Chapter 5.9 Sampling Distribution of the Mean

Jim Albert and Monika Hu

Chapter 5 Continuous Random Variables

Jar of Candies

- There is a general result about the shape of sample means that are taken from any population.
- Suppose one has a jar filled with a variety of candies of different weights.
- One is interested in learning about the mean weight of a candy in the jar.
- Instead of weighing all of the candies, suppose one selects a random sample of 10 candies from the jar and finds the mean of weights of these 10 candies.
- What has one learned about the mean weight of all candies from this sample data?

To answer this type of question, one

- assumes that one know about the weights of all candies in the jar
- looks at the pattern of means that one obtains when one takes random samples from the jar
- the group of items of interest is called the population.

The population

Assume we know exactly the weights of all candies in the jar. There are five types of candies

	Weight	Proportion
fruity square	2	0.15
milk maid	5	0.35
jelly nougat	8	0.20
caramel	14	0.15
candy bars	18	0.15

Population summaries

- Let X denote the weight of a randomly selected candy from the jar.
- This distribution is summarized by computing a mean μ and a standard deviation σ.
- One can show $\mu = 8.4500$ and $\sigma = 5.3617$.
- If one was really able to weigh each candy in the jar, one would find the mean weight to be $\mu = 8.4500$ grams.

Take a random sample

- Suppose a random sample of 10 candies is selected with replacement from the jar and the mean is computed.
- This is called the sample mean $\bar{X}$ to distinguish it from the population mean μ.

Sampling Candies

- Suppose one obtains the following weights (in grams) of a sample of 10 candies:

