The Sampling Distribution of the Mean

by James Bernhard
Suppose we would like to know the rv mean of the random variable which is height (and which we denote by $X$), taken over all UPS students.

We refer to this group that we’re interested in (consisting of all UPS students) as the *population*.

Since the rv mean we want is taken over the whole population, we’ll call it the *population mean* and denote it by $\mu$.

Also, for future use, it will be useful to refer to the standard deviation of $X$, taken over all UPS students, as the *population standard deviation* and to denote it by $\sigma$.

It is impractical to measure the height of more than a small number of students, so we choose 20 students at random and measure their heights.
We call this group of 20 students a *sample* from the population.

We call the number of students (20) in the sample the *sample size*, and denote it by \( n \).

The rv mean of \( X \) (student height) taken only over the students in the sample is called the *sample mean* of \( X \) and is denoted by \( \bar{x} \).
We call this group of 20 students a *sample* from the population.

We call the number of students (20) in the sample the *sample size*, and denote it by $n$.

The rv mean of $X$ (student height) taken only over the students in the sample is called the *sample mean* of $X$ and is denoted by $\bar{x}$.

Suppose we find that the sample mean $\bar{x}$ of the heights is 65.3 inches for our sample.

*What does this value of $\bar{x}$ tell us about the population mean $\mu$ of student heights?*
We won’t answer this question in full this chapter, but to start with at least, we first view the sample mean (for samples of size 20) as a random variable $\bar{x}$

The random process that $\bar{x}$ depends on is the selection of the sample (of size 20) from the population

In other words, $\bar{x}$ assigns a number (the sample mean) to each outcome of the random process that is “selecting a group of 20 students at random”
We won’t answer this question in full this chapter, but to start with at least, we first view the sample mean (for samples of size 20) as a random variable $\bar{x}$

The random process that $\bar{x}$ depends on is the selection of the sample (of size 20) from the population

In other words, $\bar{x}$ assigns a number (the sample mean) to each outcome of the random process that is “selecting a group of 20 students at random”

The sample space for $\bar{x}$ is GIGANTIC: it consists of all possible sets of 20 students chosen from the UPS student population

If the UPS student population has 2500 students, there are about $3.46 \times 10^{49}$ possible sets of 20 students chosen from it
We would really like to know about the distribution of $\overline{x}$, which is called the *sampling distribution* (of student heights in this case).

As a practical matter, we can’t examine anywhere near all of the possible outcomes in the sample space for $\overline{x}$.

What’s more, we often can’t examine any more than *one* of the possible outcomes in the sample space for $\overline{x}$ (the sample we chose).

We can still learn things about the sampling distribution $\overline{x}$ though if we make an additional assumption:

The samples chosen are *simple random samples*, meaning that each possible sample (of size 20) is equally likely to be chosen in the process.
If the samples as chosen as simple random samples (from a large population), we have the following key results:

1. The rv mean of $\bar{x}$ equals the population mean
2. The rv standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the size of the sample (which is 20 in this case)
3. *The Central Limit Theorem*: If the sample size is large, the distribution of $\bar{x}$ is approximately normal

In symbols, we have:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

The third item on this list holds only approximately, for large $n$ (the approximation becoming better as $n$ gets larger)
To summarize these results in English, we can use the book’s phrase: *Averages are less variable and more normal than individual observations*

Do not confuse the following concepts:

1. The *sample mean* (denoted by $\bar{x}$), which is the mean of $X$ restricted to individuals in the sample
2. The *sampling mean* (denoted by $\mu_{\bar{x}}$), which is the mean of the random variable $\bar{x}$
To illustrate these results about sampling distributions, let’s look at an:

Applet from the Rice Virtual Lab in Statistics

We can also look at a demonstration in R, in which we estimate the population mean of income among 43000+ people by sampling