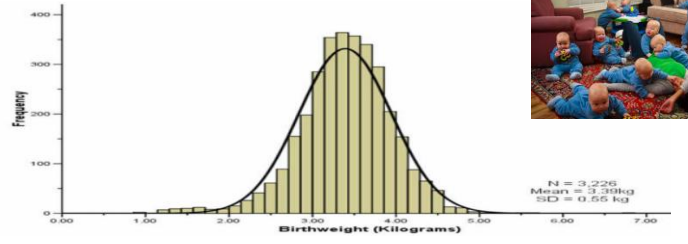
LAST CLASS: Chapter 4

Working with the normal curve..

5-1



5-2

2.6 kg

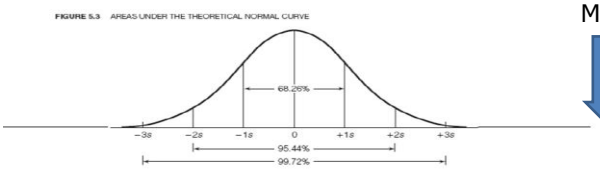
7.0 kg

Can convert using Z scores!!!

Can convert using Z scores!!!

Theoretical Normal Curve

- In Statistics we work with a "Theoretical distribution" meant to represent a normal distribution



Many SD above the mean..

- The mean is assigned a value of 0
- The standard deviation is assigned a value of 1
- Describe this distribution in terms of Z scores (Standard score
- A Z score of 1 is 1 standard deviation above the mean,..
- A Z score of -1 is 1 standard deviation below the mean,.. etc.

Between	Lies
±1 standard deviation	68.26% of the area
±2 standard deviations	95.44% of the area
±3 standard deviations	99.72% of the area

- Z scores -> also called “standard scores”

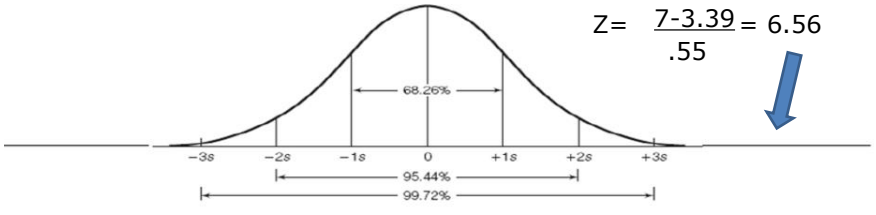
$$Z = \frac{X_i - \overline{X}}{s}$$

- Are useful in determining the exact location of any value as observed X_i in terms of this theoretical normal curve
- Correspondingly, can use Z scores to find the corresponding “proportions” of area under the curve associated with specific values
- Can be translated into percentages or probabilities.

Theoretical Normal Curve

- In Statistics we work with a “Theoretical distribution” meant to represent a normal distribution

FIGURE 5.3 AREAS UNDER THE THEORETICAL NORMAL CURVE



Z= $\frac{7-3.39}{.55}$ = 6.56

- The mean is assigned a value of 0
- The standard deviation is assigned a value of 1
- Describe this distribution in terms of Z scores
- A Z score of 1 is 1 standard deviation above the mean,..
- A Z score of -1 is 1 standard deviation below the mean,.. etc.

Between	Lies
±1 standard deviation	68.26% of the area
±2 standard deviations	95.44% of the area
±3 standard deviations	99.72% of the area

If we assume that a distribution is "normally distributed",.. We can use the Z scores to say all sorts of things about the distribution

We an use Appendix A to Describe Areas Under the Normal Curve

-> area between a *Z* score and the mean (Section 4.3)..
-> area either above or below a *Z* score (4.4)
-> area between two *Z* scores (4.5)
-> probability of randomly selected score (4.6)

5-7

Ultramarathon: 100 mile race

Mean is 680 minutes
Standard deviation is 30 minutes

What proportion ran it in less 630 minutes?

