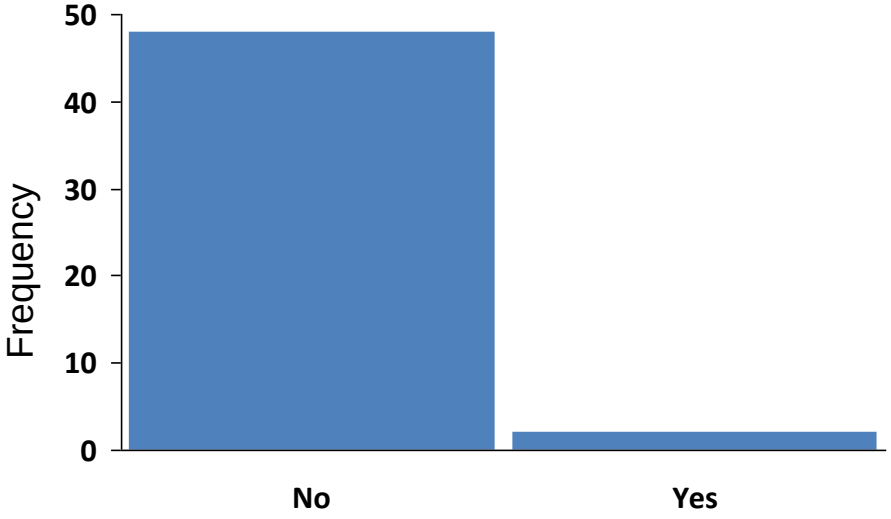
SOC 3811/5811:
BASIC SOCIAL STATISTICS

Sampling Distributions

1

Random Sample
of n=50 college students



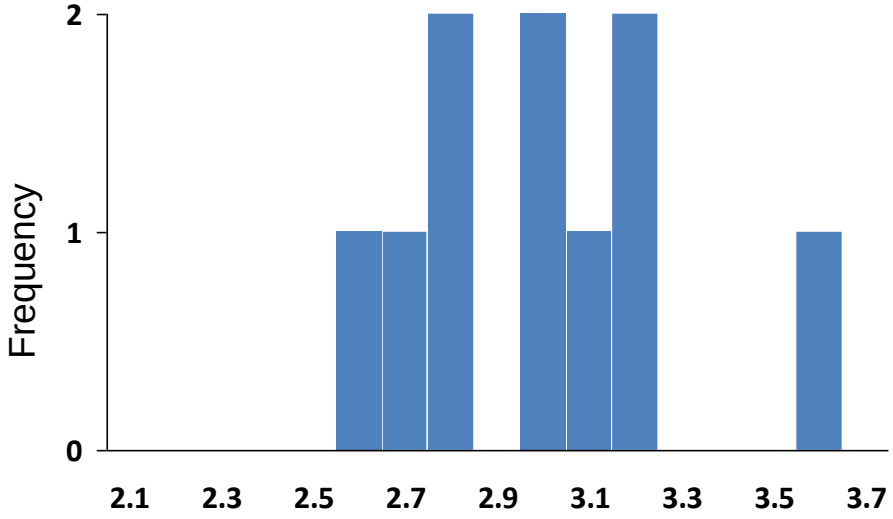
Sociology Major?

