Week 8: Sampling Distribution

Example 1

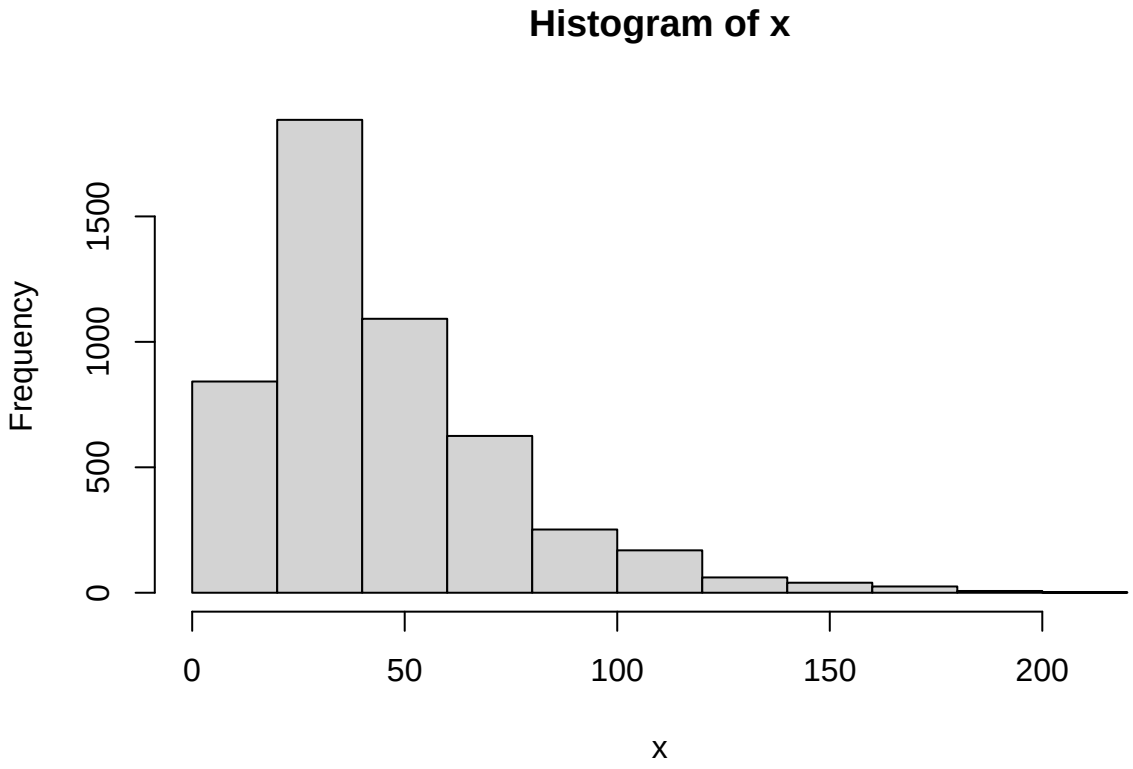
A large financial institution in NYC has about 5000 people working at the Wall Street location.

```
dat <- read.csv(url("https://nhorton.people.amherst.edu/is5/data/Population_Commute_Times.csv"))
head(dat)
```

```
##   Commute.Time
## 1          185
## 2           18
## 3           27
## 4           39
## 5          122
## 6           54
```

We assign x as the commute time, and show the histogram.

```
x <- dat$Commute.Time
hist(x)
```



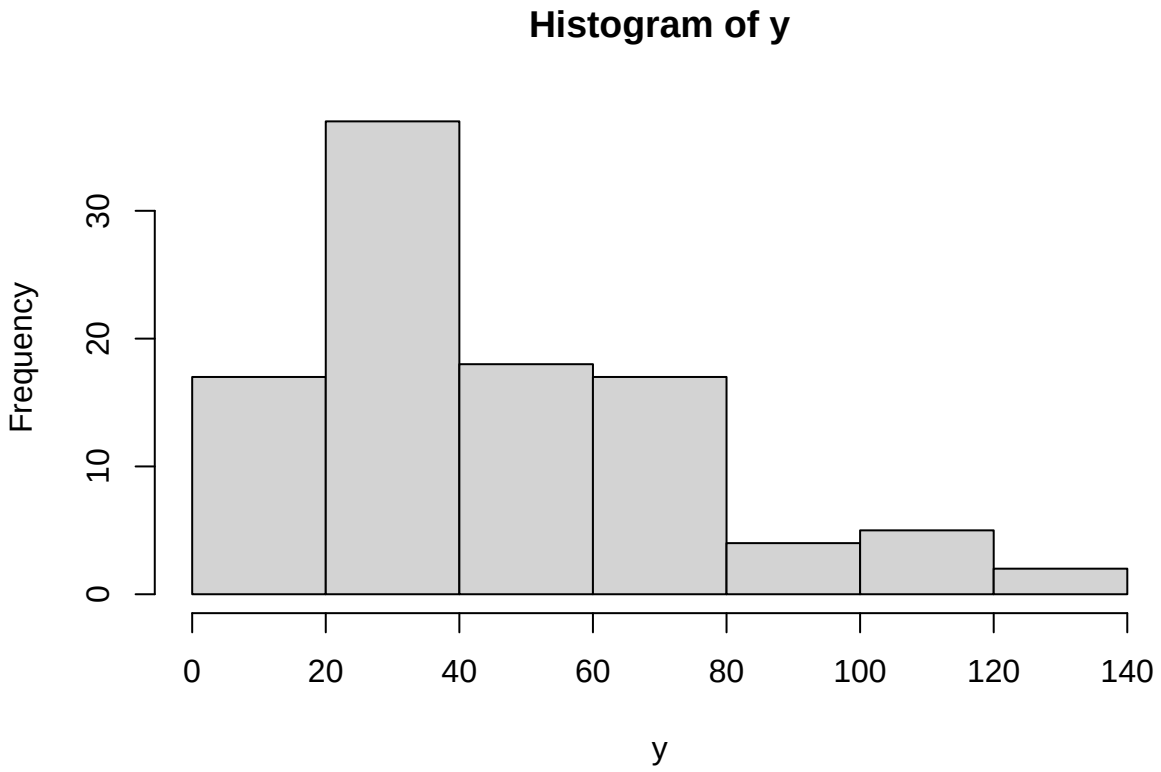
Let's find the summary statistics.

