

## Designed Experimental Study

- Manipulates the subjects somehow
- Can be used to prove causation
- Subjects randomly divided into groups
- Examples:
  - Does a coupon attached to a catalogue make recipients more likely to order?
  - Does a new medicine reduce the frequency of headaches?

## Observational Study

- Measures qualities of subjects without manipulating them
- Cannot be used to prove causation—only that the variables are related.
- Cannot be randomly assigned to groups
- Examples
  - Whether or not smoking has an effect on heart disease (can't assign groups)
  - Are higher SAT scores positively correlated with higher college GPAs?

## Types of Observational Studies

- Cross-Sectional
  - Takes a poll/survey at the current time
  - Who are you planning to vote for right now?
- Prospective
  - Studies something now, and then will revisit it in the future with the same subjects
  - Blood pressure of subjects now, versus a week later after being on a new medicine
- Retrospective Study
  - Looks into the history of the subjects
  - Comparing brain cancer & no brain cancer and how frequently subjects of both groups used cell phones in the past
- Case-Control Study
  - Subjects with a quality of interest are compared with controls on an explanatory variable (a type of retrospective)
  - Same example above

## Bias in Surveys

- Sampling Bias
  - Using nonrandom samples
  - Surveying the first 20 people we meet in downtown Athens
- Undercoverage
  - Sampling from not enough areas, so we don't have a representative sample.
  - Surveying from only the Northeast, rather than all regions of the country
- Nonresponse
  - Some sampled subjects can't be reached, or neglect to answer/return the survey
  - Sending out 1000 surveys, but only getting 61 of them back
- Response Bias
  - Confusing or "rigged" question, or an incorrect answer the subject made
  - "Do you prefer **CHOCOLATE** ice cream, or some other flavor?"

## Designed Experiments

- Experimental Unit (subject)
  - The person/object that receives the treatment
- Treatment
  - A condition/drug/etc. applied to the subject
- Response Variable
  - The categorical/quantitative variable of interest. We believe it's affected by the explanatory variable
- Explanatory Variable
  - Variable we believe to influence the response

## Designed Experiment (HW 4.1-4.4)

- Does the presence of a cute puppy in a TV advert help increase sales? A company randomly divides the 50 states into two groups, 25 each. An advert featuring the puppy is aired in one group, and a similar advert without the puppy is shown in the other group.
- Response Variable:
- Explanatory Variable:
- Treatments:
- Experimental Units:

## Designed Experiment (HW 4.1-4.4)

- We are growing one group of plants under white light, and another under blue light. It is of interest to see if the color of light received causes the plants to grow taller. Fill in the answers.
- Response:
- Explanatory:
- Treatments:
- Experimental Units:

## Experimental Designs

- Completely Randomized
  - Experimental units are randomly assigned to treatments, and no overlap occurs.
  - Example: 10 people randomly assigned to take treatment A, and another, different group of 10 subjects take treatment B (no two take both treatments)
- Matched Pairs
  - Subjects are somehow matched before the experiment happens, for measuring differences between the two
  - Twins, or same person in two treatment groups. Subjects between groups are dependent somehow.

## Experimental Designs

- Cross-over Design
  - A type of matched pairs; when a subject receives both treatments at some point in the experiment
  - So, a matched pairs in which the same person is in both groups
  - Cookie lab
- Block: a set of matched experimental units / subjects
- Randomized Block Design
  - Using blocks, but randomly assigning the order in which each block receives the treatment
  - This reduces possible bias
  - Cookie lab again (order was random)

## Blinding

- Double Blind
  - Neither the subjects nor the researcher knows who is getting which treatment (this is preferred as bias is lowest here)
  - The key is only revealed afterwards
- Single Blind
  - Subjects don't know, but the researcher does
  - We have possible "researcher bias" here

## Experimental Designs (HW 4.1-4.4)

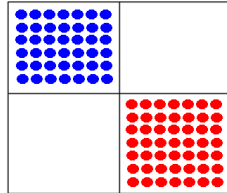
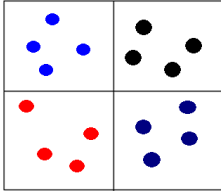
- Is there a connection between listening to classical music and reasoning skills?
- 36 students listened to a Mozart piece for ten minutes, then took a test. The same students, at another time, sat in silence for ten minutes, then took another test.
- The reasoning test scores were compared.
- Response variable:
- Levels of treatment (the two types):
- Is it matched pairs?
- Is it cross-over?

