Day 4 - Statistics

Agenda

- Activity #1: Did I trap the Median?
  - Sampling
  - Intro to Confidence Intervals
  - Estimation of population parameters
- Activity#2: How Wet is the Earth?
  - Confidence intervals
  - Estimation of population parameters
- Activity #3: Using Dice to Introduce Sampling Distributions
  - Variability
  - Sampling

Confidence

- Raise hand if you are ‘familiar’ with what a confidence interval is

Confidence

- Keep your hand raised if you have encountered the concept of confidence interval...but you lack, let's say…
- …confidence…
- …in making sense of the idea

Did I trap the median?

- Our first activity is geared around 'typical' foot size.

- Rather than collect data from class – I have provided you with a list of 50 measures of foot length (in cm).

- This will play the role of our class data.

Did I trap the median?

- Additionally, we are going to imagine that the 50 students really represents a much larger data set…let's just say 50 'million'.

- If I truly provided you with a data set this large, it would be unrealistic or impractical (at least by hand) to find the median.

Did I trap the median?

- Discussion Questions:
- 1. What are the possible ways to find out what the median foot size of the entire class is?
- 2. What are the advantages and disadvantages of collecting a sample of only seven foot sizes and not those of the entire class to access knowledge about the median foot size in the entire class?
- 3. What are the advantages and disadvantages of using the sample median to estimate the population median?
- 4. Is there any advantage in constructing an interval estimate as opposed to a point estimate (the sample median) for the population median?

- 5. Is it possible to assign a reliability value to the interval estimate (assign a probability that the interval contains the median)?
- 6. What are the factors that may affect the length and the reliability of an interval estimate?

Did I trap the median?

- Complete Questions 1 – 8 on the handout alongside your group members
  - (each of you should have a different sample, but the process you follow should be the same)

- When most of us have completed Questions 1 through 8 , we can discuss these questions

- We will complete Questions 9 and 10 as a large group

Did I trap the Median? Discussion

- Questions 1 through 8

Did I trap the Median? Discussion

- 9. Based on the median of the entire class given by your instructor and the median of your particular seven randomly chosen classmates, calculate what proportion (percent) of box plots trap the median for the entire class. This is the reliability (confidence level) of using interval estimates from Q1 to Q3.

- 10. Think about what would happen if the sample size increased. Would the proportion of box plots that would trap the median increase or decrease? Why?

Did I Trap the Median: Discussion Questions

- 1. What do you think would happen to the sample interval (Q1, Q3) if the sample size increased from 7 to 15?

- 2. Do we have any ideas of how to construct interval estimates that are shorter than the interval (Q1, Q3)? Would a shorter interval necessarily change the level of reliability?

Discussion

- Definitions:
  - In statistics, a **confidence interval** (**CI**) is a type of interval estimate of a population parameter.

Discussion

- Definitions:
  - Confidence level does not describe any single sample.
  - The Confidence level value is represented by a percentage, so when we say, "we are 99% confident that the true value of the parameter is in our confidence interval", we express that 99% of the hypothetically observed confidence intervals will hold the true value of the parameter.

Discussion

- Definitions:
  - After any particular sample is taken, the population parameter is either in the interval realized or not; it is not a matter of chance.

Discussion

- Definitions:
  - Significance levels are very much related to the concept of a confidence interval
  - If we have a 95% confidence interval, we state that in 95% of all theoretical confidence intervals – we have 'captured' the parameter
  - However, this means that in 5% of all theoretical confidence intervals – we have not.

Discussion

- Definitions:
  - We will look at the topic of 'hypothesis testing' in a later activity (probably tomorrow)
  - Unlike most textbooks, I do not see confidence intervals and hypothesis testing as all that different
  - Hypothesis testing simply analyzes data from the perspective of significance level (essentially the inverse of confidence level)
  - Much like addition and subtraction are inverses, and can be viewed as 'different'…you only need addition to get by.

How wet is the earth?

- The guiding question for this activity is:
  - What proportion of the Earth's surface is covered in water?

- We will also explore questions such as:
  - How will sample proportions vary when investigating this question.
  - What proportion of 90% confidence intervals calculated would we expect to contain the true proportion?

How wet is the earth?

- Parameter:
  - A parameter is defined as something that does not vary, and represents a numeric measure for the entire population.
  - For this context, the parameter would correspond to the *actual proportion* of Earth that is covered in water.

