

Chapter 3

Making Statistical Inferences

3.1 Drawing Inferences About Populations

3.2 Some Basic Probability Concepts

3.3 Chebycheff's Inequality Theorem

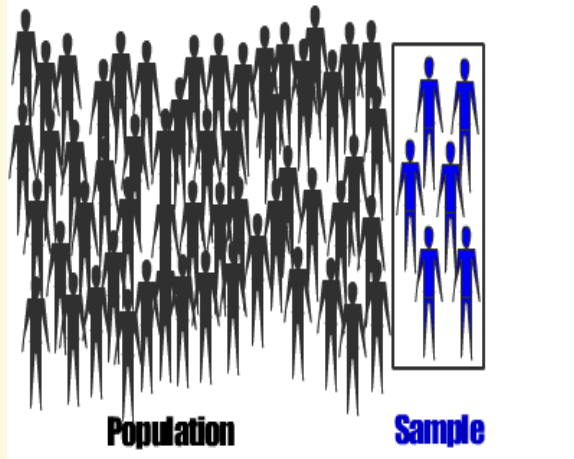
3.4 The Normal Distribution

3.5 The Central Limit Theorem

3.6 Sample Point Estimates & Confidence
Intervals

Inference - from Sample to Population

Inference: process of making generalizations (drawing conclusions) about a population's characteristics from the evidence in one sample



To make valid inferences about the population, a representative sample should be drawn from the population, preferably using **SIMPLE RANDOM SAMPLING:**

- Every population member has equal chance of selection
- Probability of a case selected for the sample is $1/N_{\text{pop}}$
- Every combination of cases has same selection likelihood

We'll treat the GSS as a s.r.s., altho it's not

Probability Theory

In 1654 the Chevalier de Méré, a wealthy French gambler, asked mathematician **Blaise Pascal** if he should bet even money on getting one “double six” in 24 throws of two dice? Pascal’s answer was no, because the probability of winning is only .491. The Chevalier’s question started an famous exchange of seven letters between Pascal and **Pierre de Fermat** in which they developed many principles of the *classical theory of probability*.



Pascal

A Russian mathematician, Andrei Kolmogorov, in a 1933 monograph, formulated the axiomatic approach which forms the foundation of the *modern theory of probability*.



Fermat

Sample Spaces

A **simple chance experiment** is a well-defined act resulting in a single event; for example, rolling a die or cutting a card deck. This process is repeatable indefinitely under identical conditions, with outcomes assumed *equiprobable* (equally likely to occur).

To compute exact event probabilities, you must know an experiment's **sample space (\mathbf{S})**, the set (collection) of all possible outcomes.



The theoretical method involves listing all possible outcomes. For rolling one die: $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$. For tossing two coins, $\mathbf{S} = \{HH, HT, TH, TT\}$.

Probability of an event: Given sample space \mathbf{S} with a set of \mathbf{E} outcomes, a probability function assigns a real number $p(E_i)$ to each event i in the sample space.

Axioms & Theorems

Three fundamental probability axioms (general rules):

1. The probability assigned to event i must be a nonnegative number:

$$p(E_i) \geq 0$$

2. The probability of the sample space \mathbf{S} (the collection of all possible outcomes) is 1:

$$p(\mathbf{S}) = 1$$

3. If two events can't happen at the same time, then the probability that *either* event occurs is the sum of their separate probabilities:

$$p(E_1) \text{ or } p(E_2) = p(E_1) + p(E_2)$$

Two important theorems (deductions) can be proved:

1. The probability of the empty (“impossible”) event is 0: $p(E_0) = 0$

2. The probability of *any* event must lie between 0 and 1, inclusive:

$$1 \geq p(E_i) \geq 0$$

Calculate these theoretical probabilities:

For rolling a single die, calculate the theoretical probability of a “4”: _____ Of a “7”: _____

For a single die roll, calculate the theoretical probability of getting either a “1” or “2” or “3” or “4” or “5” or “6”:

For tossing two coins, what is the probability of two heads: _____ Of one head and one tail: _____



If you cut a well-shuffled 52-card deck, what is the probability of getting the ten of diamonds? _____

What is the probability of any ♦ diamond card? _____

Relative Frequency

An empirical alternative to the theoretical approach is to perform a chance experiment repeatedly and observe outcomes. Suppose you roll two dice 50 times and find these sums of their face values. What are the empirical probabilities of seven? four? ten?

FIFTY DICE ROLLS

4 10 6 7 5 10 4 6 5 6
11 12 3 3 6 7 10 10 4 4
7 8 8 7 7 4 10 11 3 8 6
10 9 4 8 4 3 8 7 3 7 5 4
11 9 5 2 5 8 5

In the **relative frequency** method, probability is the proportion of times that an event occurs in a “large number” of repetitions:

$$p(E_i) = \frac{\# \text{ times event } i \text{ occurs}}{\# \text{ total events}} = \frac{N_{E_i}}{N}$$

$$p(E_7) = \underline{\hspace{2cm}}$$

$$p(E_4) = \underline{\hspace{2cm}}$$

$$p(E_{10}) = \underline{\hspace{2cm}}$$

But, **theoretically** seven is the most probable sum (.167), while four and ten each have much lower probabilities (.083). Maybe this experiment wasn't repeated often enough to obtain precise estimates? Or were these two dice “loaded?” What do we mean by “fair dice?”

Interpretation

Despite probability theory's origin in gambling, relative frequency remains the primary interpretation in the social sciences. If event rates are unknowable in advance, a "large N" of sample observations may be necessary to make accurate estimates of such empirical probabilities as:

- **What is the probability of graduating from college?**
- **How likely are annual incomes of \$100,000 or more?**
- **Are men or women more prone to commit suicide?**

Answers require survey or census data on these events.

Don't confuse formal probability concepts with everyday talk, such as "Sarah Palin will probably be elected" or "I probably won't pass tomorrow's test." Such statements express only a *personal belief* about the likelihood of a unique event, not an experiment repeated over and over.

Describing Populations

Population parameter: a descriptive characteristic of a population such as its mean, variance, or standard deviation

- Latin = sample statistic
- Greek = population parameter

Box 3.1 Parameters & Statistics

<i>Name</i>	<i>Sample Statistic</i>	<i>Population Parameter</i>
Mean	\bar{Y}	μ (mu)
Variance	S_Y^2	σ_Y^2 (sigma-squared)
Standard Deviation	S_Y	σ_Y (sigma)

3.3 Chebycheff's Inequality Theorem

If you have the book, read this subsection (pp. 73-75) as background information on the normal distribution.

Because Cheby's inequality is never calculated in research statistics, we'll not spend time on it in lecture.

