In Class Activity, Friday, September 16:

Prob 5.40[.](https://dl.dropboxusercontent.com/u/5098197/Exercises/HTML/SM2-Chapter-5-Problems/SM2-Chapter-5-Problems-A.html?access=ISMf12#540) In this activity, you are going to look at the sampling distribution and how it depends on the size of the sample. This will be done by simulating a sample drawn from a population with known properties. In particular you’ll be looking at a variable that is more or less like the distribution of human adult heights — **normally distributed** with a mean of 68 inches and a standard deviation of 3 inches.

Here’s one random sample of size n = 10 from this simulated population:

**rnorm**(10, mean=68, sd=3)

   [1] 62.842 71.095 62.357 68.896 67.494  
   [6] 67.233 69.865 71.664 69.241 70.581

These are the heights of a random sample of n = 10. The sampling distribution refers to some numerical description of such data, for example, the sample mean. Consider this sample mean the output of a single trial.

    mean( rnorm(10, mean=68, sd=3) )

  [1] 67.977

If you gave exactly this statement, it’s very likely that your result was different. That’s because you have a different random sample — rnorm generates random numbers. And if you repeat the statement, you’ll likely get a different value again, for instance:

    mean( rnorm(10, mean=68, sd=3) )

  [1] 66.098

Note that both of the sample means above differ somewhat from the population mean of 68.

The point of examining a sampling distribution is to be able to see the reliability of a random sample. To do this, you generate many trials — say, 1000 — and look at the distribution of the trials. For example, here’s how to look at the sampling distribution for the mean of 10 random cases from the population:

    s = do(1000)\*mean( rnorm(10, mean=68, sd=3) )

By examining the distribution of the values stored in s, you can see what the sampling distribution looks like.

Generate your own sample

* What is the mean of this distribution?
* What is the standard deviation of this distribution?
* What is the shape of this distribution?

Now modify your simulation to look at the sampling distribution for n = 1000.

* What is the mean of this distribution?
* What is the standard deviation of this distribution?
* What is the shape of this distribution?

Which of these two sample sizes, n = 10 or n = 1000, gave a sampling distribution that was more reliable? How might you measure the reliability?

The idea of a sampling distribution applies not just to means, but to any numerical description of a variable, to the coefficients on models, etc. Now modify your computer statements to examine the sampling distribution of the standard deviation rather than the mean. Use a sample size of n = 10. (Note: Read the previous sentence again. The statistic you are asked to calculate is the sample standard deviation, not the sample mean.)

* What is the mean of this distribution?
* What is the standard deviation of this distribution?
* What is the shape of this distribution?