1

Sample Percentage = 4.0%

1

Random Sample
of n=10 college students



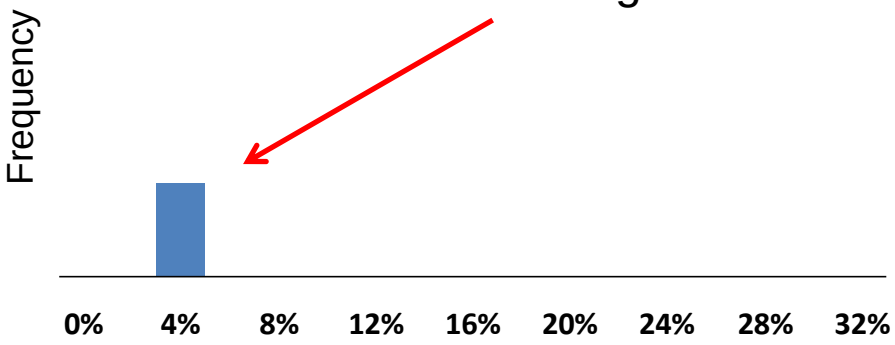
Average GPA

1

Sample Mean = 3.0

1
Random Sample
of n=50 college students

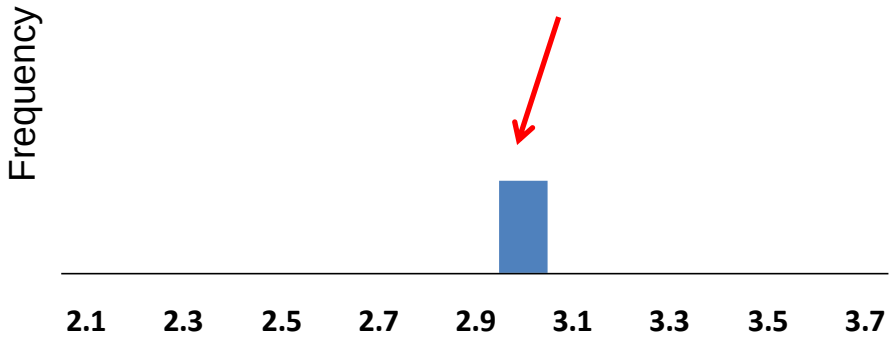
1
Sample
Percentage = 4.0%



Percent Sociology Majors

1
Random Sample
of n=10 college students

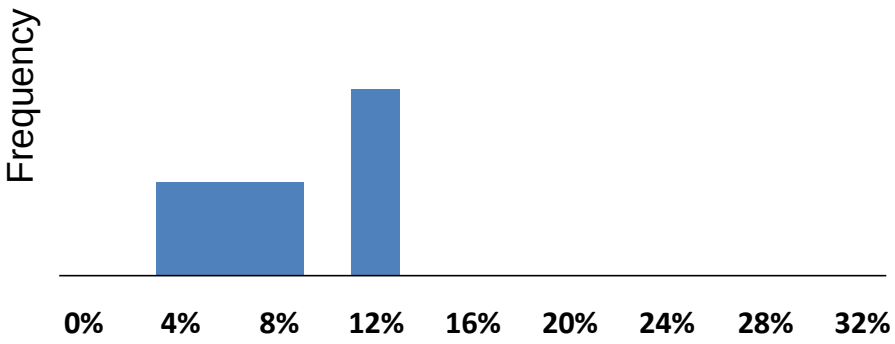
1
Sample
Mean = 3.0



Average GPA

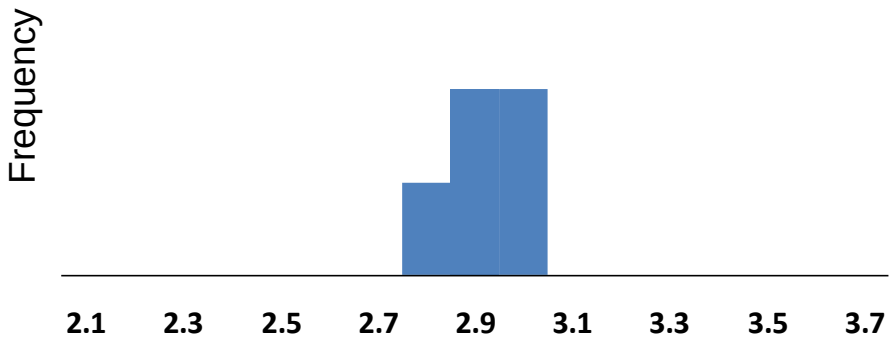
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

5
Random Samples
of n=50 college students



Percent Sociology Majors

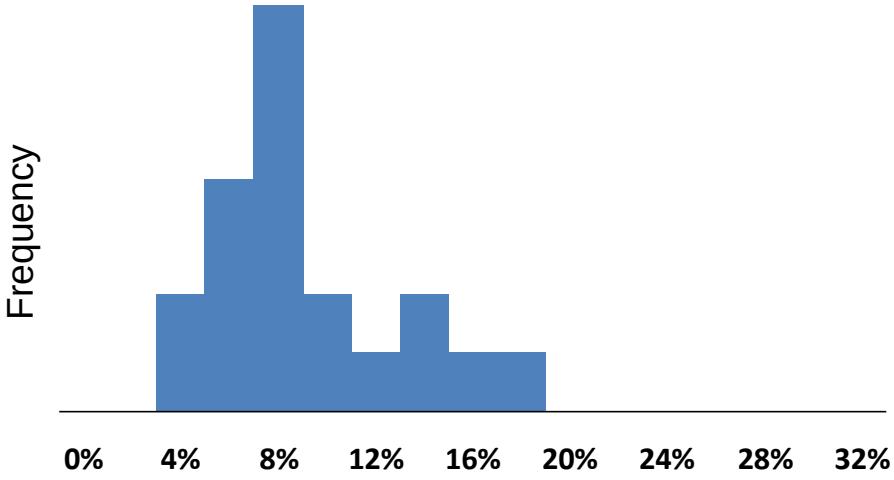
5
Random Samples
of n=10 college students