## Sampling Methods

- Simple Random Sampling
  - Each subject everywhere has an equally likely chance of being selected
  - Often done with a random number table
  - Choosing a company somewhere in the U.S.
- Systematic
  - Selecting every "k-th" subject
  - Surveying every 10<sup>th</sup> person we meet downtown
- Convenience
  - Individuals are easily found (e.g. internet surveys)
  - Often the "laziest" way, so less reliable answers

## Sampling Methods

- Stratified Sampling
  - Taking **some** subjects from **all possible groups**
- Cluster Sampling
  - Taking **all subjects** from **some possible groups**



## Sampling (HW 4.1-4.4)

- A researcher takes 3 possible classifications of companies, each of which contains 1000 businesses, and draws 100 random subjects from all three. What type of sampling is this?
- Suppose instead she draws 200 businesses at random from the whole population of companies. What type of sampling is this?
- The same researcher instead randomly selects 2 of the 3 possible classifications and then surveys all businesses in those groups. What type of sampling is this?
- Suppose instead she gets an alphabetical list of all these companies, starts with #4, and selects every 100<sup>th</sup> after that for her sample.

## Random Table (HW 4.1-4.4)

- A study will assign subjects numbered 1 – 8 into one of two groups, four in each. Use the table to decide who goes into the first group. Start with the top left, and answer in numeric order.

30494 17011  
22368 46573

## Notation

We use different letters for population parameters versus sample parameters.

$\mu$  = population mean       $\bar{x}$  = sample mean  
 $\sigma$  = population st. dev.       $s$  = sample st. dev.  
 $p$  = population proportion       $\hat{p}$  = sample proportion

Know the differences among these!

## Notation (HW 7.1-7.3)

- What symbol is used to denote the population mean?
- What symbol is used to estimate the population proportion?
- What symbol is used to describe the spread in one sample?

## Empirical Rule

- Quick Flashback to Test 1...
- The population mean is 55, and s.d. is 12. Find an interval within which about 95% of the population will fall.
- Lower Limit:  $55 - 2(12) = 31$
- Upper Limit:  $55 + 2(12) = 79$
- So our interval would be (31,79)
- Now, a similar idea for confidence intervals (but we use numbers like 1.96 instead of just 1, 2, or 3)

## Confidence Intervals

- Calculate the sample mean/proportion in your sample
- **Point Estimate** = sample mean/proportion
- Calculate the **width**, based on **level of confidence** and **standard error**
- You get a range of plausible values for the true population mean/proportion

## C.I. for Proportions

$$\left( \begin{array}{c} \text{point} \\ \text{estimate} \end{array} \right) \pm \left( \begin{array}{c} \text{confidence} \\ \text{level} \end{array} \right) \left( \begin{array}{c} \text{standard} \\ \text{error} \end{array} \right)$$

$$\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

margin of error = width of C.I.

$\hat{p}$  = point estimate  
z depends on confidence level

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \text{standard error}$$

$$z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \text{margin of error \& width of C.I.}$$

## Properties of a C.I.

- The sample proportion/mean is ALWAYS inside the confidence interval!
- In fact, it's always right in the center

$$\hat{p} \pm z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The population proportion/mean may or may not be inside the confidence interval

## Interpretation of a C.I.

- Any number falling within a confidence interval is a plausible value for the true population mean, while any number outside the interval is not.
- Example: a 98% interval is (60, 75)
  - With probability 98%, the population mean is somewhere between 60 and 75
  - We can conclude, with 98% confidence, that the true mean is above any number less than 60
  - Similarly, we can conclude that the true mean is below any number greater than 75
  - But we cannot rule out any numbers in this interval

## Interpretation of a C.I.

- The main interpretation of a 95% interval is as follows:
- We are 95% certain the population mean/proportion is somewhere inside the interval.
  - The population mean, while unknown, is fixed
  - What changes is the interval
- Warning: It is incorrect to say "The population mean is in the confidence interval 95% of the time."
  - This is because this statement implies that the population mean changes and sometimes happens to be in the interval
  - Incorrect because the population mean is fixed

## Another Way to Interpret A C.I.

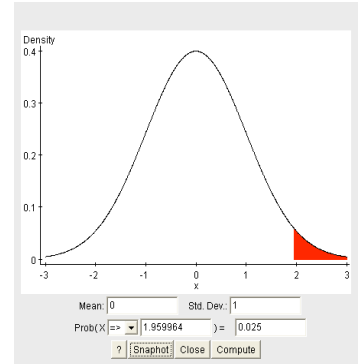
- A 95% C.I. also means that about 95% of all C.I.s constructed contain the true population proportion/mean, and about 5% do not
- A 99% C.I. means that about 99% of all C.I.s constructed contain the true population proportion/mean, and about 1% do not
- Example: 1000 intervals
  - At 95%, about 950 (maybe 940-960) contain the true proportion
  - At 99%, about 990 (maybe 985-995) contain the true proportion

## C.I. (HW 8.1-8.2)

- The annual salaries of 100 randomly selected people in Vancouver have a mean of \$49,000 and a margin of error of \$8000 with 95% confidence.
- Find the point estimate for this sample.
- Construct the 95% C.I.

