the z table!

Planning a Study to Estimate p

- We talked about finding the sample size necessary to ensure for quantitative data. This method depended on:
  - Desired SE
  - Guessed s

$$Desired\ SE_{\tilde{p}} = \sqrt{\frac{(Guessed\ \tilde{p})(1 - Guessed\ \tilde{p})}{n + z^2_{\alpha/2}}}$$

- For the proportions we use a similar idea:
  - where a guess for $\tilde{p}$ can be made on previous research or in ignorance.

Planning a Study to Estimate p

Example: Heart Attacks (cont')

How many subjects are needed if researchers want SE ≤ 0.005 for a 95% CI, and have guess based on previous research that $\tilde{p}$ would be 0.04

$$0.005 \geq \sqrt{\frac{(0.04)(0.96)}{n + 1.96^2}} = \sqrt{\frac{(0.04)(0.96)}{n + 3.84}}$$

$$0.005^2 \geq \frac{(0.04)(0.96)}{n + 3.84}$$

$$n + 3.84 \geq 1536$$

$$n \geq 1533.16 \approx 1534 \text{ subjects}$$

Example – 6.12

6.12. **Six healthy three year-old female Suffolk** sheep were injected with the antibiotic Gentamicin, at a dosage of 10 *mg/kg* body weight. Their blood serum concentration (*μg/mLi*) of Gentamicin 1.5 hours after injection were as follows: **33, 26, 34, 31, 23, 25**.

For these data, the mean is 28.7 and the standard deviation is 4.6.

(a) **Construct a 95% confidence interval for the population mean μ.**

There are five degrees of freedom. 28.7 ± 2.571 × 4.6/sqrt(6), or (23.9, 33.5). y-bar = 28.7; s = 4.5898; SE = 4.5898/sqrt[6] = 1.8738 =(approx) 1.9 micrograms/liter. 28.7 +/- (2.571)(1.8738) = (23.9,33.5) or 23.9 < mu < 33.5

(b) Define in words the population mean.

The population mean μ is the mean blood serum concentration in μg/ml of Gentamicin 1.5 hours after injection at a dosage of 10mg/kg body weight in healthy three-year-old female Suffolk sheep. The value of mu is unknown. However, it does exist and, in words, mu = mean blood serum concentration of Gentamicin (1.5 hours after injection of 10 mg/kg body weight) in healthy three-year-old female Suffolk sheep.

(c) The fact that the 95% confidence interval for μ contains nearly all the observations – will this be generally true?

The fact that, in this case, 95% confidence interval for μ contains nearly all the observations is mainly due to the small sample size. For much larger samples, confidence in the location of μ is much more concentrated and the interval will be much tighter.

Example – 6.12

StatisticalBarChartDemo2: http://socr.ucla.edu/htmls/SOCR_Charts.html