Average GPA

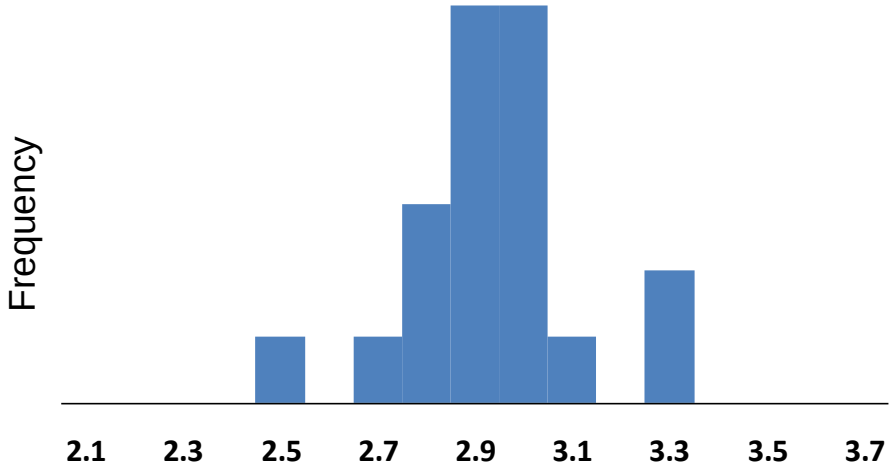
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

20
Random Samples
of n=50 college students



Percent Sociology Majors

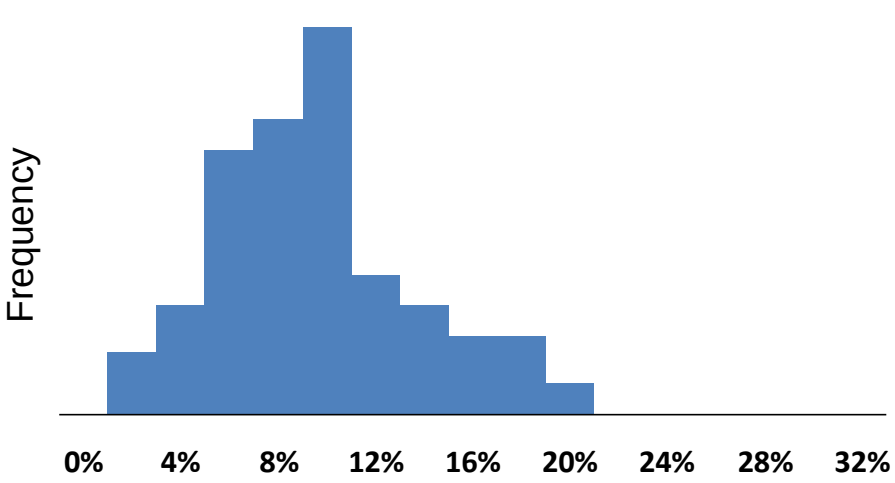
20
Random Samples
of n=10 college students



