

Statistical Sampling in Ecology

Background: In Biology in general, and in Ecology in particular, researchers constantly evaluate how many subjects they need to measure for their experiments, where the number of subjects, n , is called **sample size**. In statistics, we want the mean of our sample, \bar{Y} , to be a reasonable estimate of the true, parametric mean, μ , of the population from which we sampled. If we sampled every subject in the population, then we'd know μ , but this is usually impractical if not impossible, given our limited money, space and time. Thus, we need to determine how large n needs to be for \bar{Y} to be a reliable estimate of μ . This will depend on the true parametric variance, σ^2 , of our source population, and our ability to estimate σ^2 from our sample data, as s^2 .

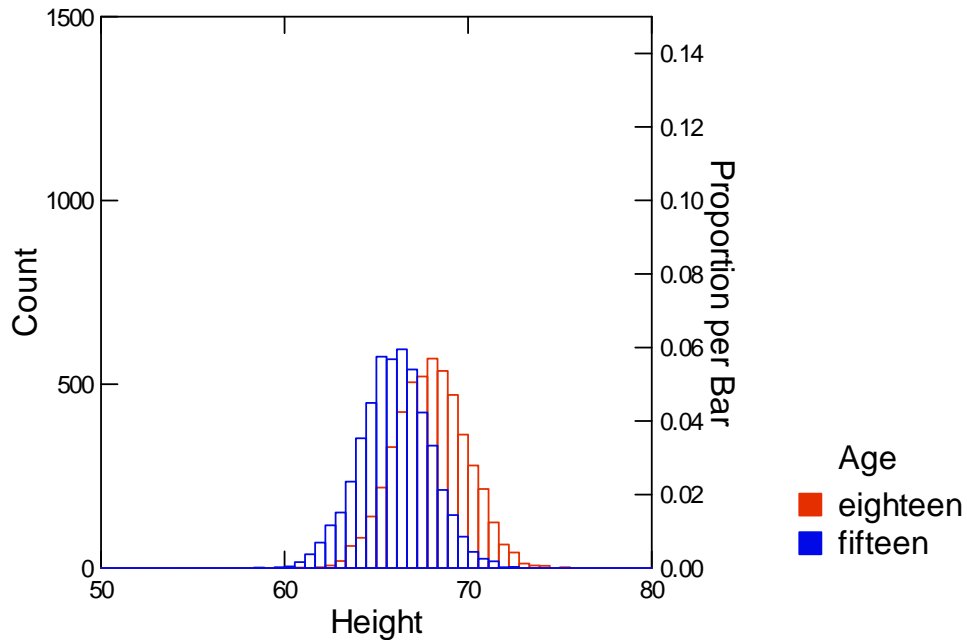
This exercise will introduce you to point (mean) and interval (variance, confidence intervals) estimation at a level covered in AP statistics class, or STA 200 or STA 291 here at UK. In addition, this exercise will introduce you to Hypothesis Testing by exploring how sample size affects our ability to discriminate the means of two distributions.

Learning Objectives: Our learning objectives are to understand how increasing our sample sizes leads to: 1. convergence between our sample means and the true parametric mean; 2. a reduction in the width of our confidence intervals about our sample means; and, 3. an increase in our ability to discriminate between two means from two different distributions.

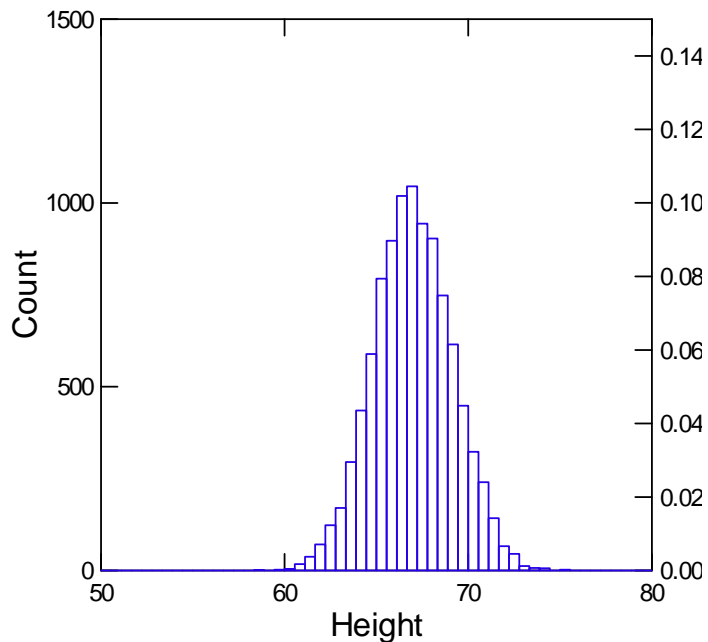
Our Data: We will be working with an Excel spreadsheet that contains human heights for two different age classes: 15 years old and 18 years old. 5000 students were sampled for each age class, and we will treat the mean and variance for each of our two distributions of 5000 people as our true parametric means and variances, which are in the table below, and the raw data are plotted as histograms on the next page.

