Chapter 11

Sampling Distributions
Sampling Terminology

◆ Parameter
  - fixed, unknown number that describes the population

◆ Statistic
  - known value calculated from a sample
  - a statistic is often used to estimate a parameter

◆ Variability
  - different samples from the same population may yield different values of the sample statistic

◆ Sampling Distribution
  - tells what values a statistic takes and how often it takes those values in repeated sampling
Parameter vs. Statistic

A properly chosen sample of 1600 people across the United States was asked if they regularly watch a certain television program, and 24% said yes. The *parameter* of interest here is the true proportion of all people in the U.S. who watch the program, while the *statistic* is the value 24% obtained from the sample of 1600 people.
Parameter vs. Statistic

- The mean of a population is denoted by \( \mu \) – this is a parameter.
- The mean of a sample is denoted by \( \bar{X} \) – this is a statistic. \( \bar{X} \) is used to estimate \( \mu \).
- The true proportion of a population with a certain trait is denoted by \( \rho \) – this is a parameter.
- The proportion of a sample with a certain trait is denoted by \( \hat{\rho} \) ("p-hat") – this is a statistic. \( \hat{\rho} \) is used to estimate \( \rho \).
The Law of Large Numbers

Consider sampling at random from a population with true mean $\mu$. As the number of (independent) observations sampled increases, the mean of the sample gets closer and closer to the true mean of the population.

($\bar{X}$ gets closer to $\mu$)
The Law of Large Numbers

Coin flipping:
The Law of Large Numbers

Rolling pair of fair dice.
Sampling Distribution

◆ The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size \((n)\) from the same population

– to describe a distribution we need to specify the shape, center, and spread

– we will discuss the distribution of the sample mean (x-bar) in this chapter
Case Study

Does This Wine Smell Bad?

Dimethyl sulfide (DMS) is sometimes present in wine, causing “off-odors”. Winemakers want to know the odor threshold – the lowest concentration of DMS that the human nose can detect. Different people have different thresholds, and of interest is the mean threshold in the population of all adults.
Case Study

Does This Wine Smell Bad?

Suppose the mean threshold of all adults is $\mu = 25$ micrograms of DMS per liter of wine, with a standard deviation of $\sigma = 7$ micrograms per liter and the threshold values follow a bell-shaped (normal) curve.
Where should 95% of all individual threshold values fall?

- mean plus or minus two standard deviations
  
  \[ 25 - 2(7) = 11 \]
  \[ 25 + 2(7) = 39 \]

- 95% should fall between 11 & 39

- What about the mean (average) of a sample of \( n \) adults? What values would be expected?
Sampling Distribution

What about the mean (average) of a sample of \( n \) adults? What values would be expected?

Answer this by thinking: “What would happen if we took many samples of \( n \) subjects from this population?” (let’s say that \( n=10 \) subjects make up a sample)

- take a large number of samples of \( n=10 \) subjects from the population
- calculate the sample mean (x-bar) for each sample
- make a histogram (or stemplot) of the values of x-bar
- examine the graphical display for shape, center, spread
Case Study

Does This Wine Smell Bad?

Mean threshold of all adults is $\mu=25$ micrograms per liter, with a standard deviation of $\sigma=7$ micrograms per liter and the threshold values follow a bell-shaped (normal) curve.

Many (1000) repetitions of sampling $n=10$ adults from the population were simulated and the resulting histogram of the 1000 $x$-bar values is on the next slide.
Case Study
Does This Wine Smell Bad?

Take many SRSs and collect their means $\bar{x}$.

Population, mean $\mu = 25$

The distribution of all the $\bar{x}$'s is close to Normal.

SRS size 10 $\rightarrow \bar{x} = 26.42$
SRS size 10 $\rightarrow \bar{x} = 24.28$
SRS size 10 $\rightarrow \bar{x} = 25.22$


Mean and Standard Deviation of Sample Means

If numerous samples of size $n$ are taken from a population with mean $\mu$ and standard deviation $\sigma$, then the mean of the sampling distribution of $\bar{X}$ is $\mu$ (the population mean) and the standard deviation is: $\frac{\sigma}{\sqrt{n}}$ (\(\sigma\) is the population s.d.)
Mean and Standard Deviation of Sample Means

- Since the mean of $\bar{X}$ is $\mu$, we say that $\bar{X}$ is an unbiased estimator of $\mu$.

- Individual observations have standard deviation $\sigma$, but sample means $\bar{X}$ from samples of size $n$ have standard deviation $\frac{\sigma}{\sqrt{n}}$. Averages are less variable than individual observations.
Sampling Distribution of Sample Means

If individual observations have the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{X}$ of $n$ independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

“If measurements in the population follow a Normal distribution, then so does the sample mean.”
Case Study
Does This Wine Smell Bad?

Mean threshold of all adults is $\mu=25$ with a standard deviation of $\sigma=7$, and the threshold values follow a bell-shaped (normal) curve.

The distribution of sample means is less spread out.

$\frac{\sigma}{\sqrt{10}} = 2.21$

Observations on 1 subject
(Population distribution)

$\sigma = 7$
Central Limit Theorem

If a random sample of size $n$ is selected from ANY population with mean $\mu$ and standard deviation $\sigma$, then when $n$ is large the sampling distribution of the sample mean $\bar{X}$ is approximately Normal:

$$\bar{X} \text{ is approximately } N(\mu, \frac{\sigma}{\sqrt{n}})$$

“No matter what distribution the population values follow, the sample mean will follow a Normal distribution if the sample size is large.”
Central Limit Theorem: Sample Size

- How large must $n$ be for the CLT to hold?
  - depends on how far the population distribution is from Normal
    - the further from Normal, the larger the sample size needed
    - a sample size of **25 or 30** is typically large enough for any population distribution encountered in practice
    - recall: if the population is Normal, any sample size will work \((n \geq 1)\)
Central Limit Theorem:
Sample Size and Distribution of $x$-bar

- $n=1$
- $n=2$
- $n=10$
- $n=25$