**Random sample**



How do you take the sample?

There are a number of ways to take a sample. In fact, we have a whole course on it! We will focus primarily on one way to take a sample that we saw earlier named a random sample.

Random sample – A sample from the population where every item has an equal probability of being included within the sample.

The easiest way to take a random sample

Number all items from your population with 1, 2, …, N. Use the sample() function in R to take a random sample of size n.

Example: Random sample (random\_sample.R)

Suppose the population is very small with N = 20 and the sample size is n = 10. Items are numbered 1 to 20.

> N <- 20

> n <- 10

> set.seed(9812)

> sample(x = 1:N, size = n)

[1] 17 14 19 10 13 3 9 16 5 7

What if I used a different seed number?

Suppose the population has a size of N = 25,000 and the sample has a size of n = 100.

> N <- 25000

> n <- 100

> set.seed(7823)

> sample(x = 1:N, size = n)

[1] 19628 13653 5074 19885 21474 20136 12240 17954 15127

[10] 22689 18885 3330 1935 9221 12763 16808 20535 847

[19] 20588 979 2521 24099 10006 8414 9066 19460 1499

[28] 23788 13400 23016 1453 16761 14654 19271 19559 15535

[37] 802 991 5570 1241 10906 3436 17237 3372 3681

[46] 3860 1350 17968 931 6032 2478 23082 1614 1595

[55] 14010 8183 13473 23012 1287 19611 16513 16467 6044

[64] 15029 24735 18273 3517 19328 18802 7715 14776 15591

[73] 18550 4553 12582 16415 15253 15408 13201 8137 13825

[82] 1491 4882 15282 20068 95 13846 15991 7351 3617

[91] 10551 9348 3830 629 15957 11921 23424 5792 5470

[100] 17549

Should you use the same seed number every time?

An older way to take a random sample

Sets of randomly generated numbers are given in a table form called a “random number table.” Textbooks sometimes still have these tables. There are a variety of ways to use this table to obtain a random sample. For example, suppose N = 20 and n = 4 and the table has the following numbers in its first row:

10480 15011 01536 02011

Look at each two consecutive digit pair. This corresponds to the number of the item to choose from the population for the sample. If a two-digit number is greater than 20, go to the next two-digit number. If you come across a repeat number less than 20, go to the next two-digit number.

My sample is 10, 01, 11, 20.

**Sampling distributions**

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.

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.

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)

Suppose a sample of students was taken 1,000 times from the population of all students on campus. For each sample, student GPAs were recorded using a sample size of 20. Also, suppose the population can be characterized by the probability distribution shown below.



Because I did the sampling from the population, I know that μ = 2.8571 and σ2 = 0.4082.

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)

> sum(ybar < 4)/1000 - sum(ybar <= 3)/1000

[1] 0.138

> #Also could just do sum(ybar > 3) since all GPAs <=

4 in data

> #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

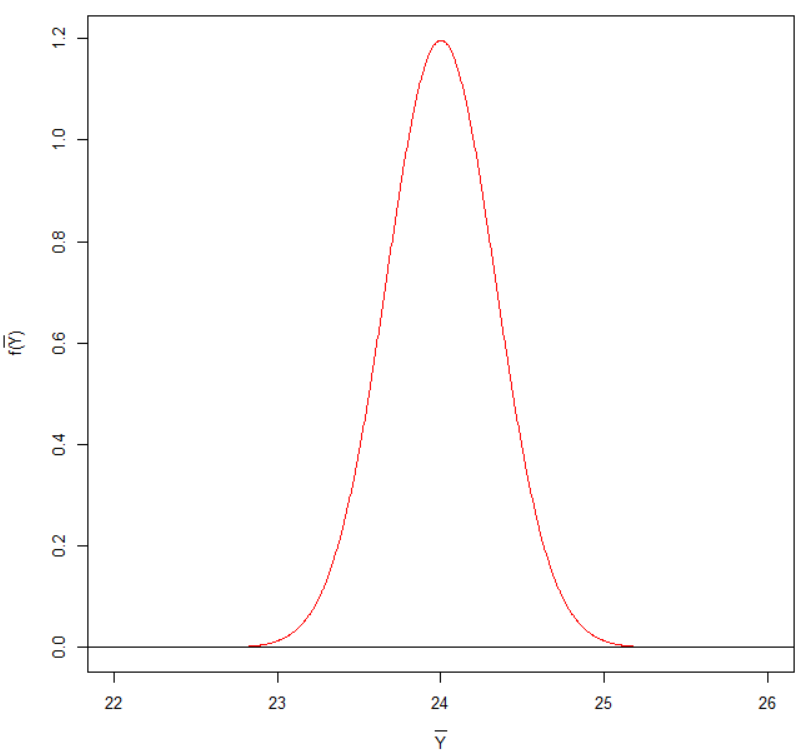
> curve(expr = dnorm(x = x, mean = mu, sd =

sigma /sqrt(n)), from = 22, to = 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.