Average GPA

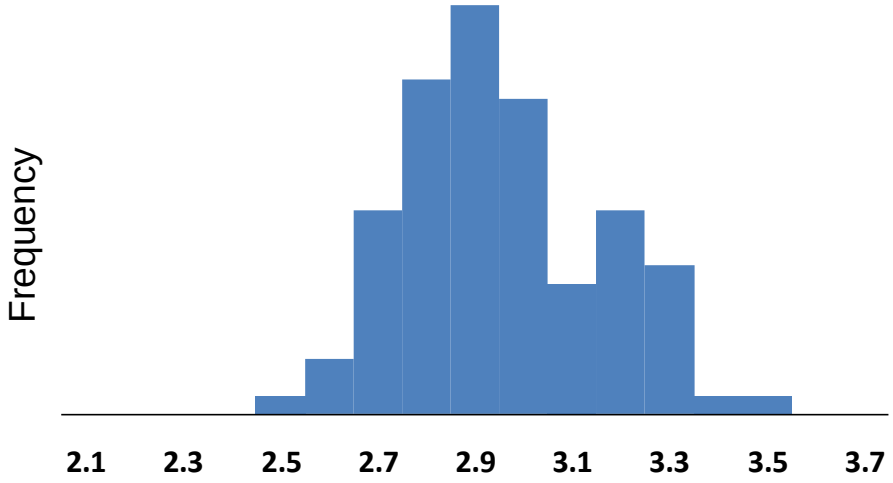
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

100
Random Samples
of n=50 college students



Percent Sociology Majors

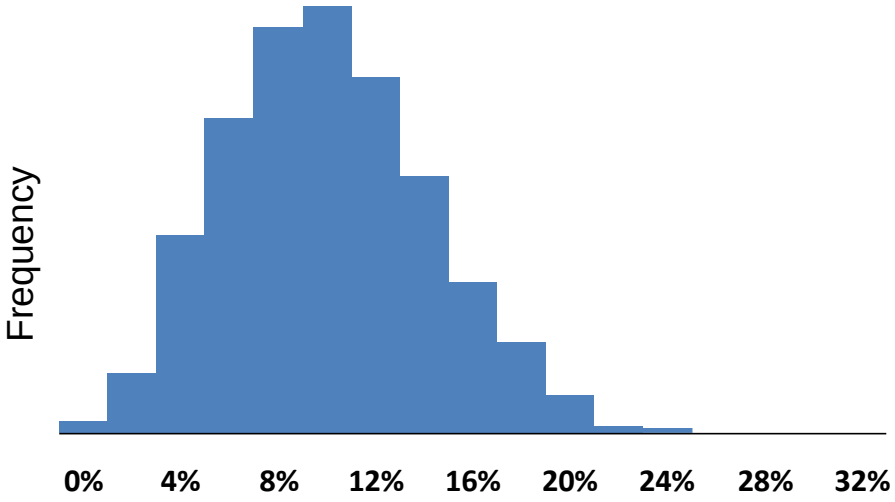
100
Random Samples
of n=10 college students



Average GPA

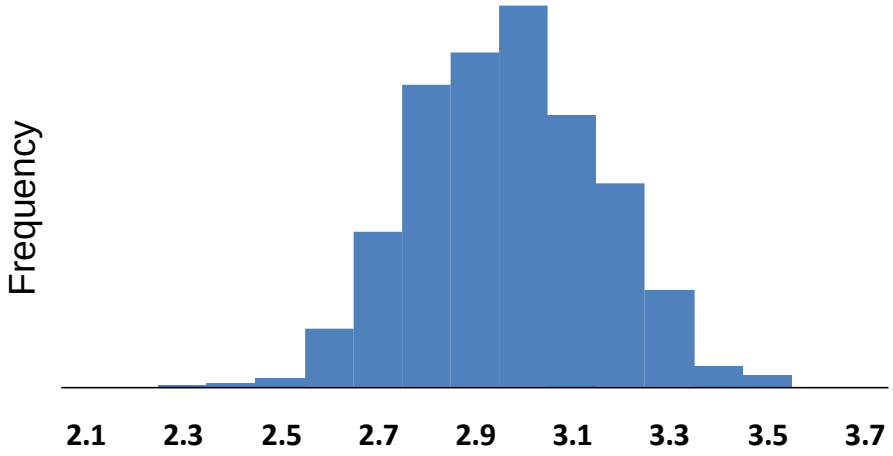
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