The Normal Distribution

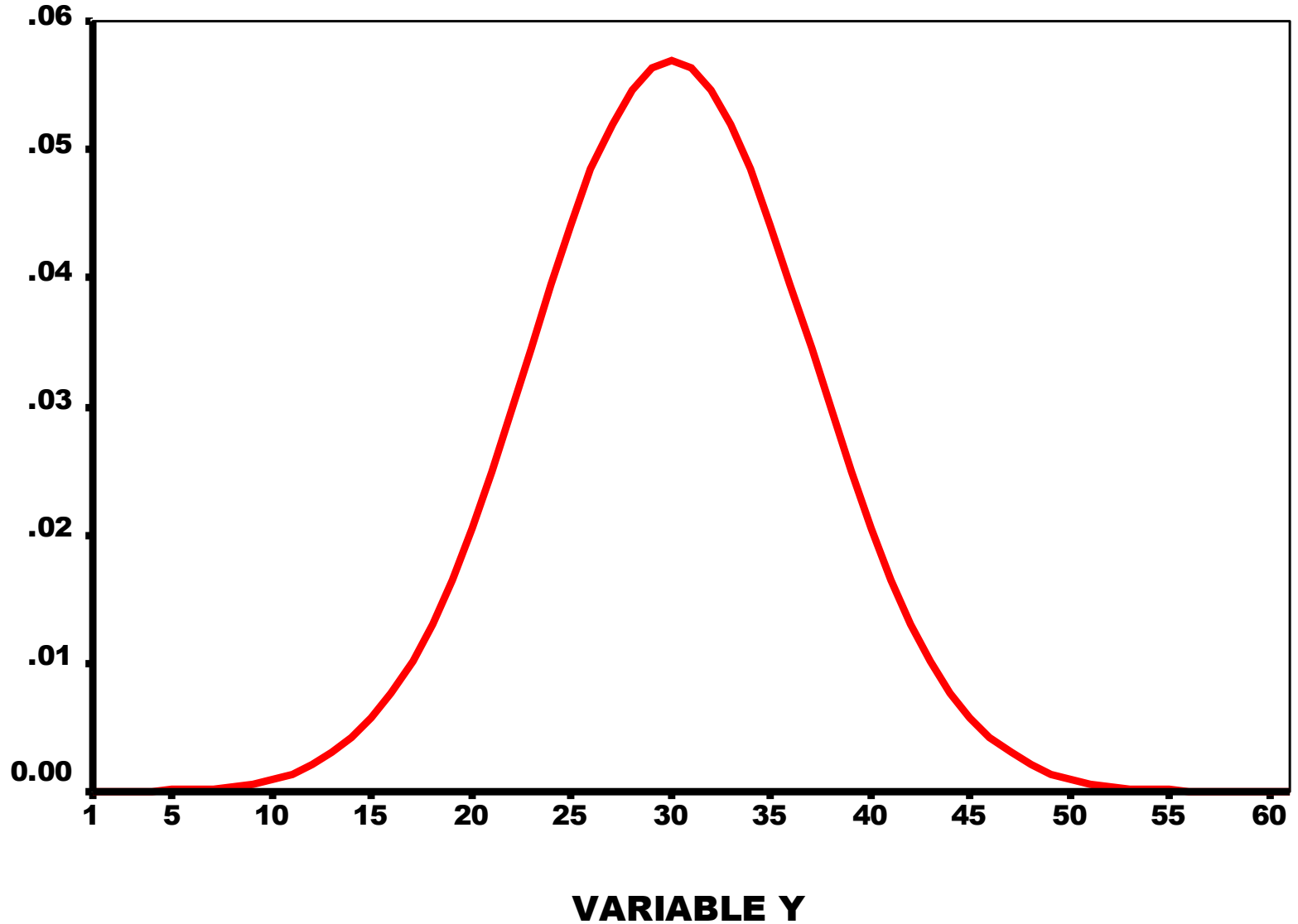
Normal distribution: smooth, bell-shaped theoretical probability distribution for a continuous variable, generated by a formula:

$$p(\mathbf{Y}) = \frac{e^{-(\mathbf{Y}-\mu_{\mathbf{Y}})^2 / 2\sigma_{\mathbf{Y}}^2}}{\sqrt{2\pi\sigma_{\mathbf{Y}}^2}}$$

where e is Euler's constant (= 2.7182818....)

The population mean and variance determine a particular distribution's location and shape. Thus, the family of normal distributions has an infinite number of curves.

A Normal Distribution
with Mean = 30 and Variance = 49



Comparing Three Normal Curves

Suppose we graph three normally distributions, each with a mean of zero: $(\mu_Y = 0)$

What happens to the height and spread of these normal probability distributions if we increase the population's variance?

Next graph superimposes these three normally distributed variables with these variances:

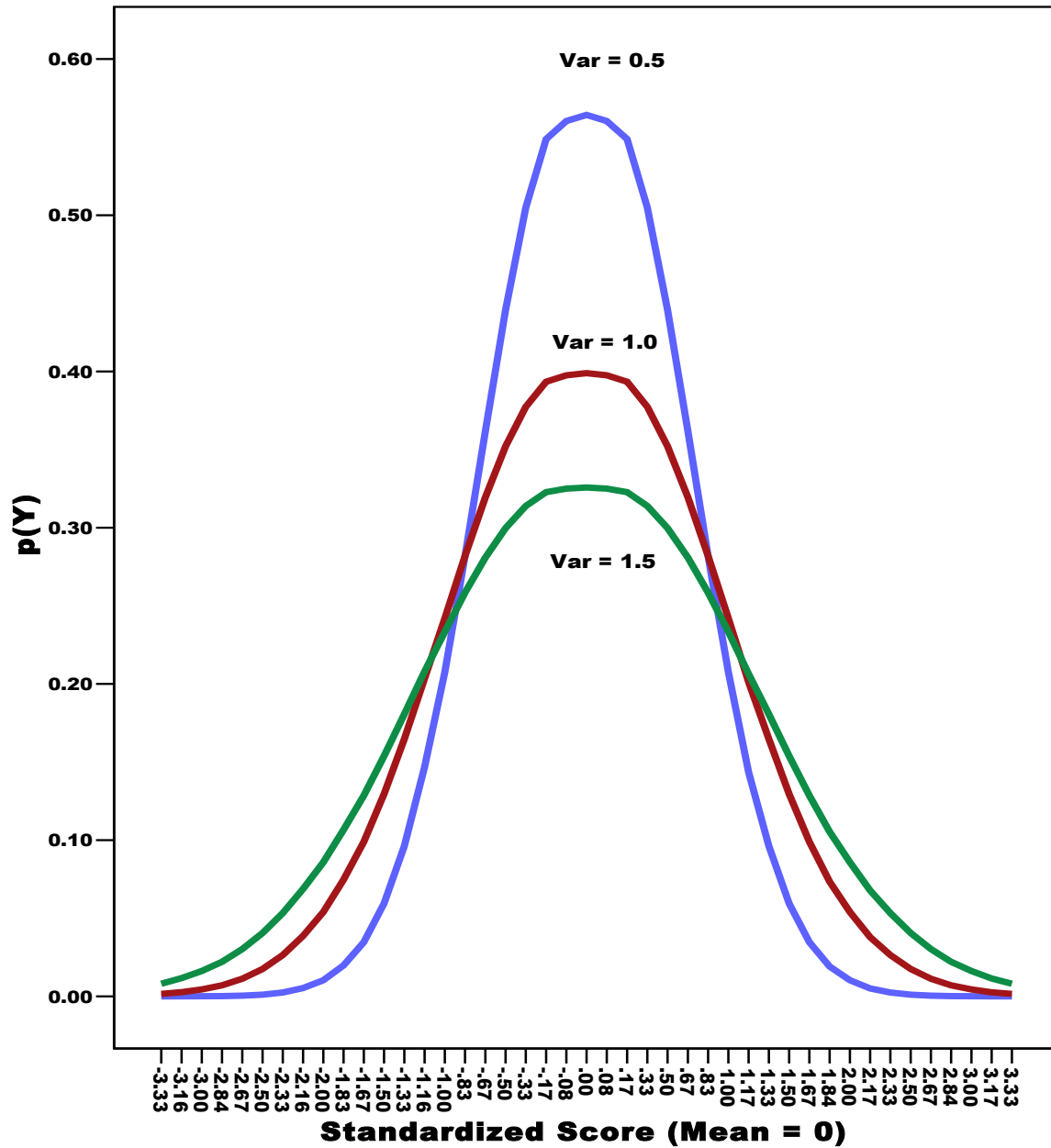
$$(\sigma_Y^2)$$

$$(1) = 0.5$$

$$(2) = 1.0$$

$$(3) = 1.5$$

Normal Curves with Different Variances



Standardizing a Normal Curve

To standardize any normal distribution, change the Y scores to Z scores, whose mean = 0 and std. dev. = 1. Then use the known relation between the Z scores and probabilities associated with areas under the curve.

