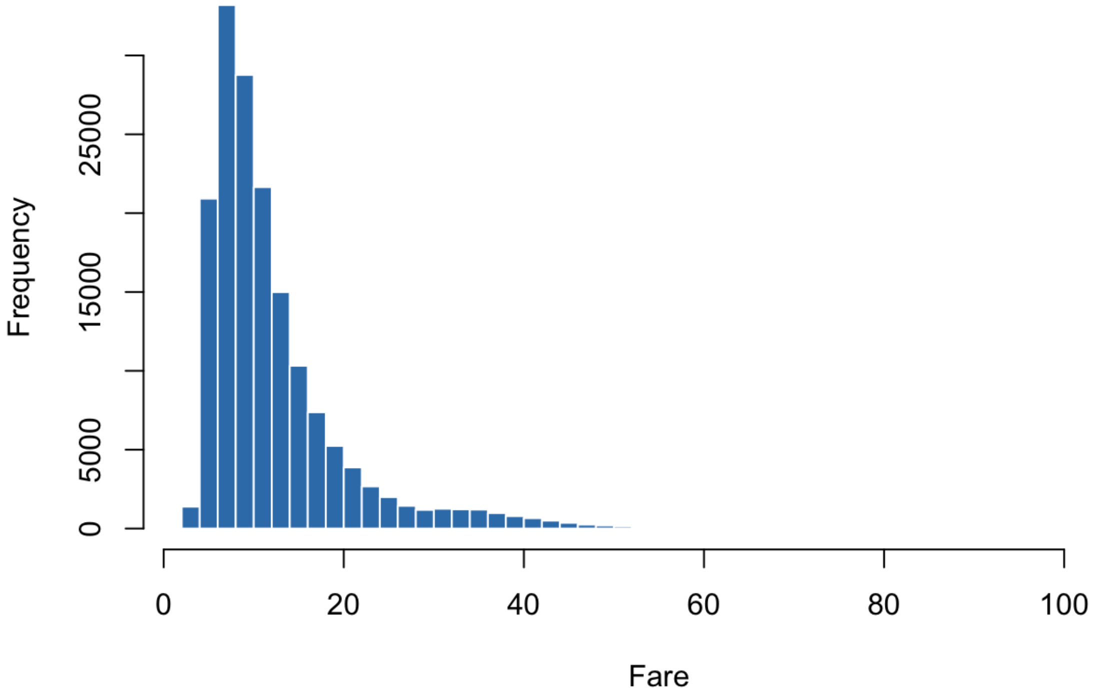
Sampling Distributions (Solutions)

STAT-UB.0003: Regression and Forecasting Models

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.

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

Solution: The histogram will look like the histogram of all 162,997 fares; the mean, median, and standard deviation will be close to the values from the complete dataset.

2. 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?

Solution: The population here is the collection of all 162,997 taxi fares. The sample mean will be within about

$$2\frac{\sigma}{\sqrt{n}} = 2\frac{7.966}{\sqrt{100}} = 1.593$$

of the population mean, $\mu = 12.424$. That is, there is roughly a 95% chance that the sample mean, $\bar{X}$ will be in the range

$$\begin{aligned}\mu \pm 2\frac{\sigma}{\sqrt{n}} &= 12.424 \pm 1.593 \\ &= (10.831, 14.017).\end{aligned}$$

(If you want to be more precise, you can use 1.96 instead of 2.)

3. 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
⋮	⋮	⋮

Solution: Each replicate is computing a sample mean from $n = 100$ samples; the population mean and standard deviation are $\mu = 12.424$ and $\sigma = 7.966$. Thus, we know that:

(a) The mean of the means will be close to

$$\mu_{\bar{X}} = \mu = 12.424.$$

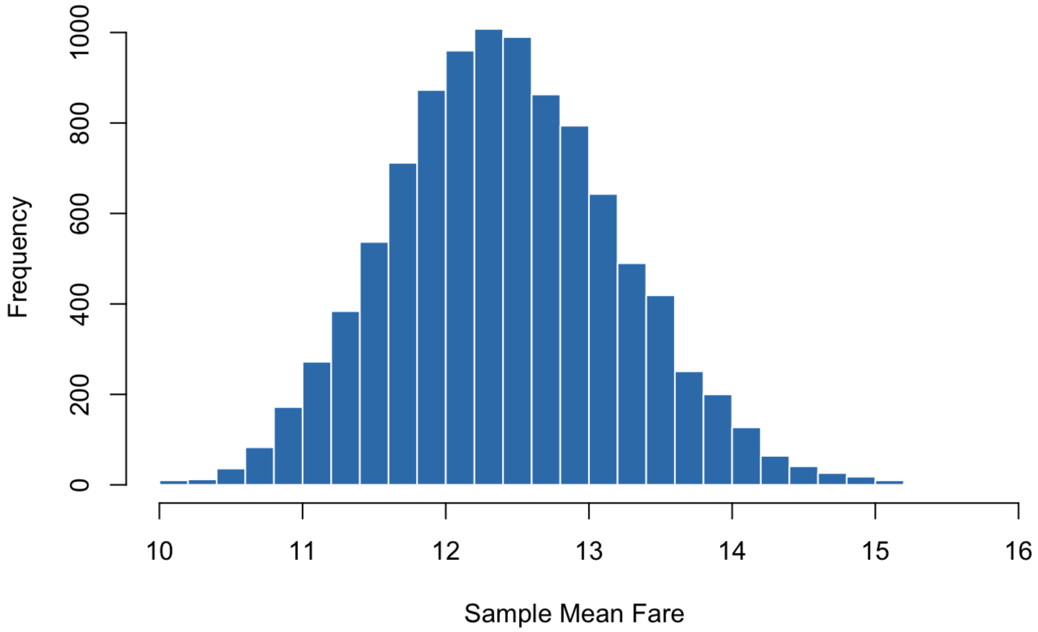
In fact, it was 12.420.

(b) The standard deviation of the means will be close to

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{7.966}{\sqrt{100}} = 0.797$$

In fact, it was 0.799.

(c) The histogram of the means will look like a bell curve. The histogram from my replicates does in fact exhibit this behavior:



4. 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?

   Solution: One reasonable choice is the taxi fares for all 2013 New York City taxi trips.

   (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 $\sigma \approx s$.*

   Solution: We can be 95% confident that the population mean is within $2\frac{\sigma}{\sqrt{n}}$ of the sample mean, where $n = 162997$. Using the approximation $\sigma \approx s$, we can get a 95% confidence interval for the population mean:

$$\begin{aligned}\bar{x} \pm 2\frac{s}{\sqrt{n}} &= (12.424) \pm 2\frac{(7.966)}{\sqrt{162997}} \\ &= 12.424 \pm 0.039 \\ &= (12.385, 12.463)\end{aligned}$$

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

   Solution: Since the sample size is large ($n = 162997$), the the Central Limit Theorem will be in force—making our confidence interval trustworthy—as long as the samples are drawn independently from the population. This will be the case if the sample is a simple random sample from the population.

   If there is any bias in the sampling procedure, for example if the sample contains a disproportionate number of weekend trips, then the interval will not be valid.