1,000
Random Samples
of n=50 college students



Percent Sociology Majors

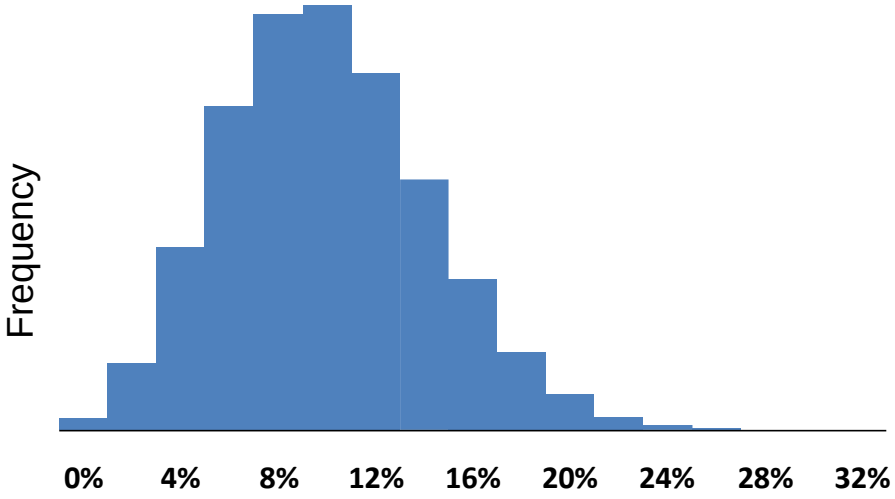
1,000
Random Samples
of n=10 college students



Average GPA

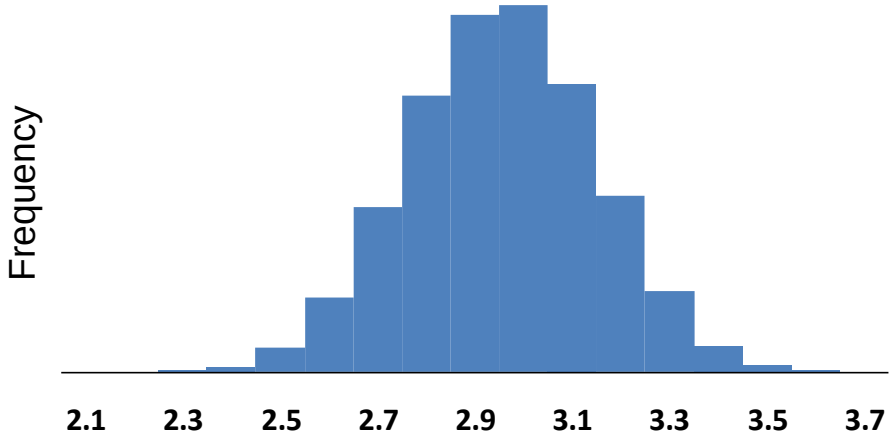
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

100,000
Random Samples
of n=50 college students



Percent Sociology Majors

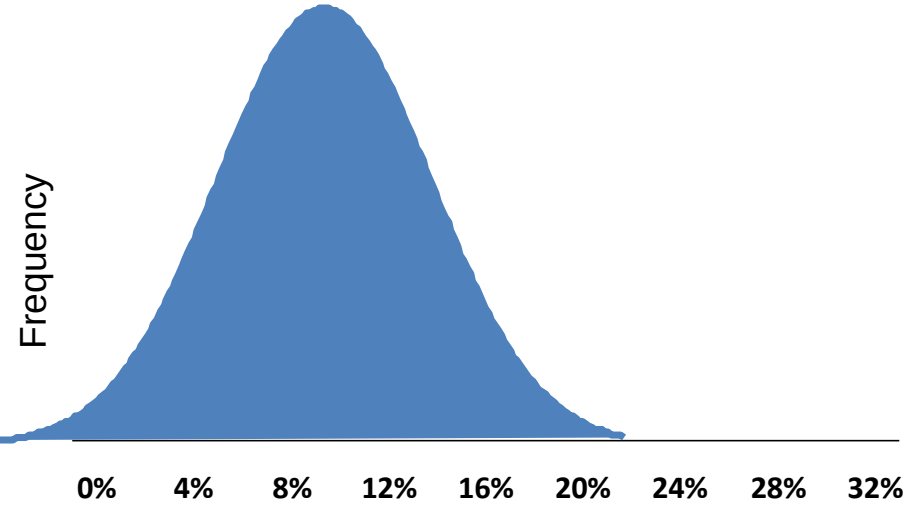
100,000
Random Samples
of n=10 college students