We previously learned how to convert a sample of Y_i scores into standardized Z_i scores:

$$Z_i = \frac{Y_i - \bar{Y}}{s_Y}$$

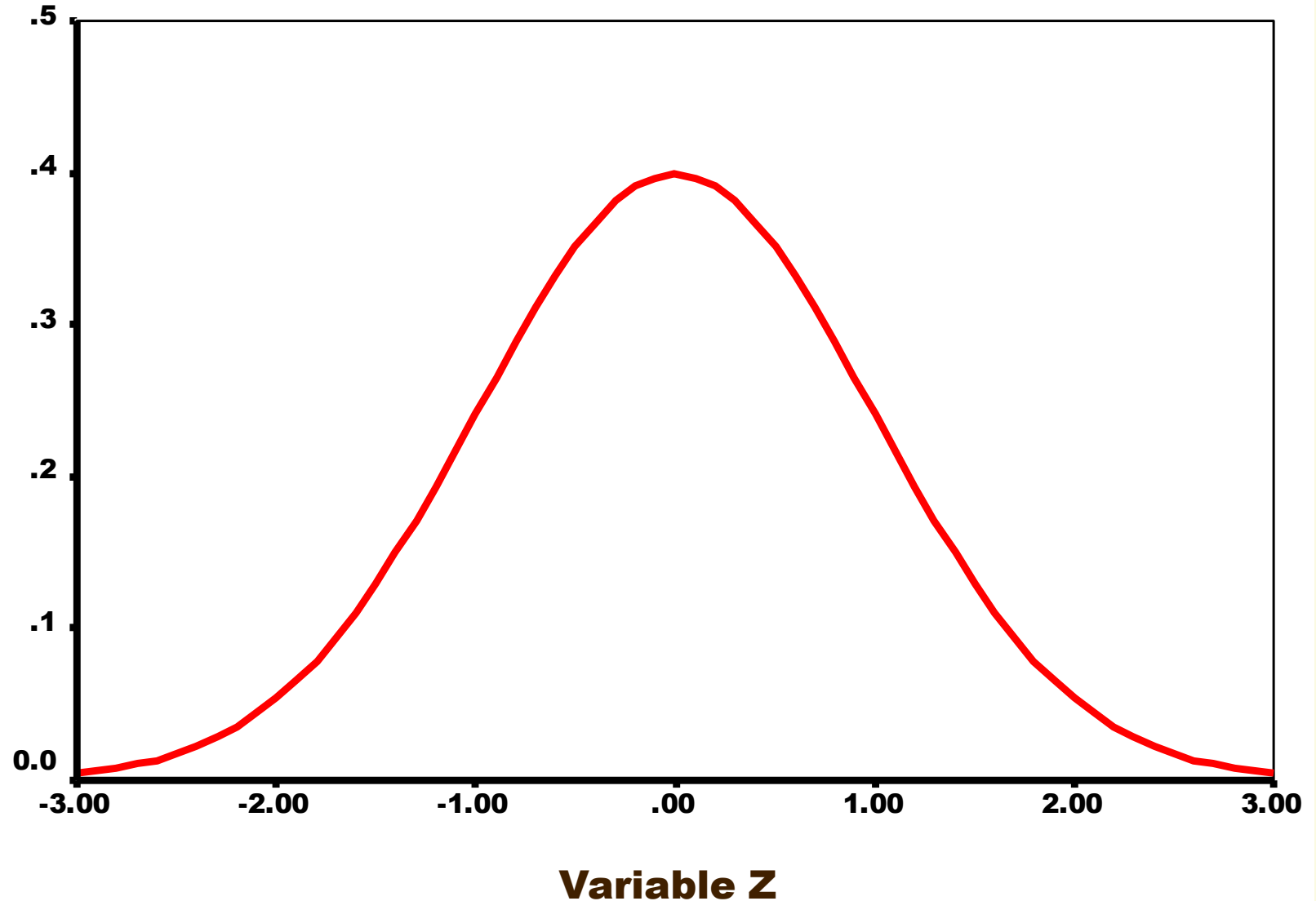
Likewise, we can standardize a population of Y_i scores:

$$Z_i = \frac{Y_i - \mu_Y}{\sigma_Y}$$

We can use a standardized Z score table (Appendix C) to solve all normal probability distribution problems, by finding the area(s) under specific segment(s) of the curve.

The Standardized Normal Distribution

with Mean = 0 and Variance = 1



Area = Probability

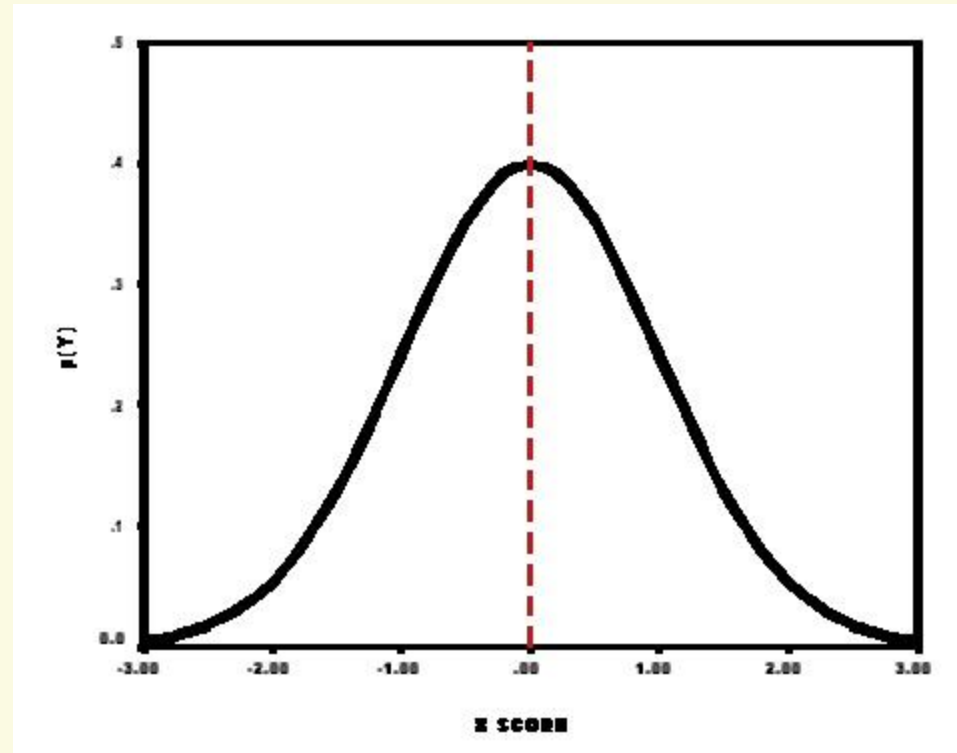
The **TOTAL AREA** under a standardized normal probability distribution is assumed to have unit value; i.e., = 1.00

This area corresponds to probability $p = 1.00$ (certainty).