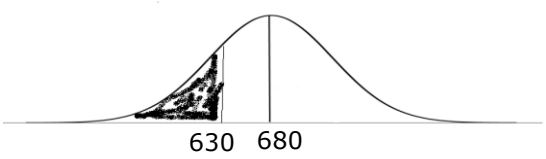
5-8

Another example

- Example:
- Mean = 680 minutes; s= 30;
- What proportion less than 630?
- Xi = 630

$$Z = \frac{X_i - \overline{X}}{s}$$

- 1. Draw it:



-
- 2. Find Z score: Z = (630-680)/30 = - 1.67

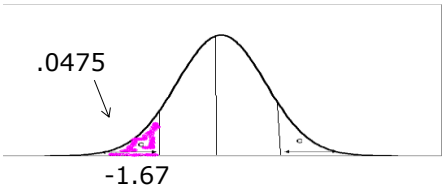
3. To find the area in the tail of a distribution, we use column c in Appendix A

AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) *Z*	(b) Area between Mean and *Z*	(c) Area beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	⋮

1.67 0.4525 0.0475

- 3. To find the area **below** a **negative** score we use column c in Appendix A



- The area below a *Z* score of -1.67 is 0.0475

- 4. Interpret :
- The proportion of all competitors who ran faster is .0474.

5-11

- Also:
- Can estimate the area between any 2 scores on the distribution:

- On opposite sides of the mean, merely add the areas (column b)
- On the same side of the mean, subtract the smaller area (column b) from the larger area (column b)

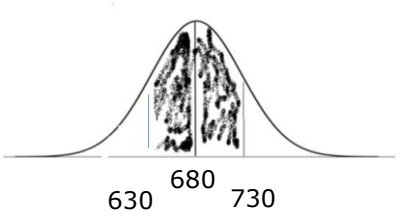
5-12

Briefly, another example

- Two persons run it with time of **730** minutes & **630** minutes in a distribution of times where the mean = **680** minutes and *s* = **30** minutes. What percentage run a time between these two results?

$$Z = \frac{X_i - \overline{X}}{s}$$

- 1. Draw it:



- 2. The two Z scores are:

$$Z = \frac{630 - 680}{30} = -1.67 \qquad Z = \frac{730 - 680}{30} = +1.67$$

3. Consult the appendix to find the relevant areas under the curve…

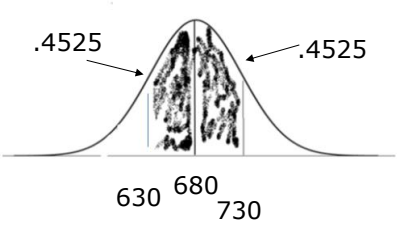
TABLE 5.3 AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) *Z*	(b) Area between Mean and *Z*	(c) Area beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	⋮

1.67 0.4525 0.0475

5-14

3. Consult the appendix to find the relevant areas under the curve…



4. Interpret :

(.4525 + .4525) *100

About 90.50% of all competitors ran the race between the times of 630 and 730.

5-15

- Also:
- Can estimate the area between any 2 scores on the distribution:

- On the same side of the mean, subtract the smaller area (column b) from the larger area (column b)

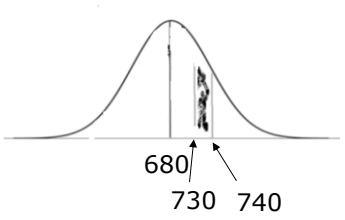
5-16

Briefly, another example

- Two persons run it with time of **730** minutes & **740** minutes in a distribution of times where the mean = **680** minutes and *s* = **30** minutes. What percentage run a time between these two results?

$$Z = \frac{X_i - \overline{X}}{s}$$

- 1. Draw it:



- 2. The two Z scores are:

$$Z = \frac{730 - 680}{30} = 1.67 \qquad Z = \frac{740 - 680}{30} = +2.00$$

3. Consult the appendix to find the relevant areas under the curve…

TABLE 5.3 AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) *Z*	(b) Area between Mean and *Z*	(c) Area beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	⋮

1.67	0.4525	0.0475
2.00	0.4772	0.0228

5-18

Find the two corresponding areas:

1.67 is .4525
2.00 is .4772, and subtract the smaller from the larger..