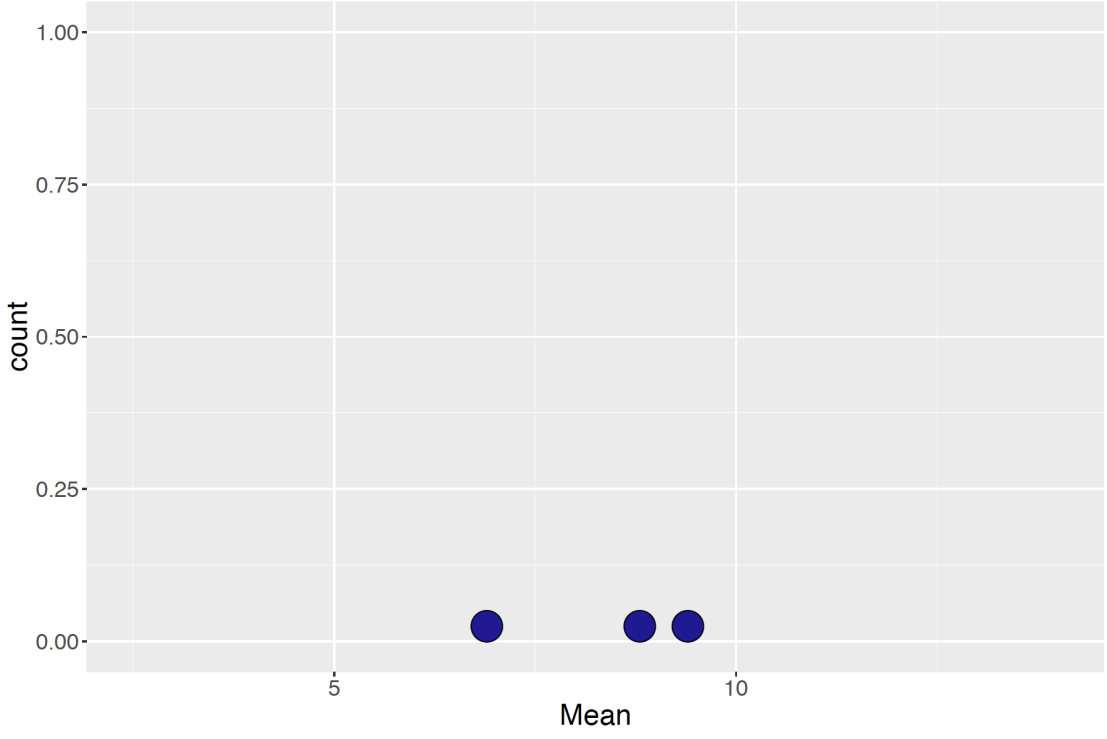
`5, 8, 5, 14, 5, 18, 8, 18, 5, 8`

One computes the sample mean

$$\bar{X} = (5 + 8 + 5 + 14 + 5 + 18 + 8 + 18 + 5 + 8)/10 = 9.4\,\mathrm{gm}.$$

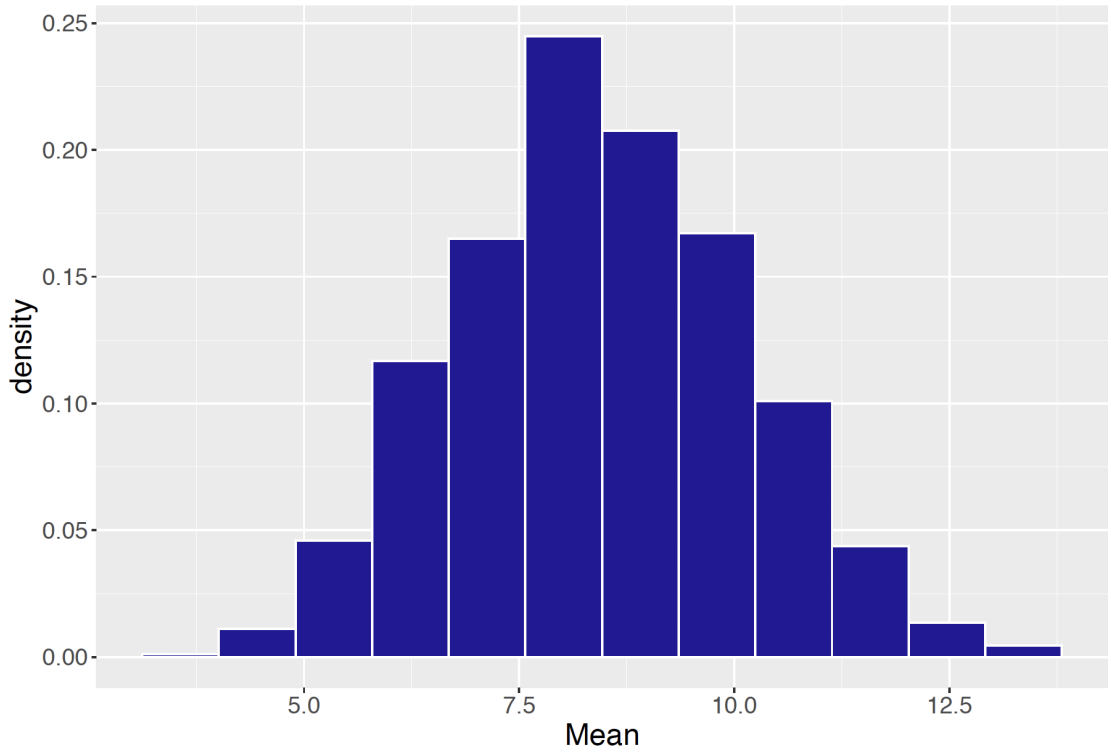
Repeated sampling

- Suppose this process is repeated two more times – in the second sample, one obtains $\bar{X}$= 6.9 gm and in the third sample, one obtains $\bar{X}$= 8.8 gm.
- Plot the three sample mean values below.



Keep sampling

- Suppose that one continues to take random samples of 10 candies from the jar and plot the values of the sample means on a graph.
- Obtain the sampling distribution of the mean $\bar{X}$, shown below.