Exactly half the total area lies on each side of the mean, ($\mu_Y = 0$)

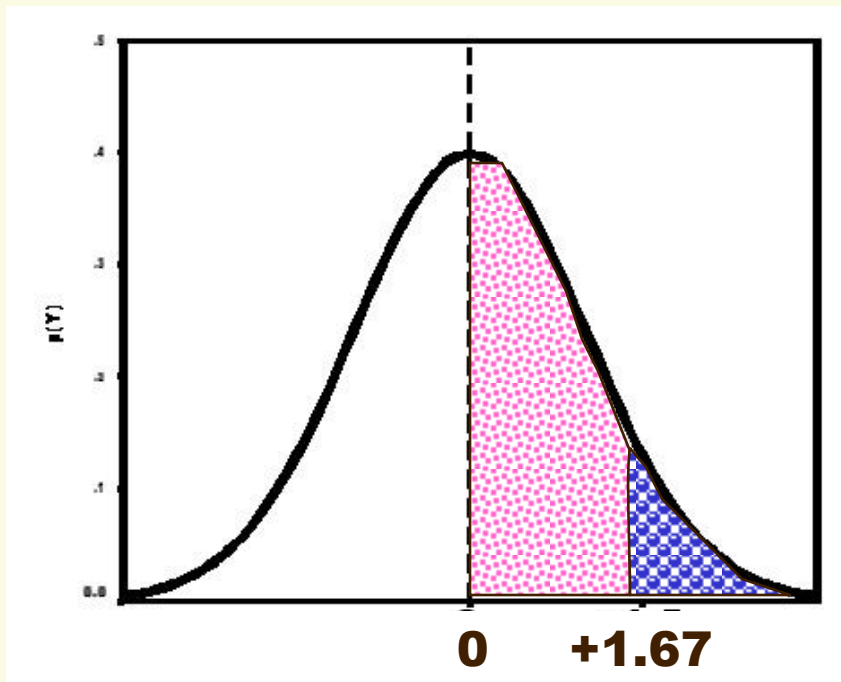
(left side negative Z , right side positive Z)

Thus, each half of the normal curve corresponds to $p = 0.500$



Areas Between Z Scores

Using the tabled values in a table, we can find an area (a probability) under a standardized normal probability distribution that falls between two Z scores

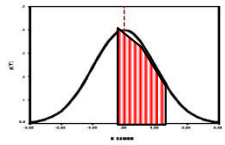
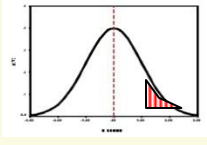


EXAMPLE #1: What is area between $Z = 0$ and $Z = +1.67$?

EXAMPLE #2: What is area from $Z = +1.67$ to $Z = +\infty$?

Also use the [Web-page version of Appendix C](#), which gives pairs of values for the areas (0 to Z) and (Z to ∞).

Appendix C The Z Score Table

Z score	Area from 0 to Z	Area from Z to ∞
		
1.50	0.4332	0.0668
...		
1.60	0.4452	0.0548
...		
1.65	0.4505	0.0495
1.66	0.4515	0.0485
1.67	0.4525	0.0475
1.68	0.4535	0.0465
1.69	0.4545	0.0455
1.70	0.4554	0.0446

For $Z = 1.67$:

Col. 2 = _____

Col. 3 = _____

Sum = _____

EX #3: What is area between $Z = 0$ and $Z = -1.50$?

EX #4: What is area from $Z = -1.50$ to $Z = -\infty$?

Calculate some more Z score areas

EX #5: Find the area from $Z = -1.65$ to $-\infty$ _____

EX #6: Find the area from $Z = +1.96$ to ∞ _____

EX #7: Find the area from $Z = -2.33$ to $-\infty$ _____

EX#8: Find the area from $Z = 0$ to -2.58 _____

Use the table to locate areas between or beyond two Z scores.

Called “two-tailed” Z scores because areas are in both tails:

EX #9: Find the area from $Z = 0$ to ± 1.96 _____

EX #10: Find the areas from $Z = \pm 1.96$ to $\pm\infty$ _____

EX #11: Find the areas from $Z = \pm 2.58$ to $\pm\infty$ _____

The Useful Central Limit Theorem

Central limit theorem: if all possible samples of size N are drawn from any population, with mean μ_Y and variance σ_Y^2 , then as N grows large, the sampling distribution of these means approaches a normal curve, with mean μ_Y and variance σ_Y^2 / N

The positive square root of a sampling distribution's variance (i.e., its standard deviation), is called the **standard error** of the mean:

$$\sqrt{\frac{\sigma_Y^2}{N}} = \frac{\sigma_Y}{\sqrt{N}} = \sigma_{\bar{Y}}$$

Take ALL Samples in a Small Population

Population ($N = 6$, mean = 4.33):

Form all samples of size $n = 2$ & calculate means:

$Y_1 = 2$
$Y_2 = 2$
$Y_3 = 4$
$Y_4 = 4$
$Y_5 = 6$
$Y_6 = 8$

$$Y_1+Y_2 = (2+2)/2 = 2 \quad Y_1+Y_3 = (2+4)/2 = 3$$

$$Y_1+Y_4 = (2+4)/2 = 3 \quad Y_1+Y_5 = (2+6)/2 = 4$$

$$Y_1+Y_6 = (2+8)/2 = 5 \quad Y_2+Y_3 = (2+4)/2 = 3$$

$$Y_2+Y_4 = (2+4)/2 = 3 \quad Y_2+Y_5 = (2+6)/2 = 4$$

$$Y_2+Y_6 = (2+8)/2 = 5 \quad Y_3+Y_4 = (4+4)/2 = 4$$

$$Y_3+Y_5 = (4+6)/2 = 5 \quad Y_3+Y_6 = (4+8)/2 = 6$$

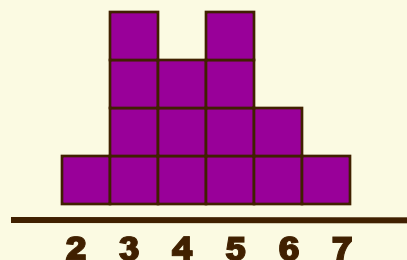
$$Y_4+Y_5 = (4+6)/2 = 5 \quad Y_4+Y_6 = (4+8)/2 = 6$$

$$Y_5+Y_6 = (6+8)/2 = 7$$

Calculate the mean of these 15 sample means = _____

Probability that a sample mean = 7? _____

Graph this sampling distribution of 15 sample means:

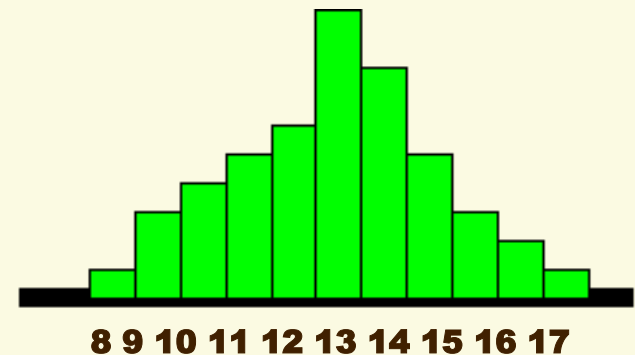
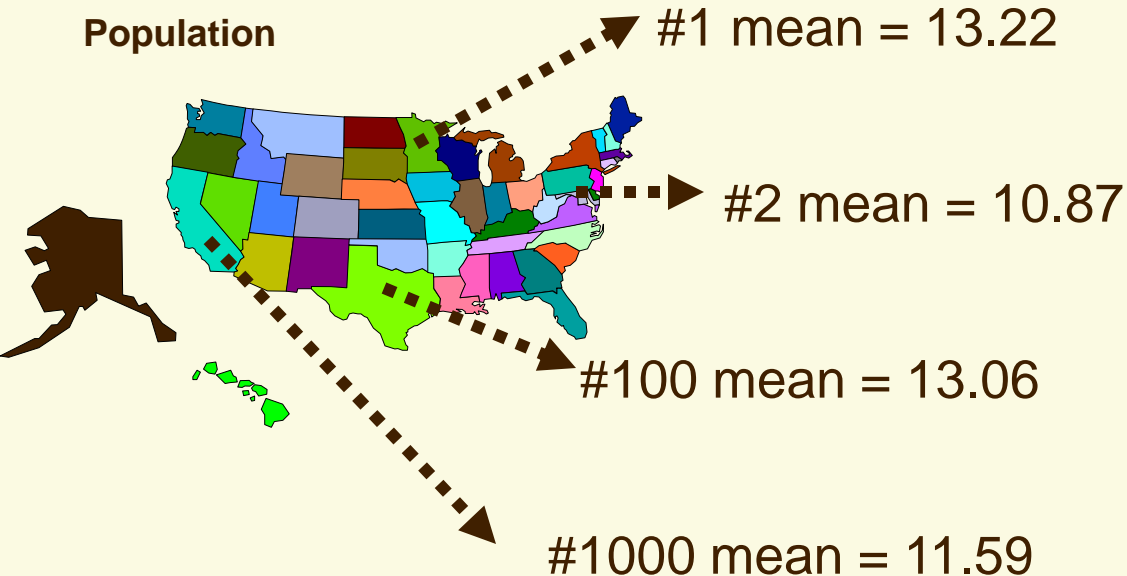


Take ALL Samples in a Large Population

A “thought experiment” suggests how a theoretical sampling distribution is built by: (a) forming every sample of size N in a large population, (b) then graphing all samples’ mean values.

Let’s take many samples of 1,000 persons and calculate each sample’s mean years of education:

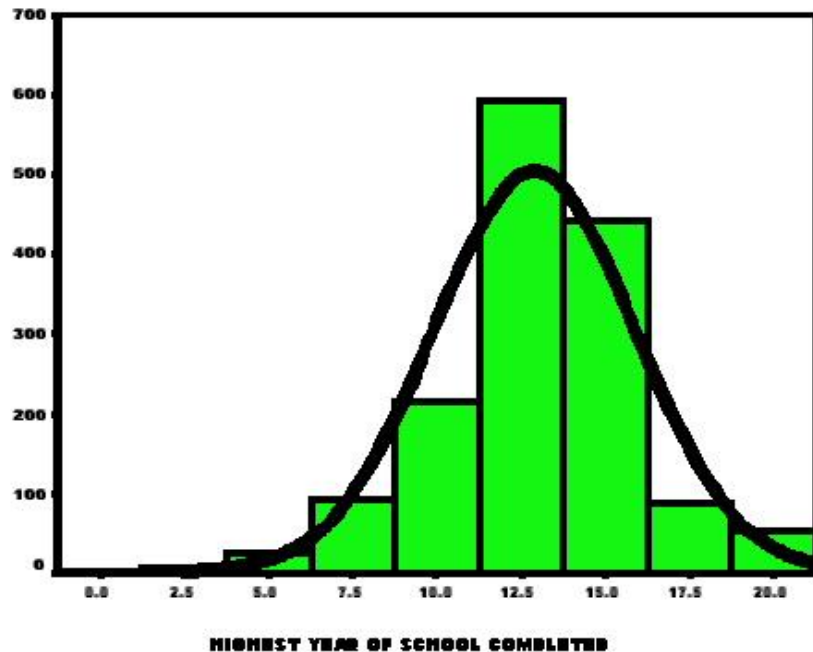
A graph of this sampling distribution of sample means increasingly approaches a normal curve:



Sampling Distribution for EDUC

Start with a variable in a population with a known standard deviation:

U.S. adult population of about 230,000,000 has a mean education = 13.43 years of schooling with a standard deviation = 3.00.



If we generate sampling distributions for samples of increasingly larger N , what do you expect will happen to the values of the mean and standard error for these sampling distributions, according to the Central Limit Theorem?

Sampling distributions with differing N s

1. Let's start with random samples of $N = 100$ observations.

CAUTION! BILLIONS of TRILLIONS of such small samples make up this sampling distribution!!!

What are the expected values for mean & standard error?

$$\mu_{\bar{Y}} = \underline{\hspace{2cm}} \quad \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}} = \underline{\hspace{2cm}}$$

2. Now double $N = 200$. What mean & standard error?

$$\mu_{\bar{Y}} = \underline{\hspace{2cm}} \quad \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}} = \underline{\hspace{2cm}}$$

3. Use GSS $N = 2,018$. What mean & standard error?

$$\mu_{\bar{Y}} = \underline{\hspace{2cm}} \quad \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}} = \underline{\hspace{2cm}}$$

Online Sampling Distribution Demo

Rice University Virtual Lab in Statistics:

http://onlinestatbook.com/stat_sim/

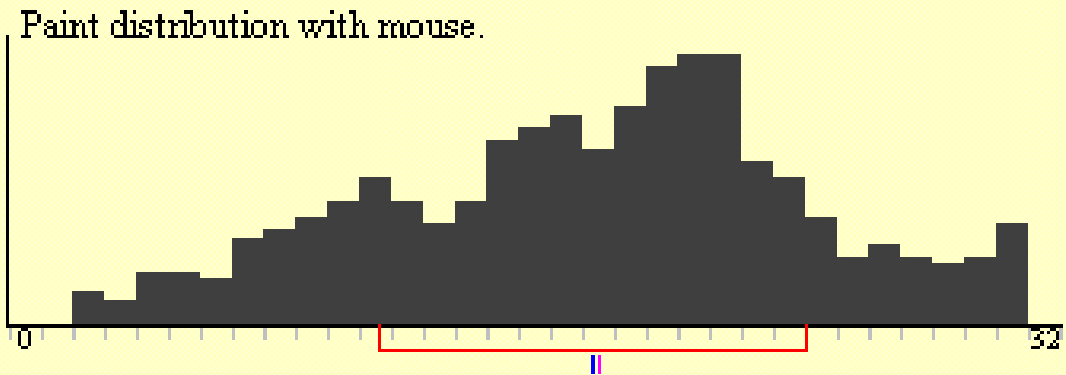
Choose & click Sampling Distribution Simulation

(requires browser with Java 1.1 installed)

