The Central Limit Theorem

STAT-UB.0001 – Statistics for Business Control

The Central Limit Theorem

1. Consider the population of all Fortune 500 CEOs and their salaries. Suppose that the mean salary (in millions of dollars) is $\mu = 20$, and the standard deviation of the salaries is $\sigma = 5$. You sample 50 CEOs and find their salaries.

   (a) Draw a histogram of what you think the population looks like.

   (b) Let X be the salary of a randomly drawn CEO. Consider X to be a random variable. What is the histogram of X? What is its expectation and standard deviation?

   (c) Consider the sample mean $\bar{X}$ to be a random variable. What is the expectation of $\bar{X}$?

   (d) What is the standard deviation of $\bar{X}$?

   (e) Draw a picture of what you think the PDF of $\bar{X}$ looks like.

2. You draw a random sample of size $n = 64$ from a population with mean $\mu = 50$ and standard deviation $\sigma = 16$. From this, you compute the sample mean, $\bar{X}$.

   (a) What are the expectation and standard deviation of $\bar{X}$?

   (b) Approximately what is the probability that the sample mean is above 54?

   (c) Do you need any additional assumptions for part (c) to be true?

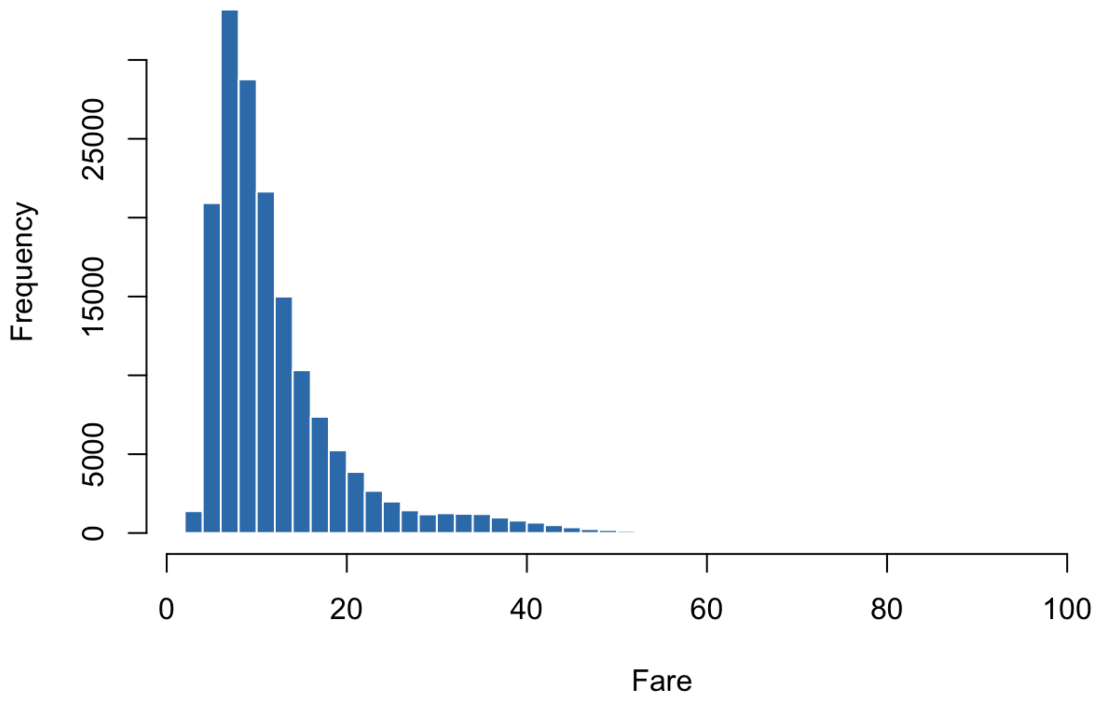
3. You draw a random sample of size $n = 16$ from a population with mean $\mu = 100$ and standard deviation $\sigma = 20$. From this, you compute the sample mean, $\bar{X}$.

   (a) What are the expectation and standard deviation of $\bar{X}$?

   (b) Approximately what is the probability that the sample mean is between 95 and 105?

   (c) Do you need any additional assumptions for part (c) to be true?

Here is a histogram of the fares (including tax and tolls) of 162,997 taxi trips taken within New York City in 2013.



The following table displays the trips with the highest and lowest fares.

Pickup			Dropoff						
Time	Borough	CD	Time	Borough	CD	Mins.	Miles	Fare ($)	Tip ($)
01-26 08:42:26	Manhattan	2	01-26 08:43:10	Manhattan	4	0.7	0.1	3.00	0.00
01-21 16:54:58	Manhattan	8	01-21 16:55:37	Manhattan	8	0.6	0.2	3.00	0.00
02-13 11:24:00	Manhattan	7	02-13 11:25:00	Manhattan	7	1.0	0.0	3.00	0.00
03-15 14:58:43	Manhattan	4	03-15 14:59:52	Manhattan	5	1.1	0.0	3.00	0.00
03-20 07:07:00	Queens	1	03-20 07:08:00	Queens	1	1.0	0.0	3.00	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
05-23 11:54:00	Queens	83	05-23 13:25:00	Brooklyn	1	91.0	28.4	87.00	15.00
06-29 01:56:00	Manhattan	1	06-29 03:03:00	Staten Is.	3	67.0	25.5	93.49	23.10
10-24 22:26:20	Manhattan	4	10-24 23:31:02	Staten Is.	3	64.7	27.9	99.49	19.89
06-12 13:04:00	Queens	83	06-12 14:02:00	Staten Is.	3	58.0	33.2	100.66	0.00
06-14 18:44:00	Queens	83	06-14 19:44:00	Staten Is.	3	60.0	36.0	107.66	15.00

The mean fare ($) is 12.424, the median is 10.000, and the standard deviation is 7.966.

4. Suppose that we randomly select 100 items from the Taxi dataset. What you say about the fares of the items in this sample?

5. Consider the (hypothetical) sample of 100 taxi fares. Will the sample mean be *exactly* equal to 12.424? Approximately how close will the sample mean be to this value?

6. I performed 10,000 replicates of the following procedure: randomly sample 100 fares from the taxi data set, then compute the mean and standard deviation of the sample. The following table lists the results from the first few replicates. What can you say about the sample means?

Rep.	Mean	Std. Dev.
1	13.093	9.034
2	12.885	8.341
3	13.079	9.033
4	10.895	7.031
5	13.478	8.905
6	13.207	7.037
⋮	⋮	⋮

Confidence Intervals

7. You can consider the dataset of 162,997 taxi fares to be a sample from a larger population.

   (a) What are some reasonable choices for this population?

   (b) Give a range of plausible values for the mean of the population you specified in part (a). *Hint: you do not know σ exactly, but since n is large, you can assume $\sigma \approx s$.*

   (c) Under what conditions will your "range of plausible values" be trustworthy?