Normal shape

- We see an interesting pattern of these sample means – they appear to have a Normal shape.
- This motivates an amazing result, called the Central Limit Theorem about the pattern of sample means.
- If one takes sample means from any population with mean μ and standard deviation σ, then the sampling distribution of the means (for large sample size) will be approximately Normally distributed with mean and standard deviation

$$E(\bar{X}) = \mu, \;\; SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \tag{1}$$

Back to candy example

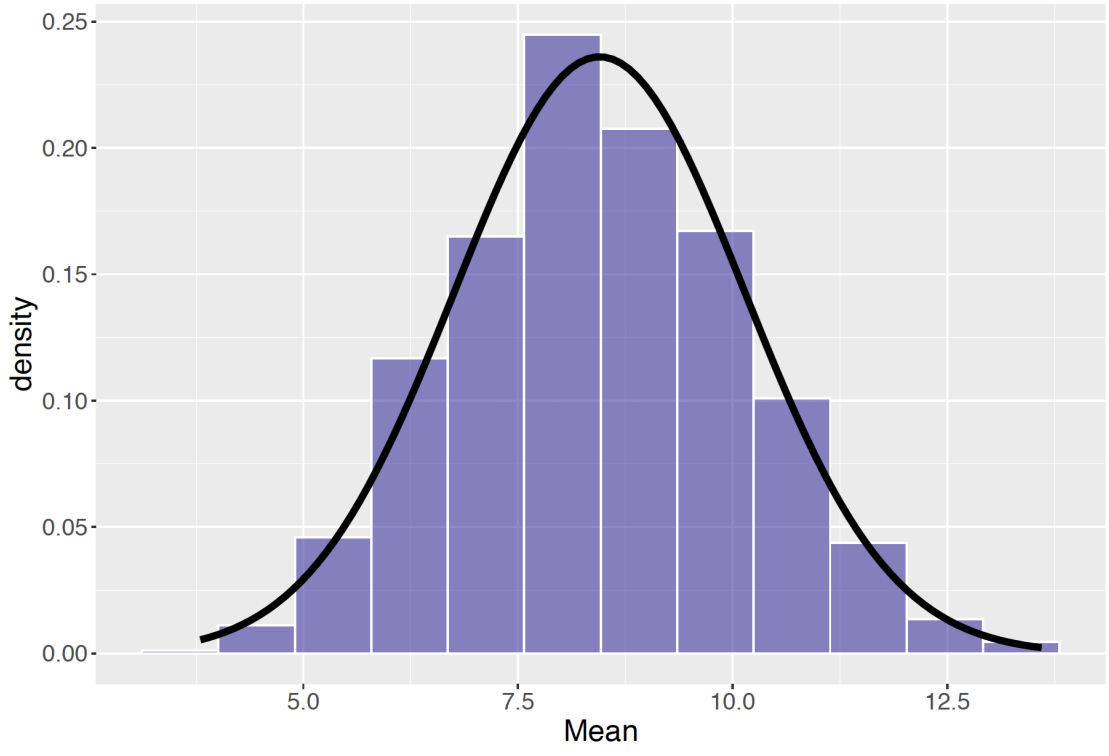
- Recall that the population of candy weights had a mean and standard deviation given by $\mu = 8.45$ and $\sigma = 5.36$.

If one takes samples of size $n = 10$, then, by this result, the sample mean $\bar{X}$ will be approximately Normal where

$$E(\bar{X}) = 8.45, \quad SD(\bar{X}) = \frac{5.36}{\sqrt{10}} = 1.69.$$

How good is the normal approximation?

- This Normal curve is drawn on top of the histogram of sample means.
- Approximation looks pretty good.



