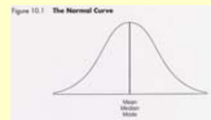
Sampling and Sampling Distributions

- Normal Distribution
- Aims of Sampling
- Basic Principles of Probability
- Types of Random Samples
- Sampling Distributions
- Sampling Distribution of the Mean
- Standard Error of the Mean
- The Central Limit Theorem

Review of the Normal Distribution

- **Normal Distribution**
- is a theoretical ideal distribution. Real-life empirical distributions never match this model perfectly.
- However, many things in life do approximate the normal distribution, and are said to be "normally distributed."



Scores "Normally Distributed?"

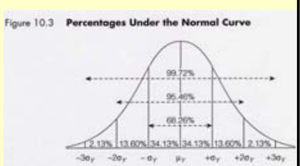
Table 10.1 Final Grades in Social Statistics of 1,200 Students (1983-1993)					
Midpoint Score	Frequency Bar Chart	Freq.	Cum. Freq. (below)	%	Cum % (below)
40	*	4	4	0.33	0.33
50	*******	78	82	6.5	6.83
60	***************	275	357	22.92	29.75
70	***********************	483	840	40.25	70
80	***************	274	1114	22.83	92.83
90	*******	81	1195	6.75	99.58
100	*	5	1200	0.42	100

- Is this distribution normal?
- There are two things to initially examine: (1) look at the shape illustrated by the bar chart, and (2) calculate the mean, median, and mode.

Example of Applying the Properties of the Normal Curve to a Population

If the mean grade is 70 and the standard deviation is 10:

- We can be confident that 68% of the students got a grade between 60 and 80
- We can be confident that 95% got a grade between 50 and 90.
- It's important to note that such conclusions can only be made when we have Population data and this is rarely the case. Typically, we must use sample data.



50 60 70 80 90

Standard Deviation = 10

Sampling

- **What is a Population?**
  - A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
- **What is a Sample ?**
  - A relatively small **subset** from a population..

Sampling

- **What is the aim of Sampling?**
  - to determine what is true of the population without having to question (or collect data on) the entire population.

Sampling

- **Parameter** – A measure (for example, mean or standard deviation) used to describe a **population distribution**.
- **Statistic** – A measure (for example, mean or standard deviation) used to describe a **sample distribution**.

Notation

Table 11.1 **Sample and Population Notations**

Measure	Sample Notation	Population Notation
Mean	$\bar{Y}$	μ_Y
Proportion	p	π
Standard deviation	S_Y	σ_Y
Variance	S_Y^2	σ_Y^2

Population inferences can be made...

...by selecting a representative sample from the population

Probability Sampling

A method of sampling that enables the researcher to specify for each case in the population the probability of its inclusion in the sample.

Typically, every case has an equal chance of being selected for the sample.

Probability Sampling: Simple Random Sampling

A sample designed in such a way as to ensure that:

every member of the population has an **equal chance of being chosen**

(This can be done using a table of random numbers, computer, or other means; Appendix A in your book provides a Table of Random Numbers)

Probability Sampling: Systematic Random Sampling

A method of sampling in which **every Kth member in the total population is chosen** for inclusion in the sample (for example every 10th member).

To determine the very first case selected use simple random sampling (e.g., if the skip interval is ten, use simple random sampling to choose the first case among the first 10 cases in the population).

Systematic Random Sampling

Figure 11.2 **Systematic Random Sampling**

From a population of 40 students, let's select a systematic random sample of 8 students. Our skip interval will be 5 (40 ÷ 8 = 5). Using a random number table, we choose a number between 1 and 5. Let's say we choose 4. We then start with student 4 and pick every 5th student:

Our trip to the random number table could have just as easily given us a 1 or a 5, so all the students do have a chance to end up in our sample.

Probability Sampling: Stratified Random Sampling

A method of sampling obtained by:

(1) **dividing the population into subgroups** based on one or more variables central to our analysis and

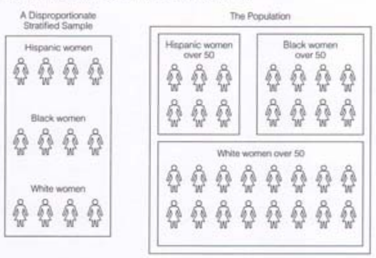
(2) then drawing a simple random **sample from each of the subgroups**

Probability Sampling: Stratified Random Sampling

- **Proportionate stratified sample** – The size of the sample selected from each subgroup is proportional to the size of that subgroup in the entire population.
- **Disproportionate stratified sample** – The size of the sample selected from each subgroup is disproportional to the size of that subgroup in the population.

