Chapter 9

Sampling Distributions

The inferential methods we will learn in the coming chapters will be based on using information from a sample to reach a conclusion about the population. In order to use this information, we must develop an understanding of how sampling information varies from sample to sample. In this chapter, we will explore the behavior of sample statistics in repeated sampling and learn one of the most important theorems in Statistics: The Central Limit Theorem.

Objectives:

- Define Sampling Distributions
- Contrast bias and variability
- Describe the sampling distribution of a sample proportion (shape, center, and spread)
- Use a Normal approximation to solve probability problems involving the sampling distribution of a sample proportion
- Describe the sampling distribution of a sample mean
- State the Central Limit Theorem
- Solve probability problems involving the sampling distribution of a sample mean.

CASE STUDY

Building better batteries

Everyone wants to have the latest technological gadget. That's why iPods, digital cameras, PDAs, Game Boys, and camera phones have sold millions of units. These devices require lots of power and can drain traditional alkaline batteries quickly. Battery manufacturers are constantly searching for ways to build longer-lasting batteries. In July 2005, Panasonic began marketing its new Oxyride battery in the United States. According to the results of preliminary testing, Oxyride batteries produced more power and lasted up to twice as long as alkaline batteries.

Battery manufacturers must constantly measure battery lifetimes to ensure that their production process is working properly. Because testing a battery's lifetime requires the battery to be drained completely, the manufacturer wants to test as few batteries as possible. As part of the quality control process, the manufacturer selects a sample of batteries to test at regular intervals throughout

production. By looking at the results from the sample, the manufacturer can determine whether the entire batch of batteries produced meets specifications.

At a particular battery production plant, when the process is working properly, AA batteries last an average of 17 hours with a standard deviation of 0.8 hour. Quality control inspectors select a random sample of 30 batteries during each hour of production and then drain them under conditions that mimic normal use. Here are the lifetimes (in hours) of the batteries from one such sample:

16.91 18.83 17.58 15.84 17.42 17.65 16.63 16.84 15.63 16.37

15.80 15.93 15.81 17.45 16.85 16.33 16.22 16.59 17.13 17.10

16.96 16.40 17.35 16.37 15.98 16.52 17.04 17.07 15.73 16.74

Do these data suggest that the process is working properly?

In this chapter, you will develop the tools you need to help answer questions like this.

Parameter vs. Statistic

The usual way to gain information about a population characteristic is to select a sample from the population. However, we must note that the sample information we gather may differ somewhat from the population characteristic we are trying to measure. Further, the sample information may differ from sample to sample. This sample-to-sample variability poses a problem when we try to generalize our findings to the population. In order to do so, we must gain an understanding of this variability.

Parameter:
A number that describes a population

A parameter always exists but in practice we rarely know its value because of the difficulty in creating a census. Parameters always use Greek letters to describe them. For instance we know that μ represents the mean of a population and σ represents the standard deviation of the population. If we are talking about a percentage parameter, we use the Greek letter ρ (rho).
Often just called p

Notation →

Example 1: If we wanted to compare the IQ's of all American and Asian males, it would be impossible. But it is important to realize that $\mu_{\text{american male}}$ and $\mu_{\text{asian male}}$ exist.

Example 2: If we were interested in whether there is a greater percentage (proportion) of women who eat broccoli than men, we want to know whether $\rho_{\text{women}} > \rho_{\text{men}}$.