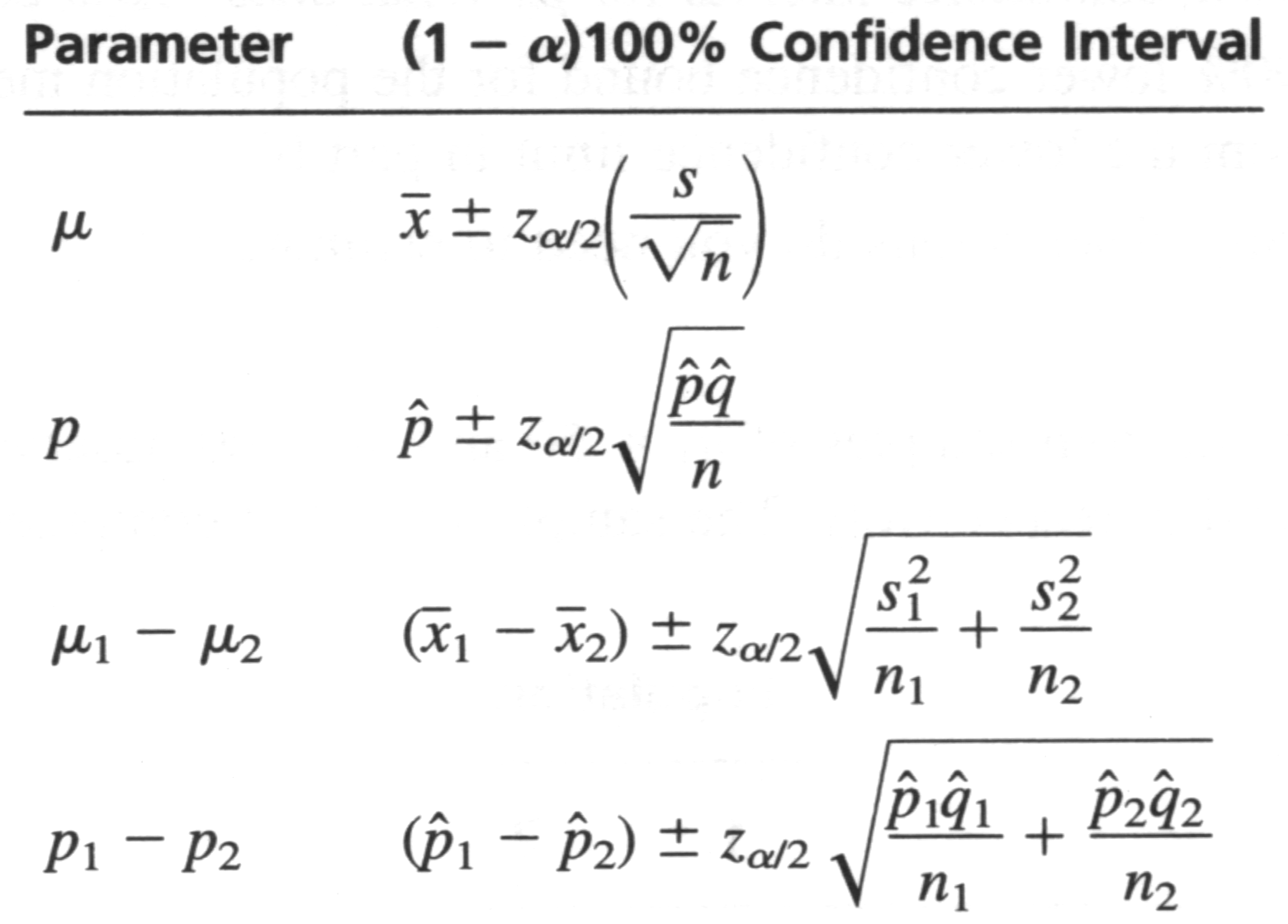
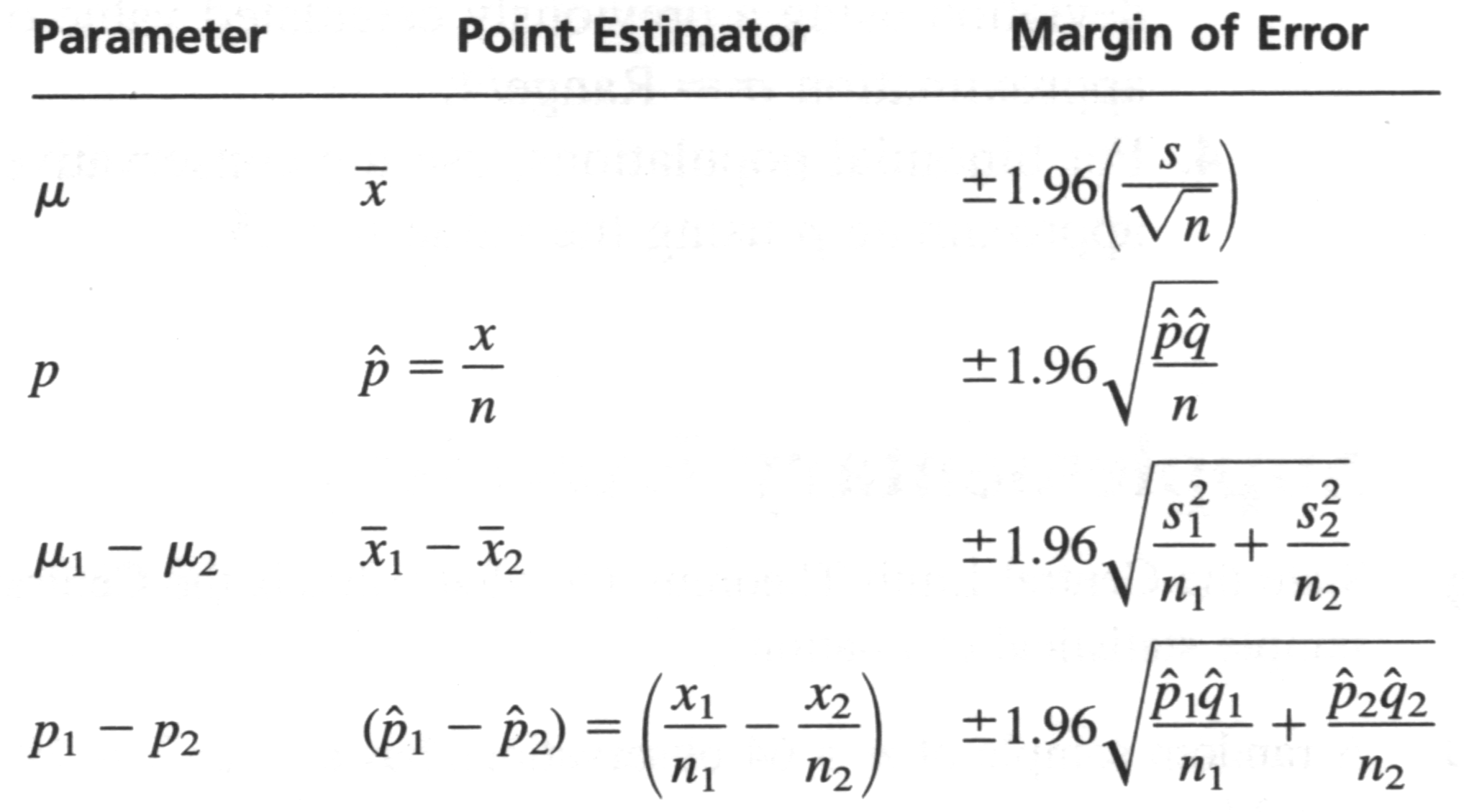
Repeat the above calculation of the distribution of the sample standard deviation with n = 1000.