.4772-.4525 = .0247,..

Interpretation:
Only 2.47% of all competitors ran a time between 730 and 740 minutes

Very briefly, "areas under the curve can also be read as probabilities".

Read section 4.6

What is the probability of …

Probability is the measure of the likelihood that an event will occur.

Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).

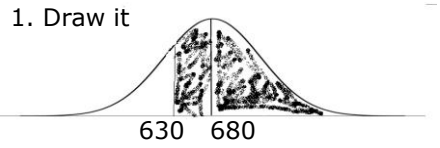
Ultramarathon: 100 mile race

Mean is 680 minutes
Standard deviation is 30 minutes

What is the "**probability**" that someone ran it in slower than 630 minutes?

Exact same steps: slightly different interpretation..

1. Draw it



2. Z score is: (630-680)/30= -1.67

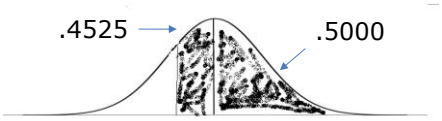
3. Consult the appendix to find the relevant areas under the curve…

TABLE 5.3 AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) Z	(b) Area between Mean and Z	(c) Area beyond Z
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	⋮

1.67	0.4525	0.0475
2.00	0.4772	0.0228

4. Find the appropriate area, and interprete



The probability is .9525 that somebody will run it slower

5-23

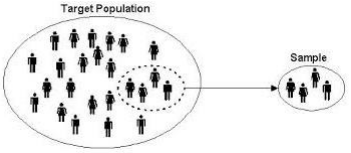
That's it for Chapter 4,..

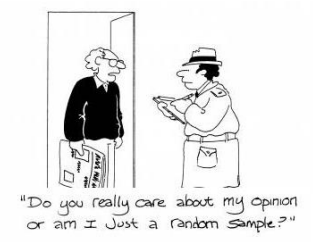
Now onto Chapter 5

Note, this chapter is rather theoretical.. Applications next week…

Chapter 5

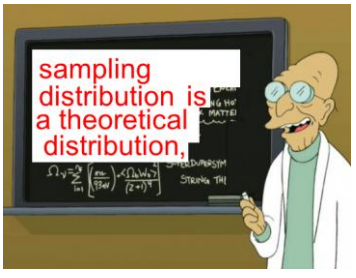
Introduction to Inferential Statistics: Sampling and the Sampling Distributions



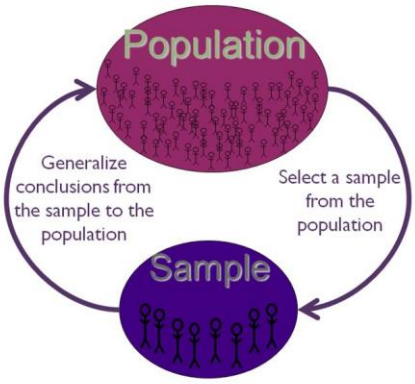


In this presentation you will learn about:

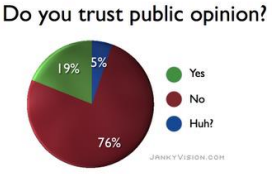
- Learn to distinguish between the:
- (i) population distribution
- (ii) sample distribution
- (iii) sampling distribution