Disproportionate Stratified Sample

Figure 11.3 **A Random Sample Stratified by Race/Ethnicity**



Predicting the Population Based on a Random Sample

A Statistical Dilemma:

How much confidence can we have that our sample estimates reflect the parameters of the larger population?

There are Two Distributions That Help Us Estimate Our Confidence in the Sample Statistic

1. The actual distribution of scores for a variable in a **sample** of the population is a **sample distribution**. We use statistics from the sample to help us **estimate** population parameters.

2. The **sampling distribution** is a theoretical distribution of all possible sample estimates of the population parameter in which we are interested (we will be examining this much closer).

Sample Distribution

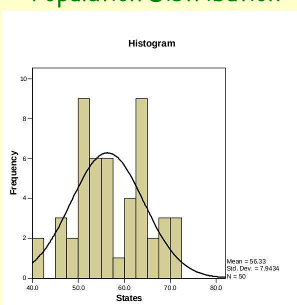
Illustration

To illustrate the distributions, let's assume our **population** is the nation's 50 states.

Our variable is "the percentage of eligible voters in the state who voted in the 1992 election" or "**%Voted**".

Population Distribution

\# of States



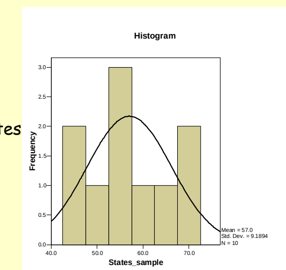
% of Eligible Voters who voted in 1992

Output

	N	Mean	Std. Deviation
%Voted	50	56.33	7.94

Sample Distribution of a Simple Random Sample of 10 States

\# of States



% of Eligible Voters who voted in 1992

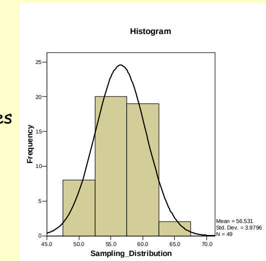
Output

	N	Mean	Std. Deviation
%Voted	10	57.0	9.19

Sampling Distribution

Sampling Distribution of the Mean

\# of samples



% of Eligible Voters who voted in 1992: sample means

Sampling Distribution

- Variables that don't have a normal distribution, do have a normal **sampling** distribution of their parameters such as the mean.
- If we take a die and role it 100 times, what will the normal distribution look like?
- If we take a die and roll it so that we have 100 sample means, what will the **sampling** distribution look like?

Sample Size

As the **sample size increases** the **sample estimates** more closely reflect the **population parameters** and:

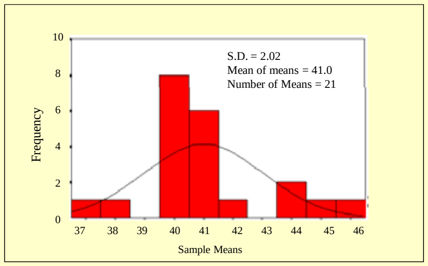
the **sample distribution** more closely reflects the **sampling distribution**

This includes both the sample **mean** and sample **standard deviation**.

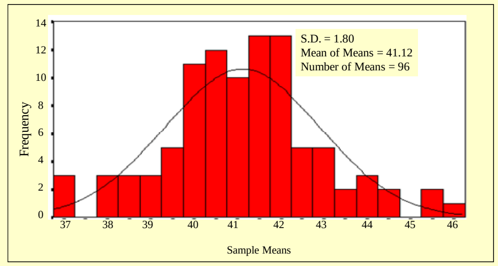
In sum a Sampling Distribution is:

- a probability distribution of all possible sample values of the **population** parameter of interest.
- **Sampling distributions** are never really observed (and consequently are considered "**theoretical**")
- To better understand the concept of the **sampling** distribution, using a limited number of samples, let's illustrate how one could begin to generate such a distribution.