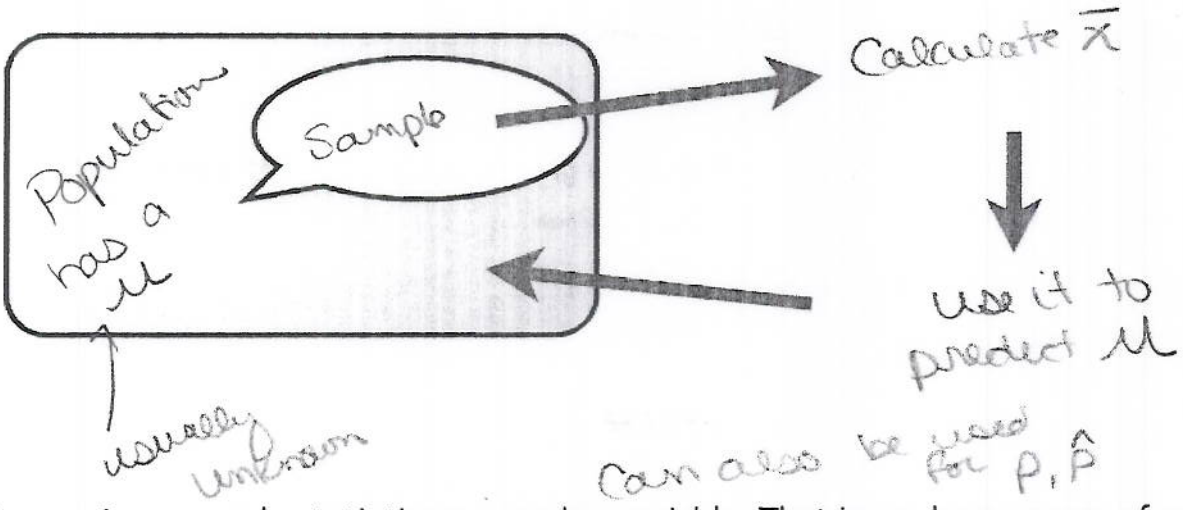
Statistic:
a number that describes a sample

The value of a statistics can always be found when we take a sample but it is important to realize that that statistic can change to sample to sample. Statistics use variables like $\bar{x}$, s, and $\hat{p}$ (non Greek). We often use statistics to estimate an unknown parameter.

Example 3: I take a random sample of 500 American males and find their IQ's. We find $\bar{x}$=103.2. We would like to be able to say that μ = 103.2. Obviously though, if we had taken different 500 males, we would have gotten a different $\bar{x}$. It is not clear that we can use $\bar{x}$ to find μ.

Example 4: I take a random sample of 200 men and find that 40 like broccoli. Then $\hat{p}_m$ = .2. From a sample of 300 women, I find that 30 like broccoli. Then $\hat{p}_w$ = .1. We know that $\hat{p}_m > \hat{p}_w$, but that is a far cry from being able to say that $\rho_m > \rho_w$.

Sampling Distribution



We can view a sample statistic as a random variable. That is, we have no way of predicting *exactly* what statistic value we will get from a sample, but, given a population parameter, we know how those values will behave in repeated sampling. If we could find *all possible* samples of a given size from a population, we could find the corresponding distribution of statistic values.

To understand why sampling variability is not fatal, we ask, "What would happen if we took many samples?" Here's how to answer that question:

- Take a large number of samples from the same population.
- Calculate the sample mean $\bar{x}$ or sample proportion $\hat{p}$ for each sample.
- Make a histogram of the values of x or $\hat{p}$.
- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

In practice it is too expensive to take many samples from a population like all adult U.S. residents. But we can imitate many samples by using simulation.

Example 5: Murphy's Law and tumbling toast If a piece of toast falls off your breakfast plate, is it more likely to land with the buttered side down? According to Murphy's Law (the assumption that if anything can go wrong, it will), the answer is "Yes." Most scientists would argue that by the laws of probability, the toast is equally likely to land butter-side up or butter-side down. Robert Matthews, science correspondent of the *Sunday Telegraph,* disagrees. He claims that when toast falls off a plate that is being carried at a "typical height," the toast has just enough time to rotate once (landing butter-side down) before it lands. To test his claim, Mr. Matthews has arranged for 150,000 students in Great Britain to carry out an experiment with tumbling toast.

Assuming scientists are correct, the proportion of times that the toast will land butter-side down is $p = 0.5$. We can use a coin toss to simulate the experiment. Let heads represent the toast landing butter-side down.

use Calcs

(a) Toss a coin 20 times and record the proportion of heads obtained, $\hat{p}$ = (number of heads)/20. Explain how your result relates to the tumbling-toast experiment.

