

UCLA STAT 13
**Introduction to Statistical Methods for the
Life and Health Sciences**

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

Teaching Assistants:
Fred Phoa, Kirsten Johnson, Ming Zheng & Matilda Hsieh

University of California, Los Angeles, Fall 2005
http://www.stat.ucla.edu/~dinov/courses_students.html

Slide 1 Stat 13, UCLA, Ivo Dinov

**Sample Size Calculations
&
Confidence Intervals for Proportions**

Slide 2 Stat 13, UCLA, Ivo Dinov

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

Slide 3 Stat 13, UCLA, Ivo Dinov

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

Slide 4 Stat 13, UCLA, Ivo Dinov

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: $SE = \frac{s}{\sqrt{n}}$

Slide 5 Stat 13, UCLA, Ivo Dinov

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

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

Slide 6 Stat 13, UCLA, Ivo Dinov

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 $SE \leq 0.6$

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

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

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

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

Slide 7

Stat 13, UCLA, Jon Dinger

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 $SE \leq 0.3$

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

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

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

Slide 8

Stat 13, UCLA, Jon Dinger

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

$$\underbrace{-t(df)_{0.025} SE}_{-1.2} \quad \bar{y} \quad \underbrace{+t(df)_{0.025} SE}_{+1.2}$$

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

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

Slide 9

Stat 13, UCLA, Jon Dinger

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 ± 1.2 is desired,

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

then SE would need to be ≤ 0.60

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

$$2SE = 1.2$$

$$SE = 0.6$$

Slide 10

Stat 13, UCLA, Jon Dinger

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

Slide 11

Stat 13, UCLA, Jon Dinger

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.

Slide 12

Stat 13, UCLA, Jon Dinger

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

Slide 13 Stat 13, UCLA, Jon Dinger

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

Slide 14 Stat 13, UCLA, Jon Dinger

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.

Slide 15 Stat 13, UCLA, Jon Dinger

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

Slide 16 Stat 13, UCLA, Jon Dinger

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

Slide 17 Stat 13, UCLA, Jon Dinger

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

Slide 18 Stat 13, UCLA, Jon Dinger

CI for a Population Proportion

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

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

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

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

Slide 19 Stat 13, UCLA, Jon Dineen

CI for a Population Proportion

- So what is the $z_{\alpha/2}^2$ 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

Slide 20 Stat 13, UCLA, Jon Dineen

CI for a Population Proportion

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

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

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

Slide 21 Stat 13, UCLA, Jon Dineen

CI for a Population Proportion

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

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

Where $100(1 - \alpha)$ is the desired confidence

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

the estimate of $\mu \pm$ (an appropriate multiplier) \times (SE)

Slide 22 Stat 13, UCLA, Jon Dineen

CI for a Population Proportion

- Incorporate that logic and we get:

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

Where $100(1 - \alpha)$ is the desired confidence

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

Slide 23 Stat 13, UCLA, Jon Dineen

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.

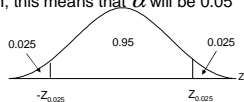
Slide 24 Stat 13, UCLA, Jon Dineen

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 $z_{\alpha/2} = 1.96$

Slide 25 Stat 13, UCLA, Jon Dineen

Application to Data

- Next, solve for \tilde{p}

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

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

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

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

Slide 26 Stat 13, UCLA, Jon Dineen

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

$$\begin{aligned} \tilde{p} \pm z_{\alpha/2}(SE_{\tilde{p}}) &= 0.038 \pm 1.96(0.0085) \\ &= 0.038 \pm 0.0167 = (0.0213, 0.0547) \end{aligned}$$

Slide 27 Stat 13, UCLA, Jon Dineen

Application to Data

- What is our interpretation of this interval?

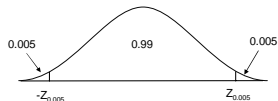
CONCLUSION: We are highly confident, at the 0.05 level (95% confidence), that the true proportion of subjects with early heart disease who have a heart attack after taking aspirin daily is between 0.0213 and 0.0547.

- Is this meaningful?

Slide 28 Stat 13, UCLA, Jon Dineen

Practice

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



So $z_{0.005}$ is 2.58

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

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

Slide 29 Stat 13, UCLA, Jon Dineen

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

$$\tilde{p} = \frac{y + 0.5\left(\frac{z_{\alpha/2}^2}{n}\right)}{n + \frac{z_{\alpha/2}^2}{n}} \quad SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + \frac{z_{\alpha/2}^2}{n}}} \quad \tilde{p} \pm z_{\alpha/2}(SE_{\tilde{p}})$$
 - 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 $df = \infty$ 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!

Slide 30 Stat 13, UCLA, Jon Dineen

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

- For the proportions we use a similar idea:

- where a guess for \tilde{p} can be made on previous research or in ignorance.

Slide 31 Stat 13, UCLA, Ivo Dinov

Planning a Study to Estimate p

Example: Heart Attacks (cont')

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

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

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

$$n + 3.84 \geq 1536$$

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

Slide 32 Stat 13, UCLA, Ivo Dinov

Planning in Ignorance

- If you have no idea what \tilde{p} will be, you can use $\tilde{p} = 0.5$ as the most conservative estimate
 - (largest value of $(\tilde{p})(1 - \tilde{p})$)

Example: Heart Attacks (cont')

How many subjects are needed if researchers wanted $SE \leq 0.005$ with 95% confidence and had no idea about \tilde{p} ?

$$0.005 \geq \sqrt{\frac{(0.5)(0.5)}{n + 3.84}}$$

$$n + 3.84 \geq 10000$$

$$n \geq 9996.16 \approx 9997 \text{ subjects}$$

Slide 33 Stat 13, UCLA, Ivo Dinov