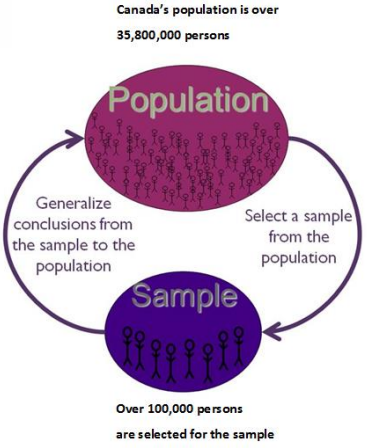
6-26

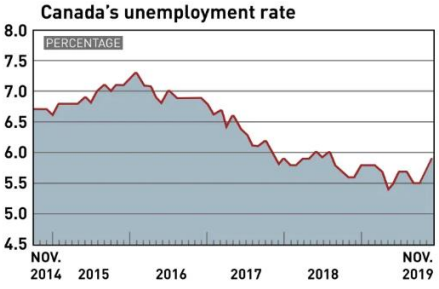


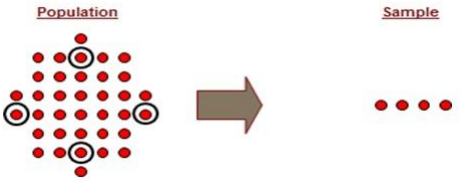
Many examples:
Political pollsters
Statistics Canada
Public Opinion Research
Market Research



Canadian Labour Force Survey







Sampling:
Most straight forward method is that of simple random selection (SRS)

Purely random, i.e. nobody in your population has a higher or lower chance of selection..

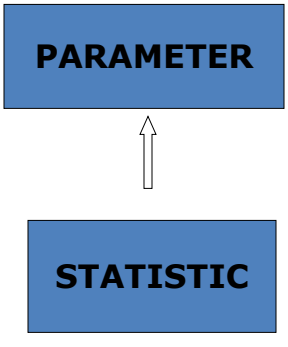
E.g. All Canadian households have an equal probability of being selected into the Canadian Labour Force Survey

Computer software which allows for random selection..

5-29

Parameter and Statistic

- ***Statistics*** are mathematical characteristics of samples.
- ***Parameters*** are mathematical characteristics of populations.
- **Statistics** are used to estimate **parameters**.



Note:
To the extent that we have a difference between the parameter and the statistic,.. we potentially have "sampling error"

6-30

Basic Logic of Estimation

- In estimation procedures, *statistics* calculated from random samples are used to estimate the value of population *parameters*, **with a varying level of success depending on:**
- **sample size and corresponding sampling error**
- ***Information on error is implied in what are referred to as "SAMPLING DISTRIBUTIONS" *to be introduced in this class…***
- ***As will be discussed:***
- ***Sampling distributions, with relatively large "standard errors" indicate lots of sampling error!!***

Basic Logic of Estimation

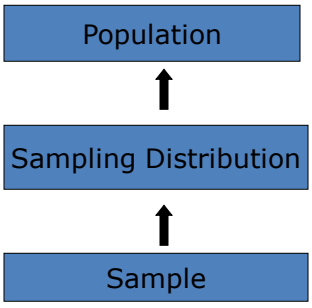
- Sampling error:
  merely the error associated with your unique sample, "by mere chance" .. discrepancy between your sample statistic and the population parameter

Example:

- *Sample 1000 Canadians (SRS).. Estimate the mean age..*
- *Not likely to be identical to the "parameter"… although likely quite close..*
- *Second sample of 10,000 Canadians (SRS).. Estimate the mean age*
- *Not likely to identical to "parameter", although likely quite close, and even closer than a sample of 1000*

The 3 types of distributions in Inferential Statistics

- Every application of inferential statistics involves 3 different distributions.
  - **Population Distribution** empirical; typically unknown
  - **Sampling Distribution:** non-empirical; known via theory
  - **Sample Distribution:**
  - empirical; known through observation



Information from the sample is linked to the population via the sampling distribution.

6-33

Population Distribution

- **A distribution on a variable for the full population that you are studying**

- **DEFINITION** of '**Population**':
- The entire pool from which a **statistical** sample is drawn. The information obtained from the sample allows statisticians to develop hypotheses about the larger **population**.
- E.g. All Canadians, All Ontario Residents,
- All Kings students…

- **Typically "unknown" (particularly for the larger populations)**
- Why? Researchers gather information from a sample because of the difficulty of studying the entire **population**.