What % of time does toast land butter side down

assign heads = butter side down

randint(0,1,20) → Li

sum Li Why?

find $\hat{p} = \frac{12}{20} = .6$ $\frac{4}{20}$

(b) Repeat this sampling process 10 times. Make a histogram of the 10 values of $\hat{p}$. Is the center of this distribution close to 0.5? yes

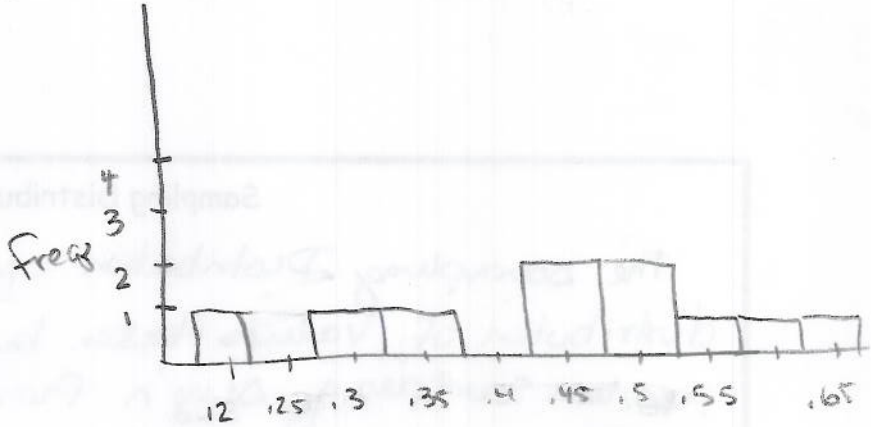
$\frac{4}{20} = .2$ $\frac{13}{20} = .65$

$\frac{11}{20} = .55$ $\frac{10}{20} = .5$

$\frac{7}{20} = .35$ $\frac{9}{20} = .45$

$\frac{9}{20} = .45$ $\frac{6}{20} = .3$

$\frac{10}{20} = .5$ $\frac{12}{20} = .6$



(c) Ten repetitions give a very crude approximation to the sampling distribution. Pool your work with that of other students to obtain several hundred repetitions. Make a histogram of all the values of $\hat{p}$. Is the center close to 0.5? Is the shape approximately Normal?

on board

(d) How much sampling variability is present? That is, how much do your values of $\hat{p}$ based on samples of size 20 differ from the actual population proportion, $p = 0.5$?

(e) Why do you think Mr. Matthews is asking so many students to participate in his experiment?

Law of Large #'s

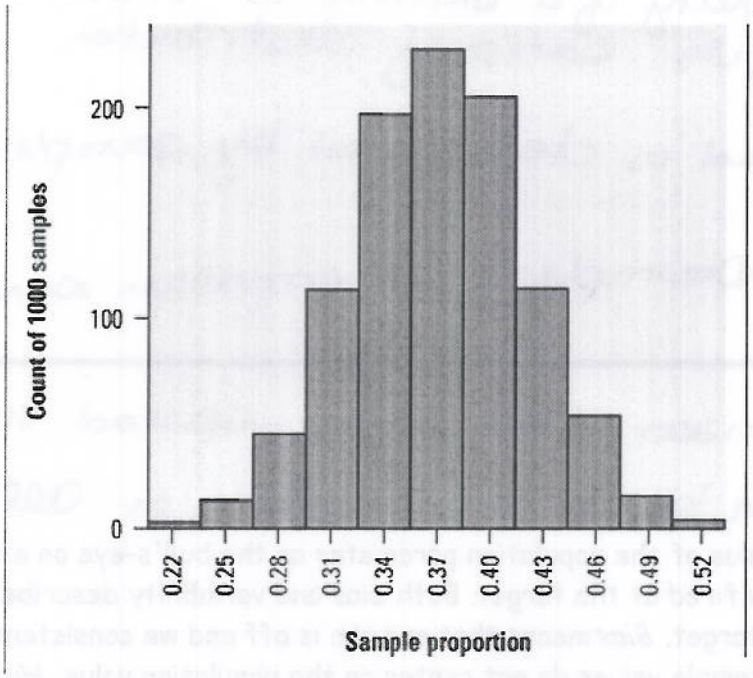
Sampling Distribution:

The sampling Distribution of a statistic is the distribution of values taken by the statistic in all possible samples of size n from the same population

Describing a sampling Distribution SOCS

* Describe overall shape
* Note center
* describe spread
* Outliers, any other unusual features

Example 6: Describe the sampling distribution below.



Proportions of samples who watched *Survivor :Guatemala* in samples of size n = 100

Symmetrical
approx Normal
Centered at 0.37
Spread .22 - .52

Sampling Bias and Variability

We have no way of knowing whether or not our statistic value is equal to the parameter we are trying to estimate. We must be aware of the bias and variability of our sampling distribution. Then we can use the information about the sample to reach a conclusion about the parameter.

Bias of a Statistic:

A Statistic Not a sampling method is biased.

*Concerns the center of the sampling dist. In other words, does our Statistic hit fairly close to the parameter

Unbiased Statistic/Unbiased Estimator:

A statistic used to estimate a parameter is unbiased if the mean of its sampling distribution is equal to the true parameter.

i.e. $\mu_{\bar{x}} = \mu$ and $\mu_{\hat{p}} = p$

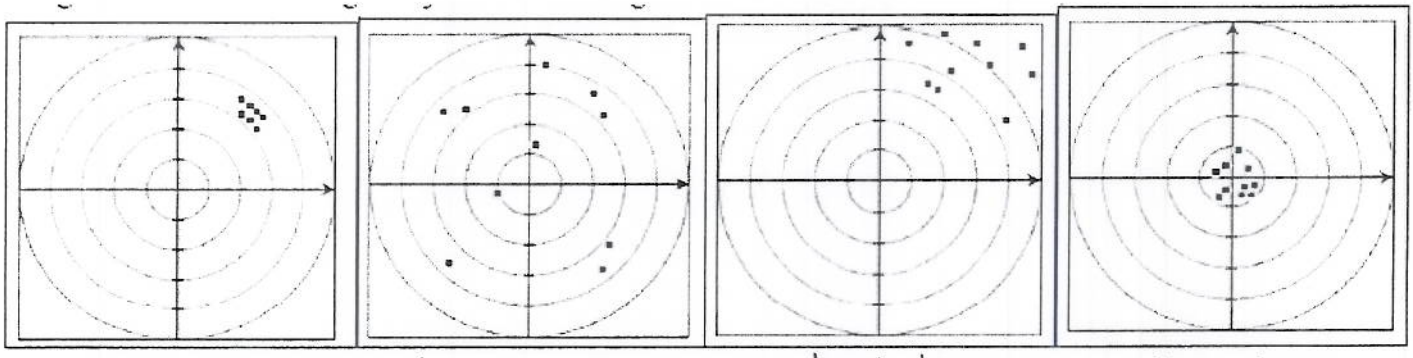
Variability of a Statistic:

* The variability of a statistic is described by the spread of its sampling distribution.
* The spread is determined by sample size
* Larger samples give smaller spreads

ex. Say n=1200 has same spread regardless of Population is all U.S. Residents or all San Fran residents

We can think of the true value of the population parameter as the bull's-eye on a target and of the sample statistic as an arrow fired at the target. Both bias and variability describe what happens when we take many shots at the target. *Bias* means that our aim is off and we consistently miss the bull's-eye in the same direction. Our sample values do not center on the population value. *High variability* means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results.

What does each target picture below in terms of variability and bias?



high bias
Low variability

Low bias
high variability

high bias
high variability

Low bias
Low Variability

Notes about bias and variability:

If we choose random samples of sufficient size the statistic will have the desired low bias + low variability