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   25.00   38.00   45.43   59.00  202.00
```

The Human Resources Department chose 100 employees and interviewed them about their commute experience. Here is a histogram of the 100 responses.

```
y <- sample(x, 100)
hist(y)
```



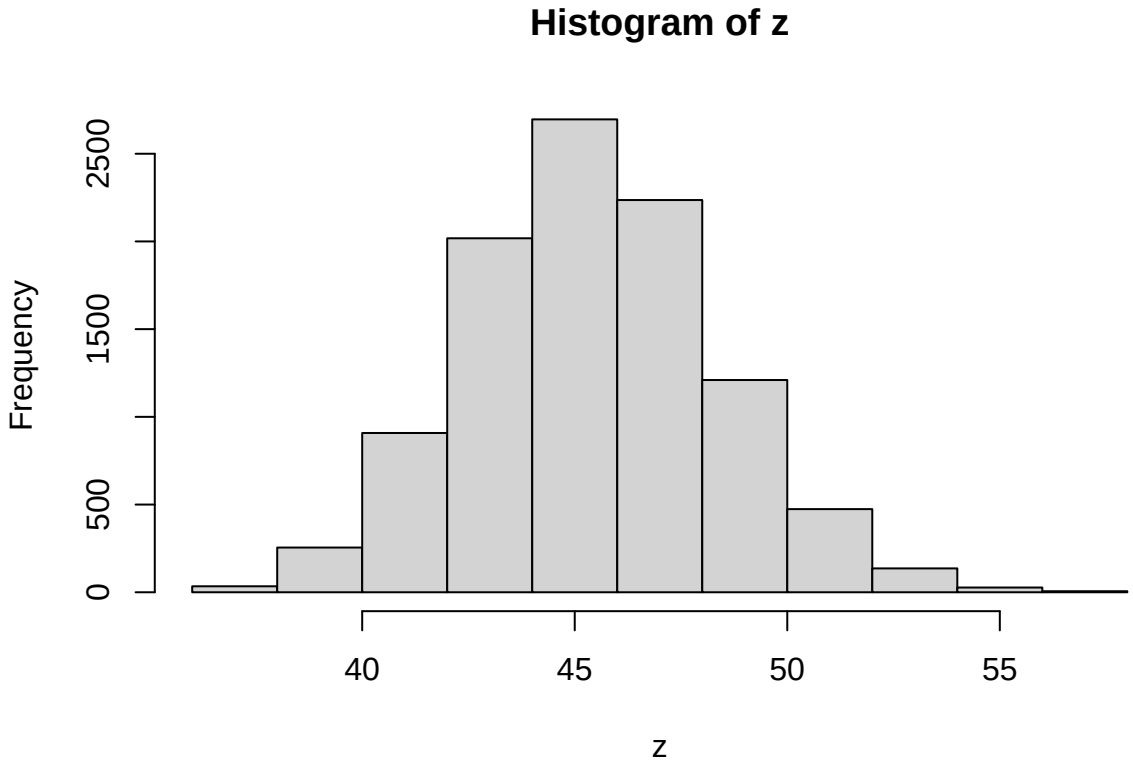

```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   24.00   37.00   45.39   64.25  127.00
```

Students should run these three lines several times and observe the variation among samples.

We use the computer to simulate 10,000 different random samples of size 100.

```
z <- replicate(10000, mean(sample(x, 100)))
hist(z)
```



We conclude that statistics vary from sample to sample, but most of the sample means are close to each other around 45 minutes. The population mean is 45.4 minutes.

Follow Up

Simulate 10,000 different random samples of size 50. Compare the histogram of the means of samples of size 50 with that of size 100 above.