Read Instructions, Click "Begin" button

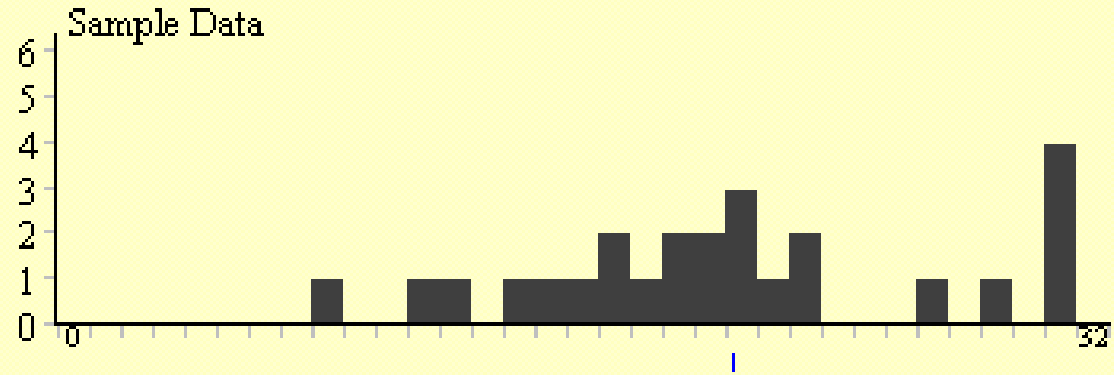
We'll work some examples in class, then you can try this demo for yourself. See screen capture on next slide:

mean= 17.76
 median= 18.00
 sd= 6.74
 skew= -0.18
 kurtosis= -0.52



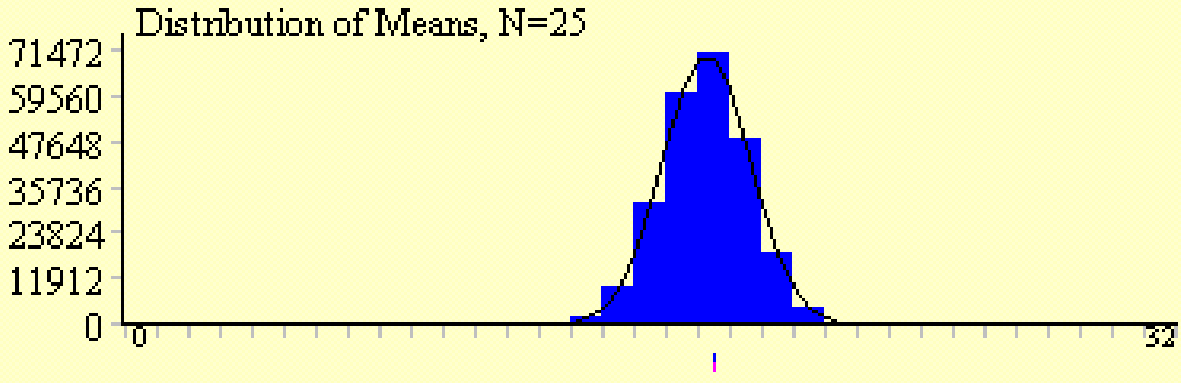
Clear lower 3
 Custom ▾

Reps = 25
 mean= 20.67



Animated Sample
 5 Samples
 1,000 Samples
 10,000 Samples

Reps = 250060
 mean= 17.96
 median= 18.00
 sd= 6.74
 skew= -0.48
 kurtosis= 0.14



Mean ▾
 N=25 ▾
 Fit normal

How Big is a “Large Sample?”

- To be applied, the central limit theorem requires a “large sample”
- But how big must a simple random sample be for us to call it “large”?

SSDA p. 81: “we cannot say precisely.”

- $N < 30$ is a “small sample”
- $N > 100$ is a “large sample”
- $30 < N < 100$ is indeterminate

The Alpha Area

Alpha area (α area): area in tail of normal distribution that is cut off by a given Z_α

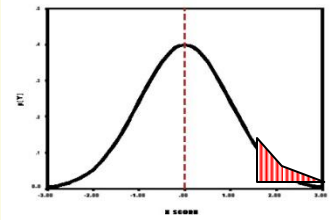
Because we could choose to designate α in either the negative or positive tail (or in both tails, by dividing α in half), we define an alpha area's probability using the absolute value:

$$p(|Z| \geq |Z_\alpha|) = \alpha$$

The alpha area is also called the **region of rejection** when used to make a decision about a hypothesis test, as shown later in this chapter.

Critical value (Z_α): the minimum value of Z necessary to designate an alpha area

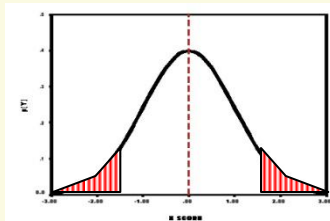
Find the critical values of Z that define six alpha areas:



$\alpha = 0.05$ one-tailed $Z =$ _____

$\alpha = 0.01$ one-tailed $Z =$ _____

$\alpha = 0.001$ one-tailed $Z =$ _____



$\alpha = 0.05$ two-tailed $Z =$ _____

$\alpha = 0.01$ two-tailed $Z =$ _____

$\alpha = 0.001$ two-tailed $Z =$ _____

These α and Z are the six conventional values used to test hypotheses.

Apply Z scores to a sampling distribution of EDUC where

$$\mu_Y = 13.43 \text{ and } \sigma_{\bar{Y}} = 0.067$$

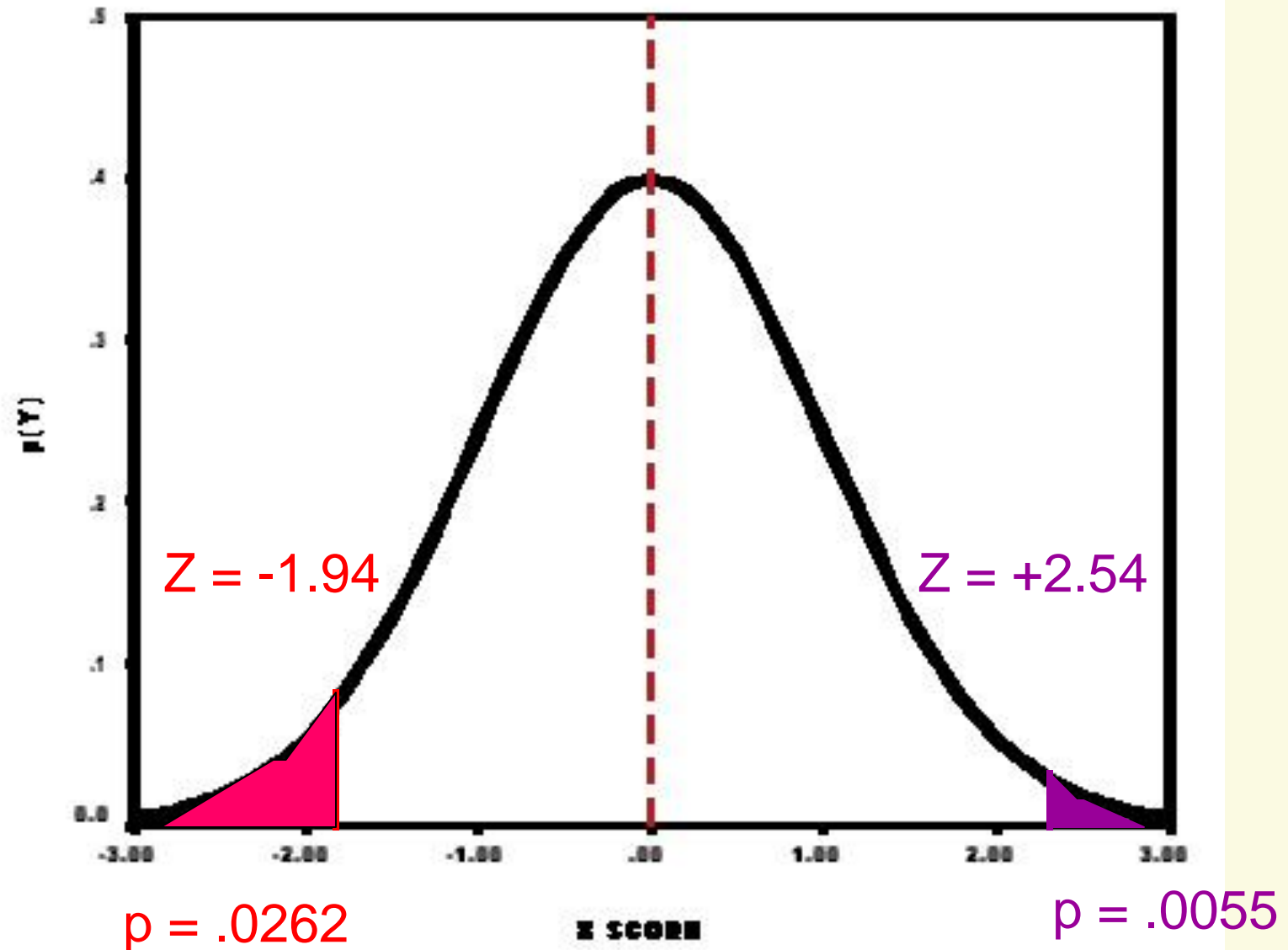
What is the probability of selecting a GSS sample of $N = 2,018$ cases whose mean is equal to or greater than 13.60?

