# Sampling distributions part 2 (means)

Thomas Scofield

Feb. 21, 2020

## Pertinent R

The `sample()` command. You can make a containter of objects and make random draws from it:

```
die = c(1,2,3,4,5,6)
sample(die, size=3, replace=TRUE)   # the optional 'replace' switch affects command's behavior
```

```
## [1] 3 5 2
```

Or, you can sample (draw randomly) rows/cases from a data frame.

```
sample(iris, size=5, replace=TRUE)
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species orig.id
## 31           4.8         3.1          1.6         0.2     setosa      31
## 47           5.1         3.8          1.6         0.2     setosa      47
## 42           4.5         2.3          1.3         0.3     setosa      42
## 54           5.5         2.3          4.0         1.3 versicolor      54
## 146          6.7         3.0          5.2         2.3  virginica     146
```

There is a related command, `resample()`, which assumes you want to sample with replacement, so it doesn't require the extra switch `replace=TRUE` to make that happen.

```
resample(iris, size=5)    # automatically draws with replacement
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Species orig.id
## 19           5.7         3.8          1.7         0.3    setosa      19
## 106          7.6         3.0          6.6         2.1 virginica     106
## 38           4.9         3.6          1.4         0.1    setosa      38
## 42           4.5         2.3          1.3         0.3    setosa      42
## 2            4.9         3.0          1.4         0.2    setosa       2
```

## Pertinent R Markdown

At times, you will want to insert math symbols into a report. One can write E=mcˆ2 as text, but it looks better if you enter the same thing in math mode. You indicate the start to math mode by including a dollar-sign $ in your source file, and once your equation is finished, you indicate math mode is over with another dollar sign. So, placing $E=mcˆ2$ in your source (.Rmd) file results in $E = mc^2$. Greek letters require math mode, and what you put between the dollar signs is a backslash character \ followed by the name (spelled out in English) of the desired Greek letter. So, including $\mu$ in the source file produces $\mu$ in your document. See if you can guess what you to include in your .Rmd file in order to get the two versions of the Greek letter sigma—that is, $\Sigma$ and $\sigma$. Math mode also allows you to add things like a "hat" to $p$, or a "bar" over an $x$.

- In .Rmd file: $\hat p$ is rendered as $\hat{p}$
- In .Rmd file: $\overline x$ is rendered as $\overline{x}$

## Sampling distribution for sample mean $\overline{x}$

The textbook illustrates this same concept using the data set **StatisticsPhD**, the first few rows of which are

```
head(StatisticsPhD)
```

```
##                        University    Department FTGradEnrollment
## 1               Baylor University    Statistics               26
## 2               Boston University Biostatistics               39
## 3                Brown University Biostatistics               21
## 4      Carnegie Mellon University    Statistics               39
## 5 Case Western Reserve University    Statistics               11
## 6       Colorado State University    Statistics               14
```

We might calculate the mean (which can be viewed as a population mean $\mu$, since this data is a *census* of all Ph.D. statistics programs in the country, not merely a sample of such programs) number of full-time students enrolled, using the `mean()` command and indicating column (variable name) and data frame name in the usual way:

```
mean(~FTGradEnrollment, data=StatisticsPhD)
```

```
## [1] 53.53659
```

If, however, we want to simulate the process of computing a mean from a random sample of 10 of these Ph.D. programs, we need only indicate a different data set on which to calculate the mean—not the full **StatisticsPhD** data frame, but a random sample of 10 cases selected from it:

```
mean(~FTGradEnrollment, data=resample(StatisticsPhD,size=10))
```

```
## [1] 47
```

This command

- draws a sample of 10 schools with replacement
- calculates $\overline{x}$, the mean number of `FTGradEnrollment` for the schools in the sample.

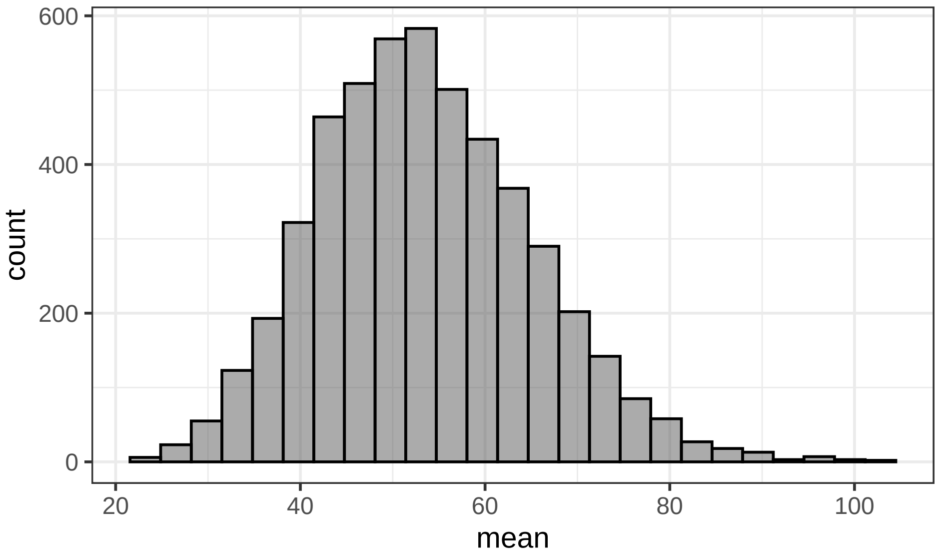
Doing it once gives you $\overline{x}$ for a single random sample with $n = 10$. Do it many times, and you'll start to see the sort of values $\overline{x}$ can take, and which ones occur more frequently, in a random sample of 10 programs. In other words, you get an idea of the sampling distribution of $\overline{x}$ for samples of size 10 taken from this population.

```
manyXbars <- do(5000) * mean(~FTGradEnrollment, data=resample(StatisticsPhD, size=10))
head(manyXbars)
```

```
##   mean
## 1 51.6
## 2 51.6
## 3 73.6
## 4 65.1
## 5 39.3
## 6 34.8
```

Seeing that the results (stowed in the data frame `manyXbars`) have been given the column name `mean`, we will incorporate those names in a call to the `gf_histogram()` command:

```
gf_histogram(~mean, data=manyXbars, color="black")
```




**Exercise 1:**

a) Find an approximate standard error of the mean, $\mathrm{SE}_{\overline{x}}$, for samples of size 10 from this population.
b) Does it appear that $\overline{x}$ is an unbiased estimator of $\mu$? How can you tell?
c) How often (give an answer in terms of *relative frequency*) is $\overline{x}$ as large as 60?

**Exercise 2:**

Repeat the work done already to generate and view an approximate sampling distribution for $\overline{x}$, again with sample sizes $n = 10$, but with the difference that you sample without replacement instead of with replacement. Are there noticeable differences between this sampling distribution and the one for samples of size $n = 10$ taken with replacement? Does `favstats()` reveal any major differences in the two distributions?

**Exercise 3:**

Now that you have looked at sampling distributions for $\overline{x}$ both with and without replacement for samples of size $n = 10$, try comparing the two for samples of size $n = 20$. Once again apply `favstats()` to see if there are notable differences. Are the differences you noted between the two distributions at $n = 10$ even more pronounced at $n = 20$?