Average GPA

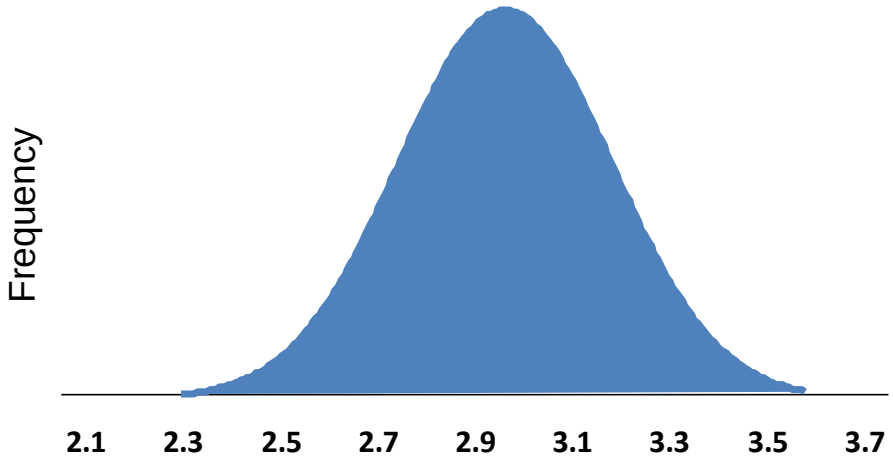
Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

Infinity
Random Samples
of n=50 college students



Percent Sociology Majors

Infinity
Random Samples
of n=10 college students



Average GPA

Note: These are **sampling distributions** … distributions of sample means and percentages… not distributions of GPA or whether people are sociology majors

Sampling Distributions

Sampling Distribution

A theoretical distribution of means or proportions, taken from an infinite number of independent random samples of size n

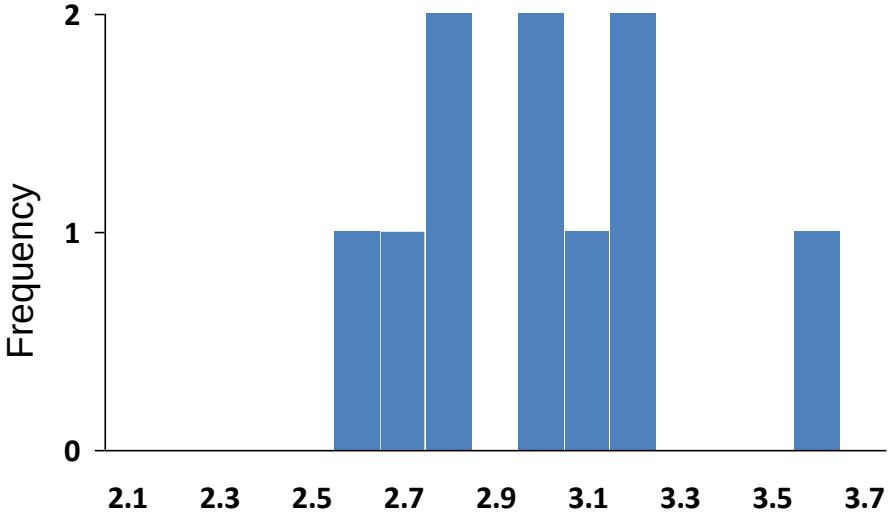
Sampling distributions of means and proportions are normal regardless of the shape of the distribution of the variable that produces the mean or proportion

Central Limit Theorem

If n is sufficiently large, then the sample means from many random samples from a population with mean μ and variance σ^2 are approximately normally distributed with mean μ and variance $\sigma^2 \Big/ \sqrt{n}$

Sampling Distributions

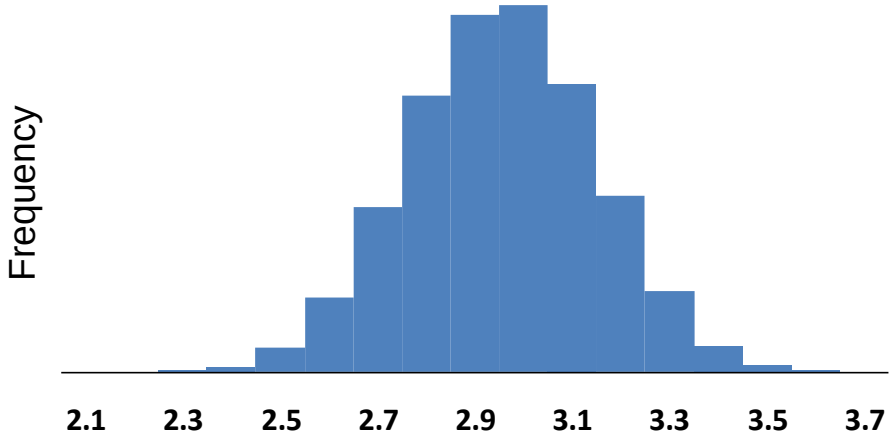
1
Random Sample
of n=10 college students