Two important points

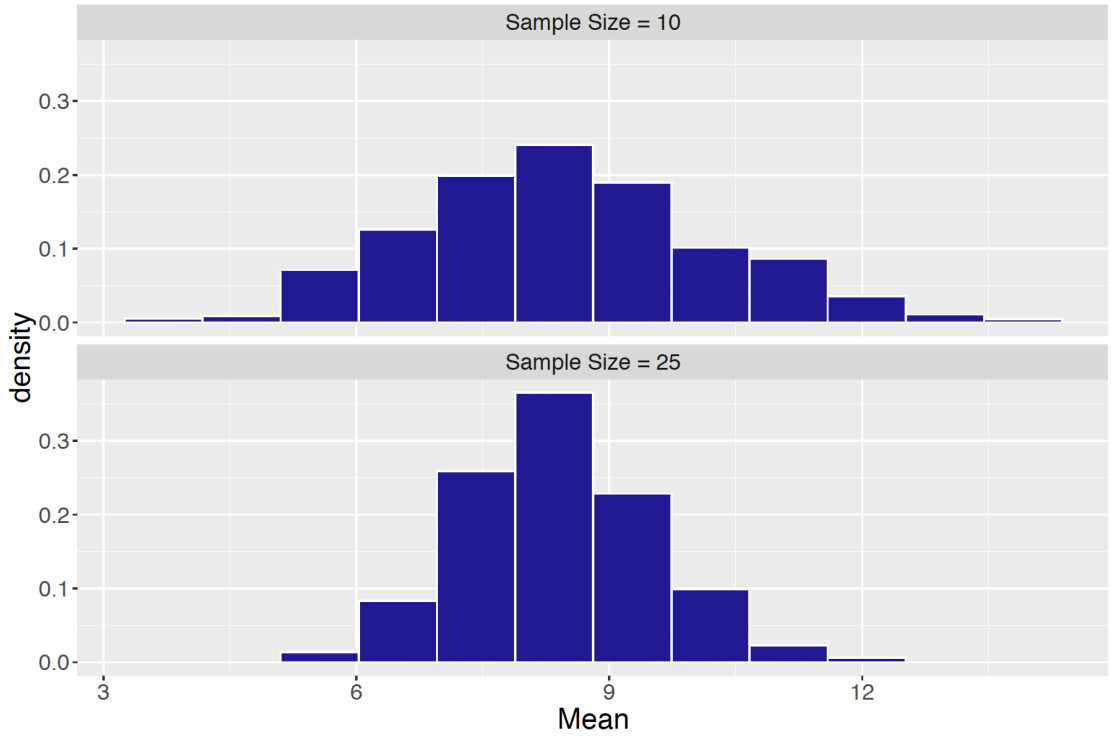
- First the expected value of the sample means, $E(\bar{X})$, is equal to the population mean μ.
- If one takes many random samples, then, on the average, the sample mean will be close to the population mean.
- Second, note that the spread of the sample means, as measured by the standard deviation, is equal to $\sigma/\sqrt{n}$.
- The spread of the sample means will be smaller than the spread of the population. Moreover, if one takes random samples of a larger size, then the spread of the sample means will decrease.

Look at different sample sizes

- We selected random samples of size $n = 10$ and computed the sample means. Suppose instead one selected repeated samples of size $n = 25$ – how does the sampling distribution of means change?

Compare the sample means for two values of n

- In a simulation, we show parallel histograms of sample means of the two sample sizes.



What do we see?

- What's the difference between sample means of size 10 and sample means of size 25?
- Both sets of sample means are Normally distributed with an average equal to the population mean.
- But the $n = 25$ sample means have a smaller spread.
- This means that as you take bigger samples, the sample mean $\bar{X}$ is more likely to be close to the population mean μ.

The Central Limit Theorem works for any population

- At one university, many of the students' hometowns are within 40 miles of the school. Also many students have homes between 80-120 miles of the university.

Distance from home

- Letting X denote "distance from home", imagine that the population of distances is described by the continuous density curve below.