6-34

Example:
Let's assume that you want to document how Canadians plan on voting in the next Federal Election

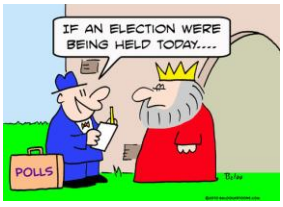
Population: All Canadians.

Can we obtain the Population Distribution in our research?

Obviously not, because we can't contact everyone!!!

Population distribution (unknown)

We estimate via a sample!!!

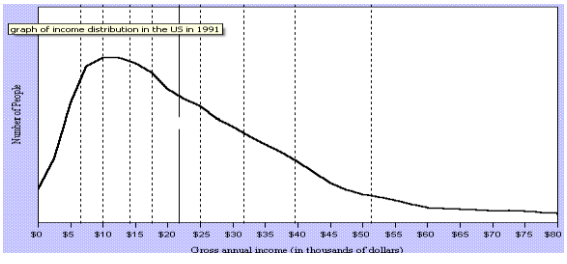


Interesting exception:

The Canadian Census allows us to document a "population distribution"
ALL Canadian households are directly contacted
Cost: 100's of millions of dollars

Advantage: Remarkable accuracy and completeness
Detailed data for all municipalities, regions, neighborhoods in Canada

Example: Distribution of Income in the US, 1991 Census



Other noteworthy exceptions: If your targeted population is relatively small: All Employees at Kings (250 persons); No need to necessarily sample

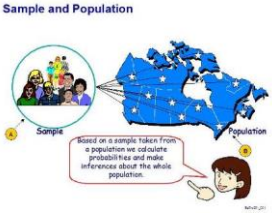
5-36

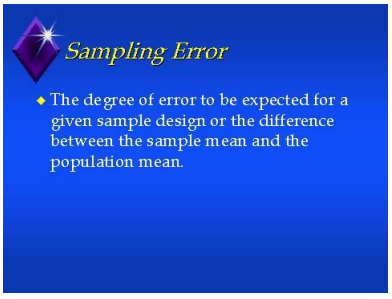
Sample Distribution

A distribution on a variable for a sample that we have selected using random methods so that the sample is "non-biased and representative!!

Typically "known" through our empirical research
E.g. -> sample (N=20000), interview 20000 respondents,.. and document income characteristics

Inevitably, we don't always Get it right!!
Issue of "sampling error"..

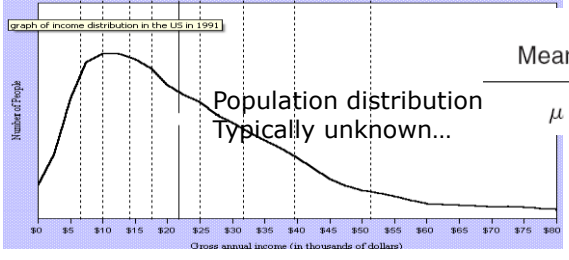




5-37

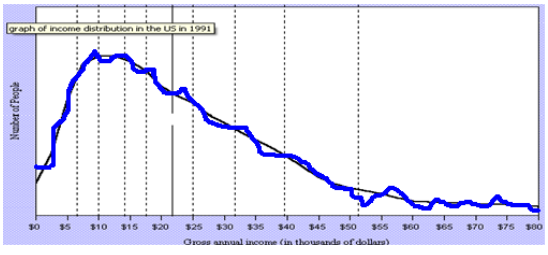
Different Distributions

- **NOTE: IF WE DO OUR RESEARCH PROPERLY> THE SHAPE OF THE SAMPLE DISTRIBUTION SHOULD BE QUIT SIMILAR TO THE SHAPE OF THE POPULATION DISTRIBUTION (IE> UNBIASED MEASUREMENT AND SAMPLING)**



Mean	Standard Deviation	Proportion
μ	σ	P_u

Difference between the two typically involves what is referred to as "sampling error"