## Determining z

- $z =$  level of confidence
- 95% C.I. :  $z = 1.96$
- 99% C.I. :  $z = 2.58$
- To get these numbers...
  - 95%, 5% is left over
  - Half of that is 2.5%
  - $P(z \geq ?) = .025$  in StatCrunch
- What about 85%?



## Proportions C.I. (HW 8.1-8.2)

- A random sample of 200 people were asked if they believed in the Loch Ness Monster. 160 said yes.
- Find a point estimate for the proportion of people who said yes.
- Find the standard error.
- Will we get a valid confidence interval?

## Proportions C.I. (HW 8.1-8.2)

- Find the margin of error for a 95% C.I.
- Construct the 95% C.I.
- Can we conclude that more than 70% of all people believe in the Loch Ness Monster?
- How about less than 88%?

## Proportions C.I. (HW 8.1-8.2)

- Same problem: now suppose another, different sample, also of size 200, is taken, and the confidence interval from this new sample is (.76, .80).
- If possible, find the population proportion and the new sample proportion.
- What is the new margin of error?
- Compare this interval with our earlier 95% interval, which was (.74456, .85544). Is this new interval more likely a 91% or a 98% interval?

## C.I. Properties

- Increasing level of confidence ( $z$ ) widens the interval
- Decreasing level of confidence ( $z$ ) shortens the interval

$$\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Intuition: narrowing your field for the true proportion means you're not as certain it really does fall inside the interval

## C.I. Properties

- Increasing the sample size shortens the C.I.
- Decreasing the sample size widens the C.I.
- This is because standard error decreases as n increases, so the margin of error (width) decreases as well.

$$\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Intuition: a larger sample size gives a more accurate estimate and allows you to zero in on the true proportion.

## Summary of C.I. Width Factors

- |                                 |                                 |
|---------------------------------|---------------------------------|
| <u>Confidence Level (z)</u>     | <u>Sample Size (n)</u>          |
| • As z increases, C.I. widens   | • As n increases, C.I. shortens |
| • As z decreases, C.I. shortens | • As n decreases, C.I. widens   |
- Assumptions for proportion C.I.:
    - Sample is randomly selected
    - $n\hat{p} \geq 15$
    - $n(1-\hat{p}) \geq 15$

## Proportions C.I. in StatCrunch

- Stat > Proportions > One Sample > With Summary
- # Successes and # Observations
- C.I., level, Standard-Wald, Calculate
- Tells you the C.I. and standard error
- Does not tell you the margin of error

## C.I. with Means

- Same general idea:  $\left( \begin{array}{c} \text{point} \\ \text{estimate} \end{array} \right) \pm \left( \begin{array}{c} \text{level of} \\ \text{confidence} \end{array} \right) \left( \begin{array}{c} \text{standard} \\ \text{error} \end{array} \right)$
- But we have a different formula:
 
$$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$$
- With proportions, use z
- With means, use t

## C.I. for Means

$$\left( \begin{array}{c} \text{point} \\ \text{estimate} \end{array} \right) \pm \left( \begin{array}{c} \text{confidence} \\ \text{level} \end{array} \right) \left( \begin{array}{c} \text{standard} \\ \text{error} \end{array} \right)$$

$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

margin of error = width of C.I.

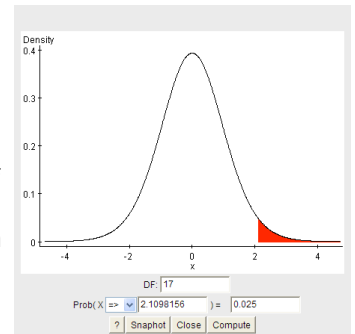
$\bar{x}$  = point estimate  
t depends on confidence level

$\frac{s}{\sqrt{n}}$  = standard error

$t \cdot \frac{s}{\sqrt{n}}$  = margin of error & width of C.I.