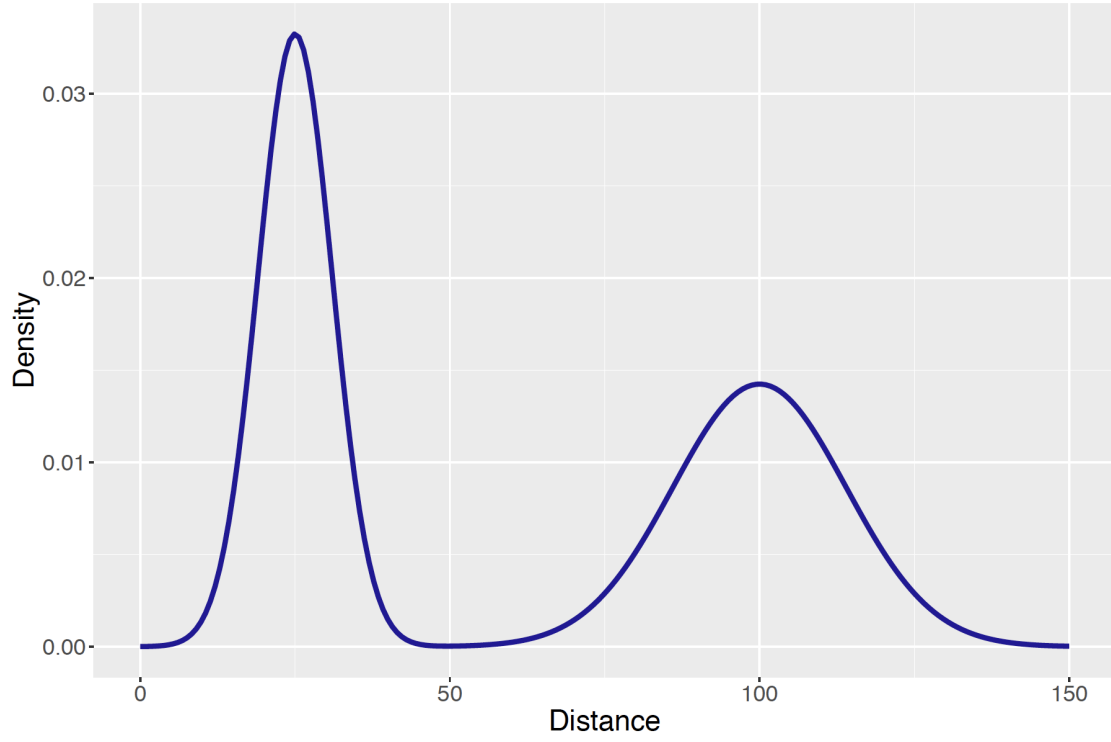


Figure 1: Density curve of the population of distances.

Sampling Students

- Take a random sample of n students from this population and computes the sample mean from this sample.
- If this sampling process is repeated many times, what will the distribution of sample means look like?
- Also, what is the effect of the sample size n?

A Simulation

- The computer simulated repeated samples of sizes $n = 1, n = 2, n = 5$, and $n = 20$. The histograms show the distributions of sample means for the four sample sizes.