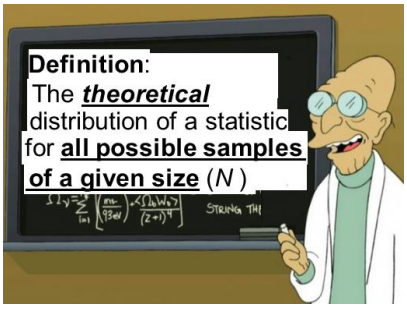


Mean	Standard Deviation	Proportion
$\overline{X}$	s	P_s

6-38

The 3rd type of distribution: sampling distribution

- The single most important concept in inferential statistics (very different from the sample and population distribution)



- The statistic could be almost anything:
- e.g. the “Mean”; or a “proportion”, etc.
- What do I mean by “***all possible samples***” of size N?

The Sampling Distribution

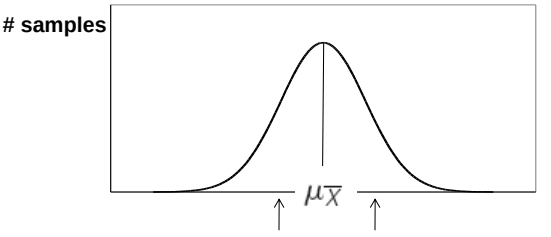
- E.g. Assume that we want the “sampling distribution” of a “mean income” with a sample of size (N) from the Canadian population…

- **Theoretically,.. assume that we:**
- ***use SRS -> obtain first sample (of size N)***
- ***calculate mean for this sample (alternatively, we could calculate a proportion or any other type of statistic)***
- **Repeat process:**
- **SRS -> 2nd sample (of size N): calculate mean**
- **SRS -> 3rd sample (of size N): calculate mean..**
- **Repeat again and again and again, until we have all potential unique samples (many many samples, right)**
- **-> this gives us our sampling distribution of this statistic (mean)**

6-40

The Sampling Distribution: Properties

1. Normal in shape



Innumerable different samples have many means

2. This sampling distribution (of means in this case) should have a mean across all samples $\mu_{\overline{X}}$ equal to the population mean μ (if unbiased)

3. The sampling distribution has a standard deviation (called the standard error) equal to the population standard deviation, *σ,* divided by the square root of *N*. $\sigma_{\overline{X}}$

The Sampling Distribution: Properties

4. Note: we can have a sampling distribution of means or a sampling distribution of proportions, as well as many other statistics…

5. Following from the formula for the "standard error", the larger the sample size (N), the smaller the standard error..

   i.e. the narrower the sampling distribution

   the less error in the statistic!!!

And this implies, if we have a larger sample, we have a lower level of error (less sampling error)..

Also referred to as an estimate that has more "efficiency"

Summary

TABLE 6.4 SYMBOLS FOR MEANS AND STANDARD DEVIATIONS OF THREE DISTRIBUTIONS

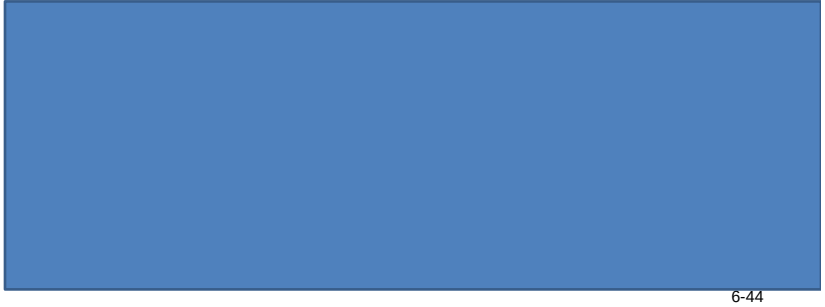
	Mean	Standard Deviation	Proportion
1. Samples	$\overline{X}$	s	P_s