	Age Class	
	15	18
Mean Height (in): μ	66.023	67.964
Variance: σ^2	3.581	3.732

The frequency distributions of height for 5000 15 year olds and 5000 18 year olds. Note that there is considerable overlap between the two distributions, and that the modes (peaks) are slightly different.



Below, the 2 distributions above are combined into one distribution. Your task is to figure out how many people you need to sample in each age group to conclude that 18 year olds are on average taller than 15 year olds.



Although we will be calculating our statistics with formulae in Excel, here are the relevant formulae.

Some Useful Sample Statistics and Their Formulae

- Mean or Average
 - The sum of the observations divided by the sample size
 - $$\bar{Y} = \frac{\sum_1^n Y_i}{n}$$
- Deviation (from the mean)
 - The difference between an observation and its mean
 - $Y_i - \bar{Y}$
- Variance
 - The sum of the squared deviations divided by n-1
 - $$s^2 = \frac{\sum_1^n (Y_i - \bar{Y})^2}{n - 1}$$
- Standard Deviation
 - The square root of the variance
 - $s = \sqrt{s^2}$
- Standard Error (of the mean)
 - The standard deviation divided by the square root of sample size
 - $$SE = s_{\bar{Y}} = \frac{s}{\sqrt{n}}$$
- 95% Confidence Interval of the Mean
 - An estimated interval about our sample mean for which we are 95% confident that it includes the true parametric mean.
 - $CI_{.95} = \bar{Y} \pm t_{.05/2, (n-1)} SE$
 - $CI_{.95} \approx \bar{Y} \pm 2SE$
 - A rule of thumb in hypothesis testing is that two means are likely to be significantly different if their 95% CIs do not overlap, though we usually perform a formal statistical test in addition, e.g. a t test.

Before we explore these human height data, we'll begin to investigate our learning objectives through a study of two online java applets that illustrate how sample size affects our estimates of means and confidence intervals. ***Answer the questions for each exercise.***

Exercises

1. Estimating the True Mean: The Law of Large Numbers: [Law of Large Numbers Applet](#)

- This applet plots on the Y-axis, the running average of a number of rolls of the dice, from $n=1$ to some upper limit you set, and on the X-axis, the number of rolls, n .
- In the "Rolls" box, type 100. Check the "Show mean" box, which plots a horizontal green line at 3.5, the true mean of 1 through 6.
- Hit the "Roll dice" button.
- To start over, hit the reset button. You'll have to reset the number of rolls, and show mean boxes. Repeat this experiment 10 times.

1. Overall, what happens to the running average of the dice as the number of rolls increases from 1 to 100?

2. What does this tell you in general about the relationship between sample means and true means as sample size increases?

2. Confidence Intervals: [Confidence Interval Applet](#)

- Under Method, choose “Means” and “z with sigma”
- For now you can leave μ and σ as is.
- n is your sample size which you will vary from 5, 10, 25 and 50.
- Set intervals to 100. Intervals is how many samples you take, and how many confidence intervals you will display simultaneously.
- Leave “conf level” at 0.95 (i.e. at 95%).
- Hitting the sample button will display the results of 100 samples of size n . The black dots are the means for each sample, and the horizontal green or red lines are the confidence intervals for each sample. The green intervals intersect the true mean; whereas the red samples do not.
- If you hit the sample button repeatedly, you’ll keep drawing 100 samples, and near the bottom, “Running total” tells you what percentage of your samples’ 95% CIs include the true mean.
- The Reset button will clear all your work so you can start over.