$$Z_i = \frac{\bar{Y}_i - \mu_Y}{\sigma_{\bar{Y}}} = \underline{\hspace{10em}} \quad \text{C: Area Beyond Z = } \underline{\hspace{10em}}$$

What is the probability of drawing a sample with mean = 13.30 or less?

$$Z_i = \frac{\bar{Y}_i - \mu_Y}{\sigma_{\bar{Y}}} = \underline{\hspace{10em}} \quad \text{C: Area Beyond Z = } \underline{\hspace{10em}}$$

Two Z Scores in a Sampling Distribution



Find Sample Means for an Alpha Area

What sample means divide $\alpha = .01$ equally into both tails of the EDUC sampling distribution?

1. Find half of alpha:

$$\alpha/2 = (.01)/2 = .005$$

2. Look up the two values of the critical $Z_{\alpha/2}$ scores:

In Table C the area beyond Z
($\alpha = .005$), $Z_{\alpha/2} =$ _____

3. Rearrange Z formula to isolate the sample mean on one side of the computation:

$$Z_i = \frac{\bar{Y}_i - \mu_Y}{\sigma_{\bar{Y}}} \Rightarrow (\pm Z_i)(\sigma_{\bar{Y}}) + \mu_Y = \bar{Y}_i$$

4. Compute the two _____

critical mean values: _____

Point Estimate vs. Confidence Interval

Point estimate: sample statistic used to estimate a population parameter

In the 2008 GSS, mean family income = \$58,683, the standard deviation = \$46,616 and $N = 1,774$. Thus, the estimated standard error = $\$46,616/42.1 = \$1,107$.

Confidence interval: a range of values around a point estimate, making possible a statement about the probability that the population parameter lies between upper and lower **confidence limits**

The 95% CI for U.S. annual income is from \$56,513 to \$60,853, around a point estimate of \$58,683.

Below you will learn below how use the sample mean and standard error to calculate the two CI limits.

Confidence Intervals

An important corollary of the central limit theorem is that the sample mean is the best point estimate of the mean of the population from which the sample was drawn:

$$\bar{Y} = \mu_Y$$

We can use the sampling distribution's **standard error** to build a confidence interval around a point-estimated mean. This interval is defined by the upper and lower limits of its range, with the point estimate at the midpoint.

Then use this estimated interval to state how confident you feel that the unknown population parameter (μ_Y) falls inside the limits defining the interval.

UCL & LCL

A researcher sets a confidence interval by deciding how “confident” she wishes to be. The trade-off is that obtaining greater confidence requires a broader interval.

- Select an alpha (α) for desired confidence level
- Split alpha in half ($\alpha/2$) & find the critical Z scores in the standardized normal table (+ and – values)
- Multiply each $Z_{\alpha/2}$ by the standard error, then separately add each result to sample mean

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}})$$

Upper confidence limit, UCL: $\bar{Y} + (Z_{\alpha/2})(\sigma_{\bar{Y}})$

Lower confidence limit, LCL: $\bar{Y} - (Z_{\alpha/2})(\sigma_{\bar{Y}})$

Show how to calculate the 95% CI for 2008 GSS income

For GSS sample $N = 1,774$ cases, sample mean: $\bar{Y} = \$58,683$

The standard error for annual income: $\sigma_{\bar{Y}} = \$1,107$

Upper confidence limit, 95% UCL: $\bar{Y} + (Z_{\alpha/2})(\sigma_{\bar{Y}})$

Lower confidence limit, 95% LCL: $\bar{Y} - (Z_{\alpha/2})(\sigma_{\bar{Y}})$

Now compute the 99% CI:

Upper confidence limit, 99% UCL: $\bar{Y} + (Z_{\alpha/2})(\sigma_{\bar{Y}})$

Lower confidence limit, 99% LCL: $\bar{Y} - (Z_{\alpha/2})(\sigma_{\bar{Y}})$

For $\bar{Y} = 40$ and $\sigma_{\bar{Y}} = 6$, find the UCL & LCL for these two CIs:

A: The 95% confidence interval: $\alpha = 0.05$, so $Z = \pm 1.96$

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}}) = 40 \pm (1.96)(6)$$

UCL = _____

LCL = _____

B: The 99% confidence interval; $\alpha = 0.01$, so $Z = \pm 2.58$

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}}) = 40 \pm (2.58)(6)$$

UCL = _____

LCL = _____

Thus, to obtain more confidence requires a wider interval.

Interpreting a CI

A CI interval indicates how much uncertainty we have about a sample estimate of the true population mean. The wider we choose an interval (e.g., 99% CI), the more confident we are.

CAUTION: A 95% CI does **not** mean that an interval has a 0.95 probability of containing the true mean. Any interval estimated from a sample either contains the true mean or it does not – but you can't be certain!

Correct interpretation: A confidence interval is not a probability statement about a single sample, but is based on the idea of repeated sampling. If all samples of the same size (N) were drawn from a population, and confidence intervals calculated around every sample mean, then 95% (or 99%) of intervals would be expected to contain the population mean (but 5% or 1% of intervals would not).

Just say: “I’m 95% (or 99%) confident that the true population mean falls between the lower and upper confidence limits.”