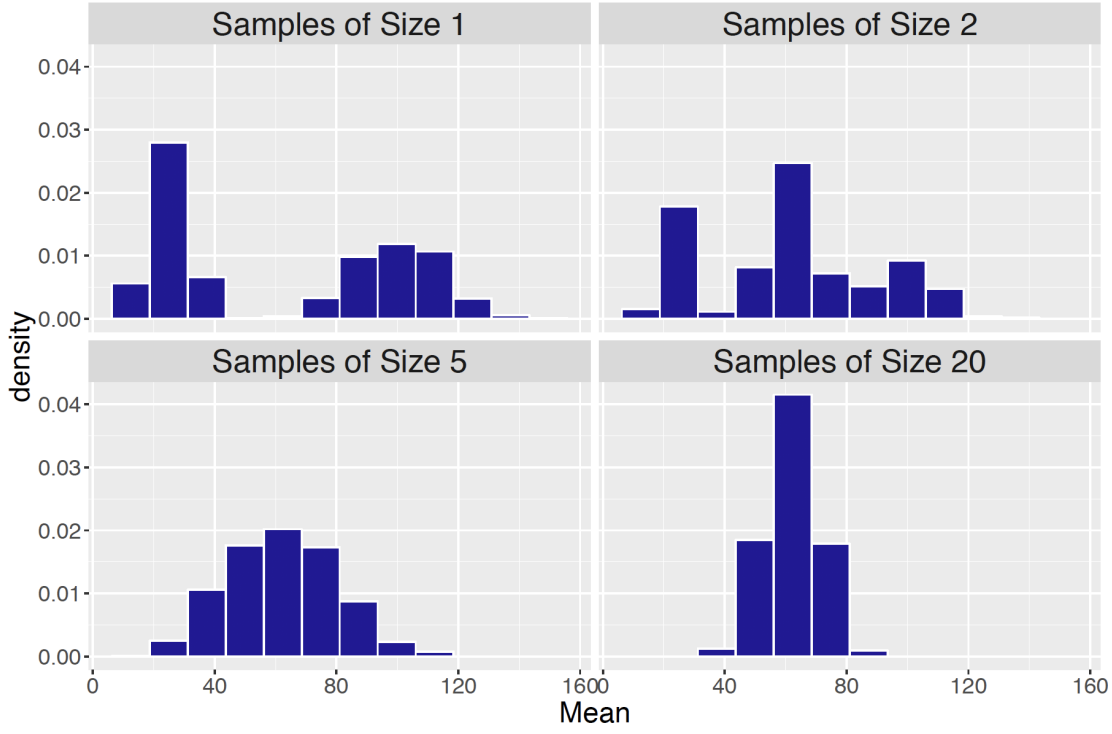


Figure 2: Histograms of random samples of distances, with sample sizes of $n = 1, n = 2, n = 5$, and $n = 20$.

What Do We See?

- Iff samples of size 1 are selected, our sample means look just like the original population.
- If samples of size 2 ares selected, then the sample means have a funny three-hump distribution.
- As one takes samples of larger sizes, the sampling distribution of means looks more like a Normal curve.
- This is what one expects from the Central Limit Theorem result – no matter what the population shape, the distribution of the sample means will be approximately Normal for a large sample size.

Distribution of Sample Means

- What is the distribution of the sample means when we take samples of size $n = 20$?
- By applying the Central Limit Theorem, the sample means will be approximately Normal with mean and standard deviation

$$E(\bar{X}) = \mu, \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \tag{2}$$

- Since one knows the mean and standard deviation of the population and the sample size, one just substitute these quantities and obtains

$$E(\bar{X}) = 60, \quad SD(\bar{X}) = \frac{41.6}{\sqrt{20}} = 9.3.$$

Using this result

- **What is the probability that a student's distance from home is between 40 and 60 miles?**
- This is a difficult question to answer exactly, since one does not know the exact shape of the population.
- Looking at the graph of the population, one sees that the curve takes on very small values between 40 and 60 miles.
- So this probability is close to zero – very few students live between 40 and 60 miles from our school.

Another question

- **What is the probability that, if one takes a sample of 20 students, the mean distance from home for these twenty students is between 40 and 60 miles?**
- That is, what is the chance that the sample mean falls between 40 and 60 miles?
- Since the distribution of $\bar{X}$ is approximately Normal with mean 60 and standard deviation 9.3, can use R.

```
pnorm(60, 60, 9.3) - pnorm(40, 60, 9.3)
```

```
## [1] 0.4842436
```

- Although it is unlikely for students to live between 40 and 60 miles from the school, it is pretty likely for the sample mean of 20 students to fall between 40 and 60 miles.

Another question

- **What is the probability that the mean distance exceeds 100 miles?**
- Here one wants to find the probability that $\bar{X}$ is greater than 100, that is $P(\bar{X} > 100)$. Using R, one computes

```
1 - pnorm(100, 60, 9.3)
```

```
## [1] 8.498565e-06
```

- This probability is essentially zero, which means that it is highly unlikely that a sample mean of 20 student distances will exceed 100 miles.