How wet is the earth?

- Additionally for this activity, we will need to develop further meaning for a ‘confidence interval’

How wet is the earth?

- Essentially the idea of a confidence interval arises from the core idea of statistics: variability exists

- Sampling is necessary in order to estimate population parameters, because it is often unrealistic to sample the entire population.

How wet is the earth?

- Generic Confidence Interval:

  Estimate ± Confidence Level x Variability

- Estimate:
  - Typically we use the 'best guess' for the parameter, which is usually the mean (or median) of the sample.
- Confidence Level:
  - Refers to the z-score (or equivalent) corresponding to a certain confidence level. For example. If you had a confidence level number of 2, this refers to the percentage of data that falls within 2 standard deviations of the mean (which is roughly 95% of the data).
- Variability:
  - Often we use the standard deviation, or some estimate of the variance

How wet is the earth?

- Generic Confidence Interval:

  Estimate ± Confidence Level x Variability

- Estimate and Variability are typically fixed or determined by our sample.
- We use our sample mean and standard deviation for these values

How wet is the earth?

- Generic Confidence Interval:
  Estimate ± Confidence Level x Variability

- Confidence Level, however, is something we can choose.

- The 'empirical rule' states that in a normal distribution roughly:
  - 68% of data falls within 1 standard deviation of mean
  - 95% of data falls within 2 standard deviations of mean
  - 99.7% of data falls within 3 standard deviations of mean

How wet is the earth?

- Generic Confidence Interval:

Estimate ± Confidence Level x Variability

- Using the empirical rule, we could use the confidence-level values of 1, 2, and 3 for a roughly 68%, 95%, and 99.7% confidence interval.

How wet is the earth?

- Generic Confidence Interval:

Estimate ± Confidence Level x Variability

- Question: As the confidence level increases, what will happen to the confidence interval itself? Will it increase, decrease, or stay the same size? Why?

How wet is the earth?

- Generic Confidence Interval:

Estimate ± Confidence Level x Variability

- Common Confidence Levels values
  - 90% confidence: 1.645
  - 95% confidence: 1.96
  - 99% confidence: 2.576

How wet is the earth?

- Generic Confidence Interval:

Estimate ± Confidence Level x Variability

- Variability for sample proportion
- The letter ‘p’ is commonly used for sample proportions
- The formula for the standard deviation for proportions is given below (note that p must be a decimal value between 0 and 1)

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

How wet is the earth?

- We will begin by going to: http://www.geomidpoint.com/random
- Select rectangular region, and whole Earth
- Generate the number of points on Earth stated on your handout (this is either 20 or 40) and determine whether or not the red pins are in water.
  - Some of these points will be obviously in water or on land. For those that are questionable – zoom in to verify.
  - Obviously do your best to be accurate, but don't spend too much time debating it.

How wet is the earth?

- Let’s compare
- n=20

- n=40

- n > 40

How wet is the earth?

- Interpreting confidence intervals
  - How many n=20 intervals contained 71%?
  - How many n=40 intervals contained 71%?
  - How many n>40 intervals contained 71%?

Discussion

Discussion

- Definitions:
  - In statistics, a **confidence interval** (**CI**) is a type of interval estimate of a population parameter.

Discussion

- Definitions:
  - Confidence level does not describe any single sample.
  - The Confidence level value is represented by a percentage, so when we say, "we are 99% confident that the true value of the parameter is in our confidence interval", we express that 99% of the hypothetically observed confidence intervals will hold the true value of the parameter.

Discussion

- Definitions:
  - After any particular sample is taken, the population parameter is either in the interval realized or not; it is not a matter of chance.

Discussion

- Definitions:
  - Significance levels are very much related to the concept of a confidence interval
  - If we have a 95% confidence interval, we state that in 95% of all theoretical confidence intervals – we have 'captured' the parameter
  - However, this means that in 5% of all theoretical confidence intervals – we have not.

Discussion

- Definitions:
  - We will look at the topic of 'hypothesis testing' in a later activity (probably tomorrow)
  - Unlike most textbooks, I do not see confidence intervals and hypothesis testing as all that different
  - Hypothesis testing simply analyzes data from the perspective of significance level (essentially the inverse of confidence level)
  - Much like addition and subtraction are inverses, and can be viewed as 'different'…you only need addition to get by.