6-43

Summary

TABLE 6.4 SYMBOLS FOR MEANS AND STANDARD DEVIATIONS OF THREE DISTRIBUTIONS

	Mean	Standard Deviation	Proportion
1. Samples	$\overline{X}$	s	P_s
2. Populations	μ	σ	P_u



6-44

Summary

TABLE 6.4 SYMBOLS FOR MEANS AND STANDARD DEVIATIONS OF THREE DISTRIBUTIONS

	Mean	Standard Deviation	Proportion
1. Samples	$\overline{X}$	s	P_s
2. Populations	μ	σ	P_u
3. Sampling distributions			
of means	$\mu_{\overline{X}}$	$\sigma_{\overline{X}}$	
of proportions	μ_p	σ_p	

6-45

Summary

TABLE 6.4 SYMBOLS FOR MEANS AND STANDARD DEVIATIONS OF THREE DISTRIBUTIONS

	Mean	Standard Deviation	Proportion
1. Samples	$\overline{X}$	s	P_s
2. Populations	μ	σ	P_u
3. Sampling distributions			
of means	$\mu_{\overline{X}}$	$\sigma_{\overline{X}}$	
of proportions	μ_p	σ_p	

Note: The standard deviations of the sampling distribution are called "standard error"

6-46

Making sense of our Sampling Distribution:

- As previously mentioned, the standard error is given as:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

- Typically we don't have anything on the mean or standard deviation of the population, right?

Formula (σ unknown)

Fortunately, we can approximate the standard error by using the sample standard deviation rather than the population standard deviation

In other words, on the basis of "one" sample, we can approximate the shape of our "corresponding sampling distribution"!!!!!!!

The Sampling Distribution: First Theorem

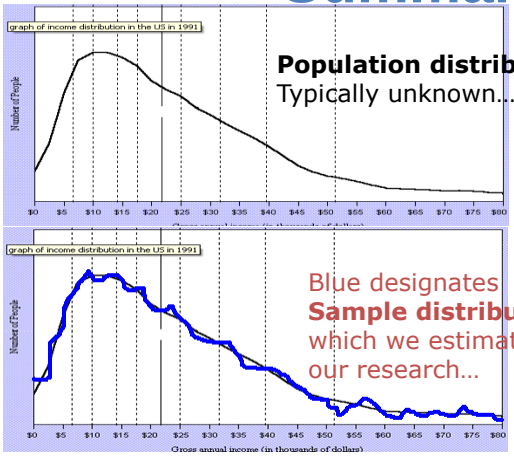
- If we begin with a trait that is ***normally*** distributed across a population (height, weight) and take an infinite number of equally sized random samples from that population, the sampling distribution of sample means will be normal.

The Sampling Distribution: Second (Central Limit) Theorem

– Again, for any trait or variable, even those that are not normally distributed in the population (e.g. income), as sample size grows larger, the sampling distribution of sample means will become normal in shape.

If repeated random samples of size *N* are drawn from ***any*** (e.g., normal or non-normal) shaped population with mean μ and standard deviation σ, then as *N* becomes large the sampling distribution of sample means will approach normality, with a mean μ and standard deviation (standard error) of $\sigma/\sqrt{N}$

Summary



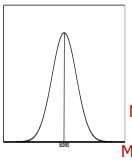
Population distribution
Typically unknown…

Blue designates
Sample distribution
which we estimate via our research…

μ ↑ mean ↓ $\overline{X}$

σ ↑ Standard deviation ↓ s

P_u ↑ proportions ↓ P_s



Sampling distribution of means (or proportions): the characteristics of which we can know from theory…

Mean of our means → $\mu_{\overline{X}}$

Mean of our proportions → μ_p

$\sigma_{\overline{X}}$ ← Standard errors

σ_p ←

Why is this useful?

The standard error from the sampling distribution can give us insight as to the quality of our "statistical estimate"!!

The larger the N, the smaller the standard error.. the more ***"efficient"*** the statistic.. likely less error in estimate

Next Week:

I will show you how to calculate "confidence intervals" using "standard errors"..