Distribution of GPA

1
Sample Mean = 3.0

100,000
Random Samples
of n=10 college students



Distribution of Sample Means

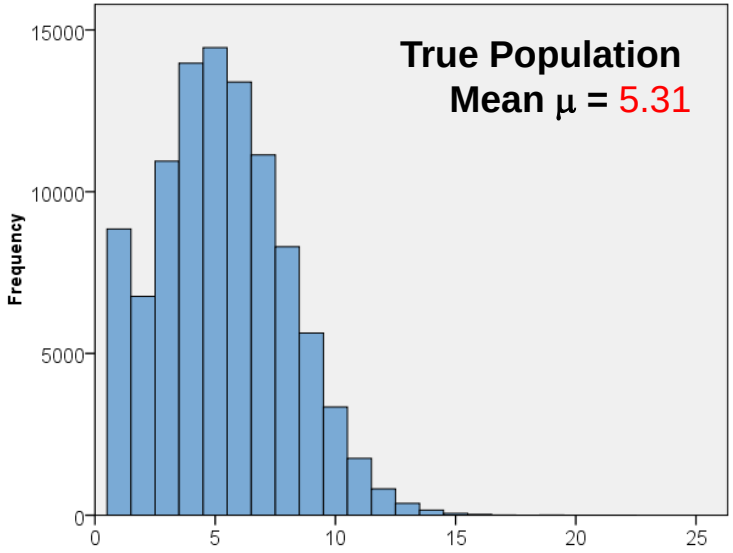
Average of all 100,000
Sample Means = 3.0

Sampling Distributions

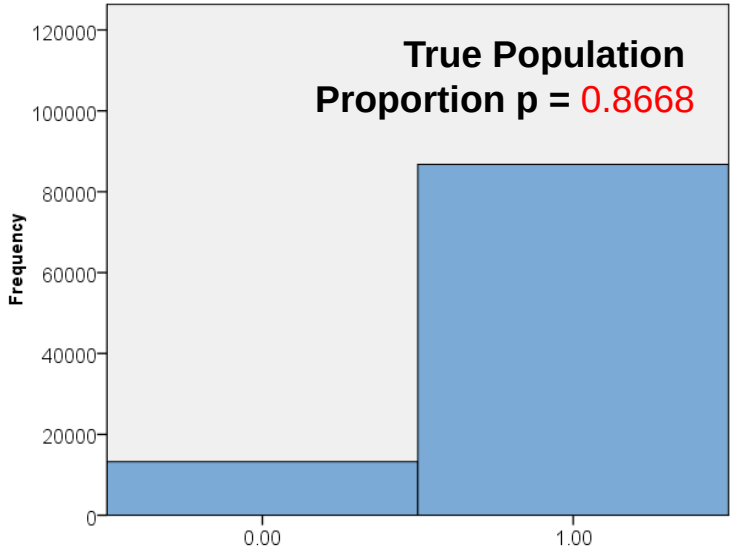
From the 1880 Census, the **population** distributions of…

"Born in the United States?"

"Number of People in the Household"



