Central Limit Theorem

The central limit theorem states that the sampling distribution of  is APPROXIMATELY a normal distribution with a mean of μ and variance of σ2/n when the sample size n is LARGE. This holds no matter what the probability distribution is for . In the course notes, we saw that n = 20 is works well for the GPA example.

Questions that we address in this lab:

* Does a smaller sample size work (e.g., n = 5 or n =10)?
* How will the approximation improve when n increases?
* Will the central limit theorem work well if the underlying probability distribution for Y is quite skewed?

Simulate observations

A tool used in statistics to answer questions like this is to simulate what a sample would look like from a population.

Here’s how we can use R to simulate observations from a population characterized by a normal distribution with μ = 2.8571 and σ2 = 0.4082.

> mu <- 2.8571

> sigma.sq <- 0.4082

> n <- 100000

> set.seed(2343)

> y <- rnorm(n = n, mean = mu, sd = sqrt(sigma.sq))

> head(y)

[1] 2.329982 3.037242 2.998464 2.869211 3.307891 2.274544

> hist(x = y, freq = FALSE, xlab = "y", col = NA, breaks =

25)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq)),

Added after making video:

Where does quant come from? Suppose Z is a standard normal random variable. One can find that P(Z < -1.96) = 0.025. For this example, I wanted to fill in the blank: P(Y < \_\_) = 0.025 instead. To find this missing value, I use the Z = (Y-μ)/σ relationship. Thus,

P(Z < -1.96) = P((Y-μ)/σ < -1.96)

 = P(Y < -1.96σ + μ).

Why did I choose -1.96? There was no particular reason. I could have chosen some other value and worked with it instead.

add = TRUE, col = "red")

> # Compare to mu and sigma

> mean(y)

[1] 2.85754

> var(y)

[1] 0.4082285

> # Check P(Y < \_\_\_)

> quant <- -1.96\*sqrt(sigma.sq) + mu

> quant

[1] 1.604846

> pnorm(q = quant, mean = mu, sd = sqrt(sigma.sq))

[1] 0.0249979

> sum(y < quant)/n

[1] 0.02498

> mean(y < quant)

[1] 0.02498



The actual normal distribution is plotted upon the histogram so that you can see the observations are what we would expect. Note that the y-axis is re-scaled from a regular histogram so that the normal distribution can be plotted.

GPA example

We will re-examine the GPA example in your class notes with different sample size n. The population is characterized by the following probability distribution



The population mean and variance are μ = 2.8571 and σ2 = 0.4082.

This distribution is based on a beta probability distribution with parameters a = 5 and b = 2 – see the textbook “Probability and Statistics for Engineers and Scientists” for more information. You are NOT required to know about the beta distribution, how to plot it, or how the observations are sampled in the upcoming discussion.

We simulate 1000 samples of size n = 10 from this population using the following code:

> n <- 10

> set.seed(8129)

> set1 <- 4\*matrix(data = rbeta(n = 1000\*n, shape1 = 5, shape2

 = 2), nrow = 1000, ncol = n)

> head(round(set1,2))

 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 3.41 2.79 2.65 1.87 3.21 2.99 2.96 3.83 3.28 2.15

[2,] 3.01 3.15 3.83 2.25 3.75 3.76 2.22 1.99 2.50 3.29

[3,] 2.90 2.23 2.95 3.17 3.25 3.85 3.59 3.37 3.57 3.11

[4,] 3.06 2.59 2.50 2.60 2.33 2.19 2.01 2.29 2.94 2.89

[5,] 2.38 2.74 3.48 3.25 3.24 3.34 3.17 2.79 2.25 3.39

[6,] 2.46 3.51 2.21 2.88 3.75 2.40 3.50 3.04 2.63 2.98

> ybar <- rowMeans(set1)

> head(round(ybar,2))

[1] 2.91 2.97 3.20 2.54 3.00 2.94

Now ybar contains 1000 sample means. We can use the following code to make a histogram of all 

> mu <- 2.8571

> sigma.sq <- 0.4082

> hist(x = ybar, main = "Histogram of sample means", xlab =

 expression(bar(y)), freq = FALSE)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq/n)),

 add = TRUE, col = "red")

> # Check P(Ybar < \_\_\_)

> quant <- -1.96\*sqrt(sigma.sq/n) + mu

> quant

[1] 2.461102

> pnorm(q = quant, mean = mu, sd = sqrt(sigma.sq/n))

[1] 0.0249979

> sum(ybar < quant)/length(ybar)

[1] 0.021

> mean(ybar < quant)

[1] 0.021



The central limit theorem works well for n = 10!

What happens for n = 5? I use the same code but change n.



There may be too few bars in the histogram to see the shape of the distribution. Also, the top of the normal distribution curve is cut off. Thus, we modify our code to increase the number of bars and the limit for y-axis.

> hist(x = ybar, ylim = c(0, 1.5), main = "Histogram of sample

 means", xlab = expression(bar(y)), freq = FALSE, breaks

 = 15, col = NA)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq/n)),

 add = TRUE, col = "red")



With n = 5, the histogram of  starts to reflect some skewness from the population. As we expect, the central limit theorem does not work as well for smaller n.

Question: What happens if n = 1? What will the histogram look like?

What would happen with a much more skewed distribution?

Suppose the probability distribution now is given as shown below.



The distribution is heavily skewed to the left and has population mean and variance of μ = 3.2787 and σ2 = 0.3331. We let n = 20 first to see how good the approximation is.

> n <- 20

> set.seed(8129)

> set1 <- 4\*matrix(data = rbeta(n = 1000\*n, shape1 = 5, shape2

 = 1.1), nrow = 1000, ncol = n)

> ybar <- rowMeans(set1)

> head(round(ybar,2))

> head(round(ybar,2))

[1] 3.68 3.39 3.41 3.40 3.18 3.41

> #Histogram with normal distribution overlay

> mu <- 3.278689

> sigma.sq <- 0.3330923

> hist(x = ybar, main = "Histogram of sample means", xlab =

 expression(bar(y)), freq = FALSE)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq/n)),

 add = TRUE, col = "red")



Like what we did in the first example, we make the histogram a little better:

> hist(x = ybar, ylim = c(0, 3.2), main = "Histogram of sample

 means", xlab = expression(bar(y)), freq = FALSE, breaks

 = 15)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq/n)),

 add = TRUE, col = "red")



The normal approximation works rather well.

For n = 10, the skewness is much more obvious:

> hist(x = ybar, xlim = c(2.4, 4), main = "Histogram of sample

 means", xlab = expression(bar(y)), freq = FALSE, col =

 NA)

> curve(expr = dnorm(x = x, mean = mu, sd = sqrt(sigma.sq/n)),

 add = TRUE, col = "red")



What happens for a larger sample size? Here’s n = 40.



Questions/Comments:

* Does the distribution of  look very symmetric?
* Is n = 20 enough for the central limit theorem to work well?
* What happens as n changes?
* The histogram of  for n = 100 is



Our conclusions are

* There isn’t a universal smallest number that guarantees the central limit theorem to work.
* If the distribution of the population is relatively symmetric, then the sample mean reaches approximate normality for smaller samples than if the probability distribution for Y is skewed or has a large departure from normality.
* The sample size needed for the observed  to achieve approximate normality depends on the probability distribution for Y in the population.