Run 10 samples of 100 intervals for each sample size: 5, 10, 25, 50

3. What percentage of the 95% CIs include the true mean for each sample size?

4. How does the width of the 95% CIs change as you increase sample size?

5. How does the average distance between the sample means and the true mean change as you increase sample size?

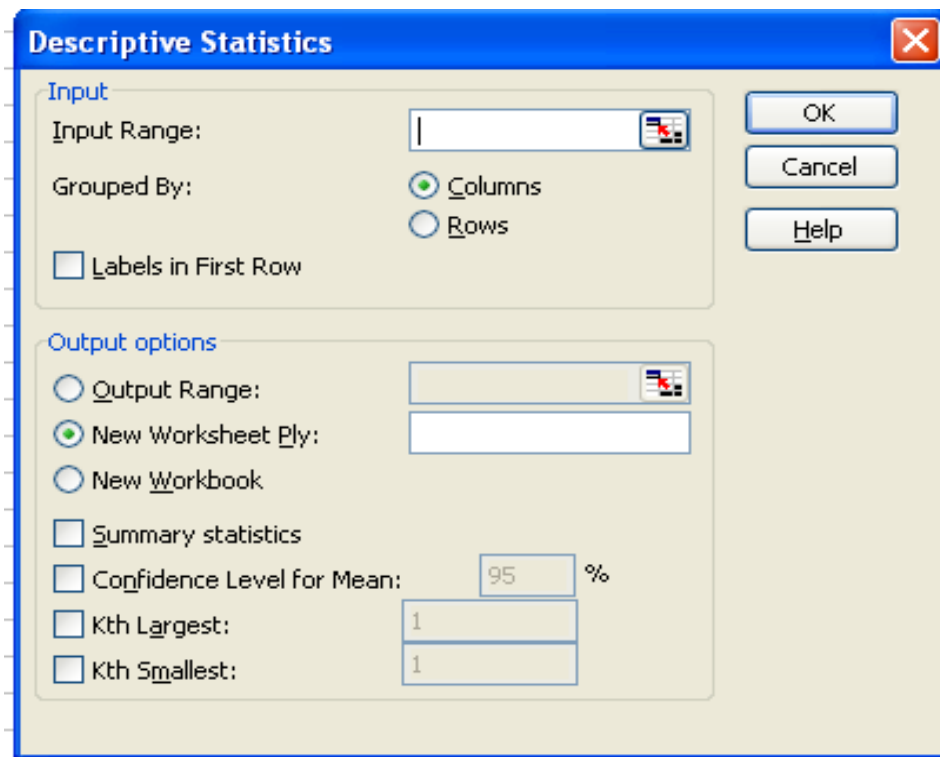
3. Human Height Data: For this exercise, work in pairs. Open sheet 1 of your Excel spreadsheet. Here's the link:

<http://darwin.uky.edu/~sargent/Bio325/StatisticalSamplingLabData.xls>

You will be selecting random samples of 5, 10, 25 and 50 individuals from each age class, and using the Excel Data Analysis Tool to calculate your sample statistics. Each student pair should pick a row at random in the spreadsheet (don't all pick the same rows at the top of the table; your TA can assign you regions of the table to explore), and starting with that row, work downward with sample sizes of 5, 10, 25 & 50.

To calculate your sample statistics..,

- open the Data Analysis Tool and select Descriptive Statistics.
- This will open a box that looks like this...



- Click on the red and black icon to the right of the Input Range Box, which opens another box that looks like this...



- Highlight the sample you want to analyze, then click the icon on the right side of the box.
- Check the Summary Statistics and Confidence Level Boxes, and then OK.
- Your output will be placed into a new worksheet and will look something like this (for the first 5 15 year olds)...

<i>Column1</i>	
Mean	67.80172
Standard Error	0.758672
Median	67.10487
Mode	#N/A
Standard Deviation	1.696442
Sample Variance	2.877915
Kurtosis	-1.11982
Skewness	0.504164
Range	4.26386
Minimum	65.86807
Maximum	70.13193
Sum	339.0086
Count	5
Confidence Level(95.0%)	2.106411

- You'll need to widen the first column to be able to read what's in it. The relevant statistics are in **Bold**.
- Count is your sample size, in this case 5.
- You will want to save the Mean and $CI_{.95}$ from each analysis you run, and enter them into the Table on the next page.
- To fill in the table, you will repeat this process 10 times for each sample size for each of the two age classes. If you enter the data into the Excel Output sheet, it may facilitate your answering the questions. Save a copy of this file and give it to your TA when you're done (emailing it is fine).

Exercise 3 Questions:

6. How does increasing sample size influence how close your sample means are to the true parametric mean for each age class?(Hint: what's the variance of your sample means for each sample size within each age class?)

7. How does increasing sample size affect the width of your 95% CIs ?(Hint: what's your average CI for each sample size within each age class?)

8. For each sample size column, how much overlap is there between the two age classes for their mean \pm CIs?(Hint: You may need to plot these to answer this question) How does increasing sample size affect this overlap?

9. For these Human Height data, what sample size for each age class would you recommend that would allow one to safely conclude that on average, 18 year olds are taller than 15 year olds? Why?

10. Why is it not necessary to measure every individual in a population in order to estimate its mean?

References:

The following, [Statistics Online Computational Resources](#), is an excellent resource at the level of high school AP statistics and STA 291 here at UK.