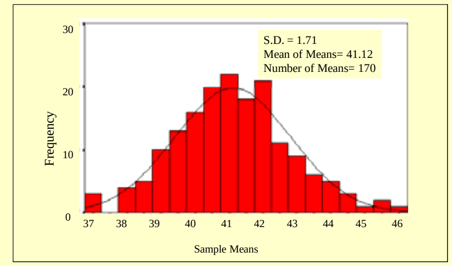
Here is a Sampling Distribution of Sample Means with 21 Samples



Here is a Sampling Distribution of Sample Means with 96 Samples



Here is a Sampling Distribution of Sample Means with 170 Samples



The Central Limit Theorem summarizes what we have learned:

- If **all possible random samples** of size N are drawn from a population then, as the number of samples increases, the **sampling distribution** of a statistic (such as the mean) becomes approximately normal.

A Single Sample and the Sampling Distribution

If we take only a single **random** sample, and the sample size is **large** (50 is okay, 150 is better), then we can assume that the sample distribution will be very similar to the population distribution and also the **sampling distribution** (this is referred to as the **Law of Large Numbers**).

Therefore, the properties of the **sampling distribution** can be applied to our single, large random sample.

The Central Limit Theorem (continued)

Characteristics of the sampling distribution include:

--68% of the sample means fall within ± 1 standard error of the average or mean of the means (the SE is similar to the SD and will be discussed further)

--95% fall within ± 1.96 standard errors of the mean

A Single Sample and the Sampling Distribution

Thus, with a single, large random sample we can identify **confidence intervals** within which our population parameter is likely to fall.

Sampling Distribution

In sum, properties of the **"sampling distribution"** tell us that the distribution of multiple sample statistics (such as the mean) is likely to be normal (have a **normal distribution)**.

Consequently, we can use the properties of the **normal distribution** to help us determine our level of confidence that our sample statistic reflects the population parameter.

Applying Properties of the Sampling Distribution:

Since the distribution of a single large sample is very similar to the sampling distribution and we don't have the actual sampling distribution, we use a single sample in place of the sampling distribution.

We can use the number of cases and the standard deviation of a single sample to **calculate the standard error of the sampling distribution and subsequently the level of confidence** for our sample statistics.

Calculating the Level of Confidence: An Example

1. We **take a sample** of 100 new Assistant Professors of sociology and determine each person's income.
2. In our sample, the **mean** income is $50,000 (for nine months) and the **standard deviation** is $7,000.

In this example, we want to know **how much confidence can we have** that our sample mean income reflects the mean for the whole population of new sociology Assistant Professors.

Properties of the sampling distribution tell us that by:

(1) calculating the standard error of the mean and then

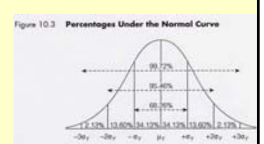
(2) applying it to the normal curve (much like we did for the population's standard deviation) we can:

(3) determine levels of confidence in our sample statistic.

Example of Applying the Properties of the Sampling Distribution

If the sample mean is $50K and the standard deviation is $7K for a sample of 100:

- **We first calculate the standard error and then apply it to our problem** (we will learn how to calculate the SE in the next chapter). In this case 1 SE = $700; 2 SE = $1,400
- We know that 95% of the sample means would fall between two standard errors of the mean (actually 1.96 not 2):
- We can be 95% confident that the average income ranges between $48,600 (50,000 – 1,400) and $51,400 (50,000 + 1,400)



48.6K 50K 51.4K

2 Standard Errors = $1,400

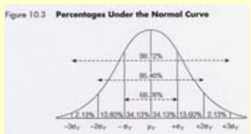
Example of Applying the Properties of the Sampling Distribution

If the sample mean is $50K, the standard deviation is $7K, and N=100:

- We can be 99% confident that the **average income of the population** ranges somewhere between

  $47,900 (50,000-2,100) and $52,100 (50,000 + 2,100).

- Our confidence intervals at the 99% level are $47.9K and $52.1K.



47.9K 50K 52.1K

3 Standard Errors = $2,100

(to be precise it is 2.54 rather than 3 * SE)

In sum, properties of the sampling distribution tell us that:

We can use a sample mean and standard deviation to calculate a **standard error** and subsequently identify the **level of confidence** we have in our sample findings.

What exactly is a sampling distribution?

And

What is standard error?

www.ruf.rice.edu/~lane/stat_sim

Review Homework

The Central Limit Theorem

σ_y σ_y / N