* What is the mean of this distribution?
* What is the standard deviation of this distribution?
* What is the shape of this distribution?

For this simulation of heights, the population standard deviation was set to 3. You expect the result from a random sample to be close to the population parameter. Which of the two sample sizes, n = 10 or n = 1000 gives results that are closer to the population value?



|  |
| --- |
| **Point Estimate** |
| A **point estimate** of a population characteristic is a single number (hence the name “point” estimate) that is based on sample data and represents a plausible value of the characteristic. |



Conditions to use these approximations: n > 30 and data not skewed, or use t-distribution for n < 30 and normal data. For proportions, need np and n(1-p) > 5, as a rule of thumb.

5.01p: Using the Galton data, and the parametric approximation above, write down a confidence interval for the mean height. How does your confidence interval compare to the one from bootstrapping?

5.02p: Using the Galton data, and the parametric approximation above, write down a confidence interval for the difference in mean height between men and women. How does your confidence interval compare to the one from bootstrapping?

5.41: Using the Galton data, and the parametric approximation above, write down a confidence interval for the proportion of individuals with height > 70.

5.42: Using the Galton data, and the parametric approximation above, write down a confidence interval for the difference in proportions between men and women with height > 70.

Effect Size…



Suppose the SwimRecords data was only a pre-study. Use the standard deviation from that dataset to determine the n needed to get a 95% CI for the mean with margin of error < .5