**Central Limit Theorem (named “Sampling distributions” in video)**



Suppose Y1, Y2, …, Yn are random variables representing a random sample from a population with probability distribution possibly unknown. Also, suppose each random variable has the same population mean μ and population variance σ2. What does this mean?

* E(Y1) = μ, E(Y2) = μ, , E(Yn) = μ
* Var(Y1) = σ2, Var(Y2) = σ2, , Var(Yn) = σ2
* Relating to the GPA example, we could have μ = 2.8571 and σ2 = 0.4082.
* The random sample part means Y1, Y2, , Yn are independent random variables.

Questions:

* Is  a random variable?
* Is  a random variable?

What would we expect the sample mean to be on average if we repeatedly took random samples and calculated  each time? One can show using properties of expected values that



Some books will use  to denote .

In a similar manner, one could also show that



Some books will use  to denote .

We can say the standard deviation is . Notice the effect that n has on these quantities.

Proofs:

Remember that for two independent random variables X and Y and constants a and b,

* E(aX + bY) = aE(X) + bE(Y)
* Var(aX + bY) = a2Var(X) + b2Var(Y)







Why are these items important? This helps to determine the characteristics of sample mean BEFORE we even take a sample!!!

We can go even further! One can show that the probability distribution of  is APPROXIMATELY a normal distribution with a mean of μ and variance of σ2/n when the sample size is LARGE. This result is known as the central limit theorem!

Notes:

* Terminology: The distribution of a statistic, like  here, is often referred to as a sampling distribution. This is still a probability distribution, just a different name that some people use.
* What is a large sample? A general “rule of thumb” is for n 30. A normal distribution may do a good job of approximating the distribution for  even for smaller sample sizes. It is dependent on what the underlying probability distribution is for Y1, Y2, …, Yn.
* For emphasis, notice the central limit theorem holds no matter what the underlying probability distribution is for Y1, Y2, …, Yn. You just need a large enough sample.
* The central limit theorem result is often written as



where Z has an approximate normal distribution with mean 0 and standard deviation 1.

* There is one problem with the central limit theorem – you need to know μ and σ2. How to get around this problem will be discussed in later in the course.

Example: GPA and the central limit theorem (CLT\_GPA.R)

Let Y be a random variable representing GPA. Suppose the population of student GPAs can be characterized by the following PDF



for 0 ≤ y ≤ 4 and f(y) = 0 for all other possible values of y. One can show that μ = E(Y) = 2.8571 and σ2 = Var(Y) = 0.4082.



A random sample of size 20 is taken from this population. Thus, we have Y1, …, Y20 now has our random variables. And, suppose we repeat this random sampling process 1,000 times!

Below are the first 6 samples of size 20 and the first 6 sample means (you are not responsible for how I took the sample):

> head(round(set1,2))

 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]

[1,] 3.41 2.79 2.65 1.87 3.21 2.99 2.96 3.83 3.28 2.15 1.11 2.73 3.68 3.34

[2,] 3.01 3.15 3.83 2.25 3.75 3.76 2.22 1.99 2.50 3.29 3.64 3.55 3.45 2.45

[3,] 2.90 2.23 2.95 3.17 3.25 3.85 3.59 3.37 3.57 3.11 3.69 1.55 3.38 3.47

[4,] 3.06 2.59 2.50 2.60 2.33 2.19 2.01 2.29 2.94 2.89 3.34 2.33 3.81 3.72

[5,] 2.38 2.74 3.48 3.25 3.24 3.34 3.17 2.79 2.25 3.39 2.04 2.39 3.08 3.06

[6,] 2.46 3.51 2.21 2.88 3.75 2.40 3.50 3.04 2.63 2.98 2.00 1.90 3.98 2.51

 [,15] [,16] [,17] [,18] [,19] [,20]

[1,] 3.56 2.12 3.27 3.76 2.91 3.59

[2,] 3.57 3.26 3.04 2.77 3.33 3.34

[3,] 3.37 3.80 3.68 3.33 2.42 3.58

[4,] 2.83 3.27 3.74 2.10 2.95 2.79

[5,] 2.85 2.29 3.38 3.60 2.89 2.61

[6,] 2.41 1.74 3.20 2.60 2.17 2.81

> ybar <- rowMeans(set1)

> head(round(ybar,2))

[1] 2.96 3.11 3.21 2.81 2.91 2.73

Histogram of the  with normal distribution overlay:



μ and :

> #Estimate of mu

> mean(ybar)

[1] 2.854353

> #Estimate of sigma/sqrt(n)

> sd(ybar)

[1] 0.1337018

> #Actual sigma/sqrt(n)

> sqrt(sigma.sq/n)

[1] 0.1428636

Note that if ALL possible samples of size 20 were taken, the mean and standard deviation of all of them would be μ = 2.8571 and  = 0.1429

:

> #Estimate of P(3 < Ybar < 4)

> mean(ybar <= 4) - mean(ybar <= 3)

[1] 0.138

> #Estimate of P(3 < Ybar < 4) using a normal

 distribution approximation

> pnorm(q = 4, mean = mu, sd = sqrt(sigma.sq/n)) –

 pnorm(q = 3, mean = mu, sd = sqrt(sigma.sq/n))

[1] 0.1585936

Notice how close the probability resulting from the central limit theorem is to the probability resulting from the simulated probability distribution of . Thus, the central limit theorem allows us to calculate these probabilities without taking 1,000 samples of size 20, finding the mean, finding the variance, … . Of course, taking 1,000 samples of size 20 is not feasible for the vast majority of real-life applications!

Because we thoroughly discussed finding probabilities with the normal probability distribution earlier in the chapter, many of the same techniques with finding these probabilities apply here. **Remember the main advantage of using**  **is that you do not need to know the probability distribution** **for Y!**

Example: Cereal boxes (cereal\_boxes.R)

A cereal manufacturer claims that it fills boxes on average with 24 oz. of cereal. Also, suppose a FDA official wants to determine if the cereal boxes truly have the advertised weight of 24 oz. The FDA official randomly samples 36 boxes of cereal.

Suppose the cereal boxes are truly being made with μ = 24 oz. of cereal and σ = 2 oz. of cereal.

1. What is the approximate probability the sample mean weight is greater than 23 oz.?

Notice that nothing is said about the probability distribution for each box here!!!

Plot of the approximate normal distribution for 
(μ = 24 and ).

> mu <- 24

> sigma <- 2

> n <- 36

> #Approximate normal distribution for Ybar - not

 responsible for expression() part

> curve(expr = dnorm(x = x, mean = mu, sd =

 sigma /sqrt(n)), xlim = c(22 26), col =

 "red", n = 1000, xlab = expression(bar(Y)),

 ylab = expression(paste("f(", bar(Y), ")")))

> abline(h = 0)



Find P(> 23) = 1 – P(< 23). The resulting probability is 0.9987.

> 1 - pnorm(q = 23, mean = mu, sd = sigma/sqrt(n))

[1] 0.9986501

1. What is the approximate probability the sample mean weight of the boxes is between 23 and 25 oz.?

Find P(23 <  < 25). The resulting probability is 0.9973.

> pnorm(q = 25, mean = mu, sd =sigma/sqrt(n)) –

 pnorm(q = 23, mean = mu, sd =sigma/sqrt(n))

[1] 0.9973002

1. The company will be fined if the sample mean weight of the boxes is not within ±1 oz. of the advertised true mean. What is the approximate probability the company will receive a fine?

Find P(< 23 or  > 25). This is 1 – P(23 <  < 25) = 1 - 0.9973 = 0.0027.