## The T Calculator

- Only new feature: degrees of freedom
- DF = n - 1
- Same strategy as before: with 95%...
- 5% left over, and half of that is 2.5%
- $P(X \geq ?) = .025$  with 18 observations (DF = 17)



## C.I. with Means

- T-values change as degrees of freedom change (unlike normal calculator)
- Degrees of freedom =  $n - 1$  **ALWAYS**
- Assumptions for doing C.I. for means:
  - Random sample
  - One of these two should be true:
    - Sampling from a normal population
    - $n > 30$

## C.I. with Means (HW 8.3-8.4)

- 480 people responded to a question on how many children they have:  
Mean = 3    Median = 2    S.D. = 1.78
- Find the point estimate for the sample.
- Find the standard error of the sample.
- Suppose the 95% C.I. was (2.84, 3.16). Choose an answer: With 95% confidence, the true mean lies (above, below, within) this interval.
- Is it plausible that the true population mean is 2? Why or why not?

## C.I. Means (HW 8.3-8.4)

- New Problem: The starting salaries of a sample (from a normal distribution) are \$34000, \$44000, \$54000.
- Mean = 44000, S.D. = 10000
- Find the point estimate.
- Find the standard error.
- How many degrees of freedom?

## C.I. Means (HW 8.3-8.4)

- Find the margin of error for a 95% C.I. ( $t = 4.30265$ )
- Find the 95% C.I.
- What effect will increasing the sample size have on the 95% interval?

## C.I. Means (StatCrunch)

- |  |   |
|--|---|
| • With data: enter data in one column          | • With Summary:                                   |
| • Stat > T-Statistics > One Sample > With Data | • Stat > T-Statistics > One Sample > With Summary |
| • Select var1, Next                            | • Enter sample mean, s.d., sample size            |
| • C.I., level of confidence, calculate         | • C.I., level of confidence, calculate            |

## Choosing Sample Size

- Idea: We have a given confidence level and a desired margin of error
- What sample size is needed to achieve that?
- Formula is different for proportions and means (see formula sheet)

## Sample Size Needed Formulas

$$n = \frac{\hat{p}(1-\hat{p})z^2}{m^2}$$

$$n = \frac{\sigma^2 z^2}{m^2}$$

$n$  = sample size needed

$n$  = sample size needed

$\hat{p}$  = "guess" on sample proportion

$\sigma$  = previous standard deviation

$z$  = z-score for confidence level

$z$  = z-score for confidence level

$m$  = desired margin of error

$m$  = desired margin of error

- What do we choose for the sample proportion?
  1. Proportion of a previous study
  2. If nothing is known,  $\hat{p} = .50$

## Sample Size (HW 8.3-8.4)

- We are interested in the proportion of students at a college that are affiliated with Greek life. We want to estimate it with probability .95 and within 0.06.
- What sample size do we need, if no previous study is known?
- All other things the same, if we instead estimate with probability .98, will we need more or fewer subjects?

## Sample Size (HW 8.3-8.4)

- Now suppose we're told that a previous study at Harvard said that 20% of students are involved with Greek life.
- Before doing any calculations, will the required sample size for a margin of error of .06 be larger or smaller than before?
- Find the new sample size, again with probability .95.
- What is the advantage to knowing a previous study proportion?

## Sample Size (HW 8.3-8.4)

- We are estimating the average number of acres on a farm to within 23, with probability 95%. In a previous study, the sample s.d. was 202 acres.
- Find the sample size needed.
- All other things being equal, if we wanted a smaller margin of error, will we need more or fewer subjects?

## Proportions Summary

### Assumptions for a Valid Confidence Interval

- Random sample
- We need  $n\hat{p} \geq 15$
- We need  $n(1-\hat{p}) \geq 15$

### Finding Sample Size

$$n = \frac{\hat{p}(1-\hat{p})z^2}{m^2}$$

### Confidence Interval

Point Estimate =  $\hat{p}$

Standard Error =  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Level of Confidence: use  $z$

Margin of Error =  $z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Lower Limit =  $\hat{p} - z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Upper Limit =  $\hat{p} + z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

## Means Summary

### Assumptions for a Valid Confidence Interval

- Random Sample
- One of these two:
  - Normal population, or...
  - $n > 30$

### Finding Sample Size

$$n = \frac{\sigma^2 z^2}{m^2}$$

### Confidence Intervals

Point Estimate =  $\bar{x}$

Standard Error =  $\frac{s}{\sqrt{n}}$

Level of confidence depends on  $t$

Margin of Error =  $t \cdot \frac{s}{\sqrt{n}}$

Lower Limit =  $\bar{x} - t \cdot \frac{s}{\sqrt{n}}$

Upper Limit =  $\bar{x} + t \cdot \frac{s}{\sqrt{n}}$