Calculate another CI example

If $\bar{Y} = 50$ and $\sigma_{\bar{Y}} = 3.16$ find UCL & LCL for two CIs:

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}}) = 50 \pm (1.96)(3.16)$$

$$\text{LCL} = \underline{\hspace{2cm}}$$

$$\text{UCL} = \underline{\hspace{2cm}}$$

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}}) = 50 \pm (2.58)(3.16)$$

$$\text{LCL} = \underline{\hspace{2cm}}$$

$$\text{UCL} = \underline{\hspace{2cm}}$$

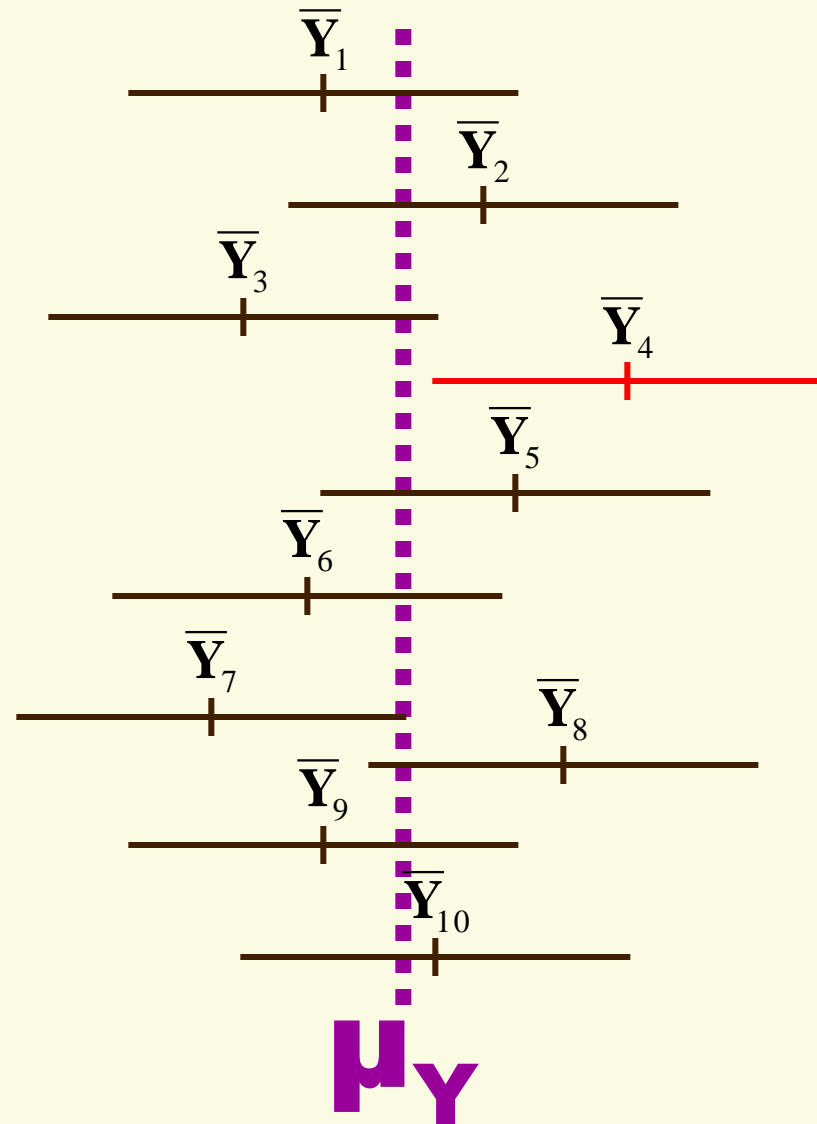
INTERPRETATION: For all samples of the same size (N), if confidence intervals were constructed around each sample mean, 95% (or 99%) of those intervals would include the population mean somewhere between upper and lower limits.

Thus, we can be 95% confident that the population mean lies between 43.8 and 56.2. And we can have 99% confidence that the parameter falls into the interval from 41.8 to 58.2.

A Graphic View of CIs

The confidence intervals constructed around 95% (or 99%) of all sample means of size N from a population can be expected to include the true population mean (dashed line) within the lower and upper limits.

But, in 5% (or 1%) of the samples, the population parameter would fall outside their confidence intervals.



Online CI Demo

Rice University Virtual Lab in Statistics:

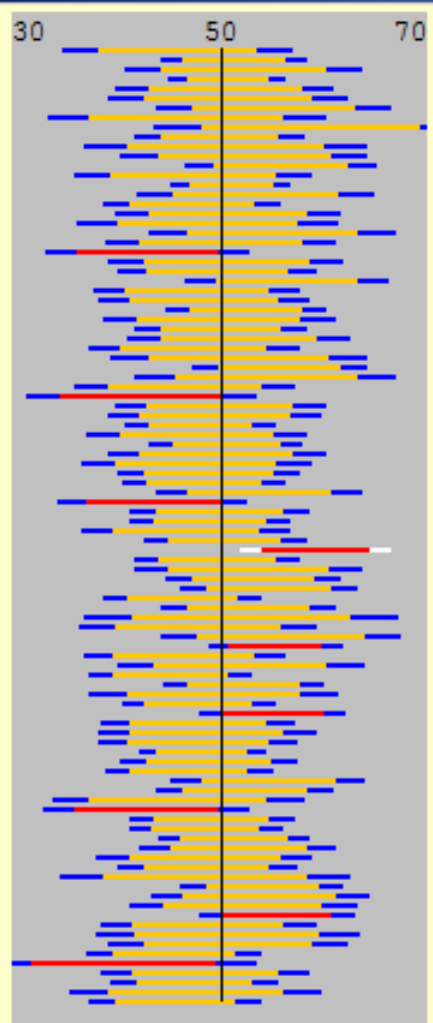
http://onlinestatbook.com/stat_sim/

Choose & click Confidence Intervals

(requires browser with Java 1.1 installed)

Read Instructions, Click "Begin" button

We'll work some examples in class, then you can try this demo for yourself. See screen capture on next slide:



Sample size:

When you click the sample button, 100 samples of the specified sample size (10, 15, or 20) will be taken from a population with a mean of 50 and a standard deviation of 10. The confidence interval on the mean will be computed for each. If the 95% confidence interval contains the population mean of 50 then a line will show the 95% confidence interval in orange and the 99% confidence interval in blue. If the 95% confidence interval does not contain the population mean then it will be shown in red. If the 99% interval does not contain the population mean it will be shown in white.

Cumulative Results:

	99% Conf. Int	95% Conf. Int
Contained 50	298	286
Did Not Contain 50	2	14
Proportion Contained	0.993	0.953

What is a “Margin of Error”?

Opinion pollsters report a “margin of error” with their point estimates:



The Gallup Poll’s final survey of 2010, found that 48% of the 1,500 respondents said they approved how Pres. Obama was doing his job, with a “margin of sampling error” = ± 3 per cent.

Using your knowledge of basic social statistics, you can calculate --

(1) the **standard deviation** for the sample point-estimate of a proportion:

$$s_p = \sqrt{p_1 p_0} = \sqrt{(0.48)(0.52)} = \sqrt{0.2496} = 0.4996$$

(2) Use that sample value to estimate the sampling distribution’s **standard error**:

$$\sigma_{\bar{p}} = s_p / \sqrt{N} = 0.4996 / \sqrt{1500} = 0.4996 / 38.73 = 0.0129$$

(3) Then find the upper and lower **95% confidence limits**:

$$LCL = p_1 - (Z_{\alpha/2})(\sigma_{\bar{p}}) = 0.48 + (-1.96)(0.0129) = 0.48 - .025 = 0.455$$

$$UCL = p_1 + (Z_{\alpha/2})(\sigma_{\bar{p}}) = 0.48 + (+1.96)(0.0129) = 0.48 + .025 = 0.505$$

Thus, a “margin of error” is just the product of the standard error times the critical value of $Z_{\alpha/2}$ for the 95% confidence interval!