Number of People in the Household



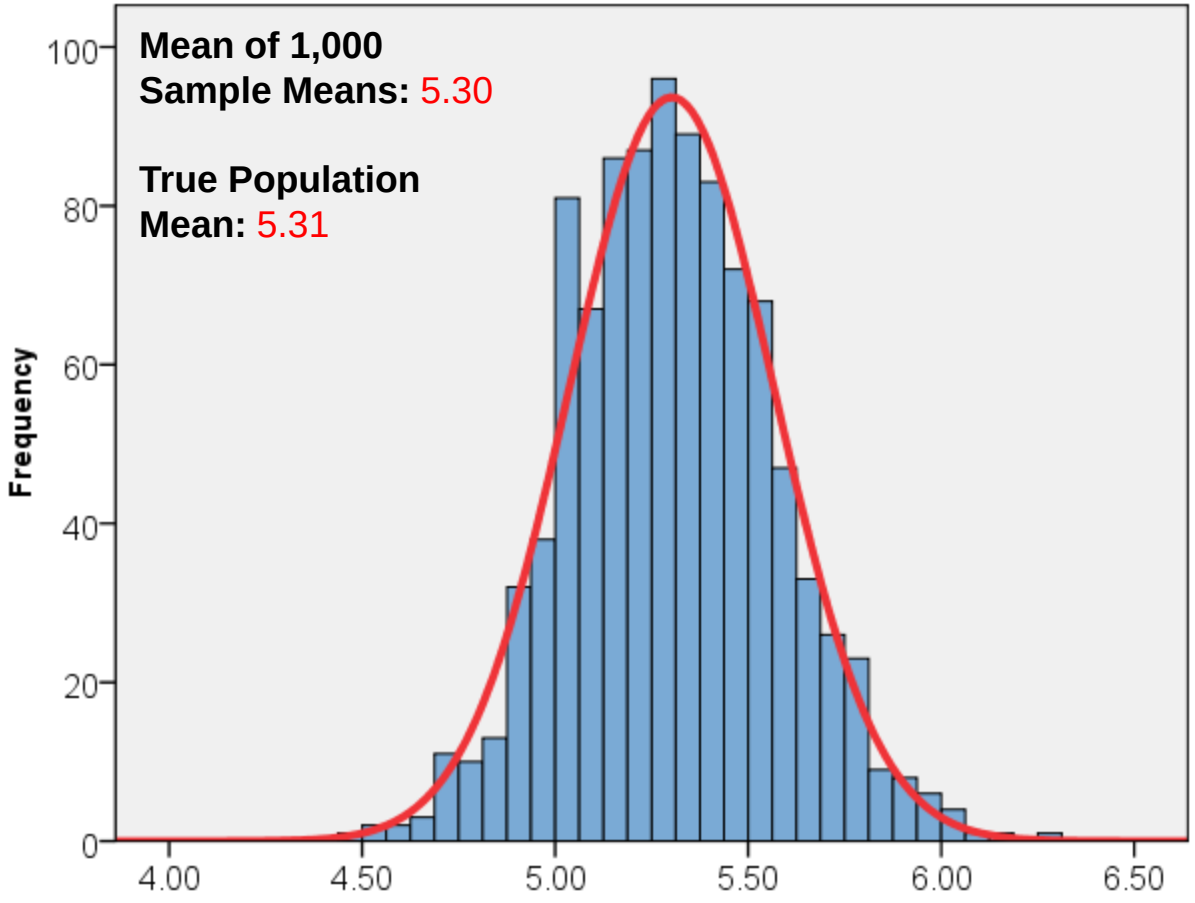
Born in the US? (0=No, 1=Yes)

Sampling Distributions

1. I sampled 100 people, and computed…
   … the sample proportion born in the U.S. ($\hat{p}$)
   … the sample mean number of people in the household ($\bar{x}$)
2. I repeated that exercise 999 more times

On the next slides are the distributions of those 1,000 sample proportions ($\hat{p}$) and sample means ($\bar{x}$); all 1,000 samples have n=100

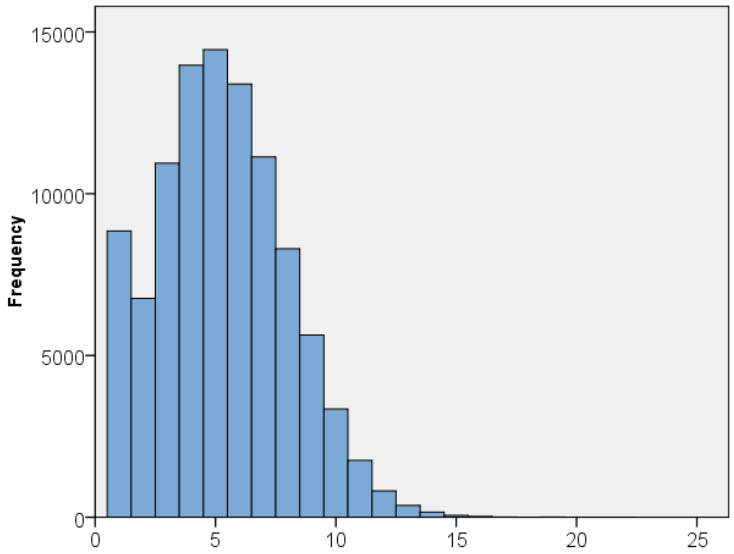
Sampling Distributions



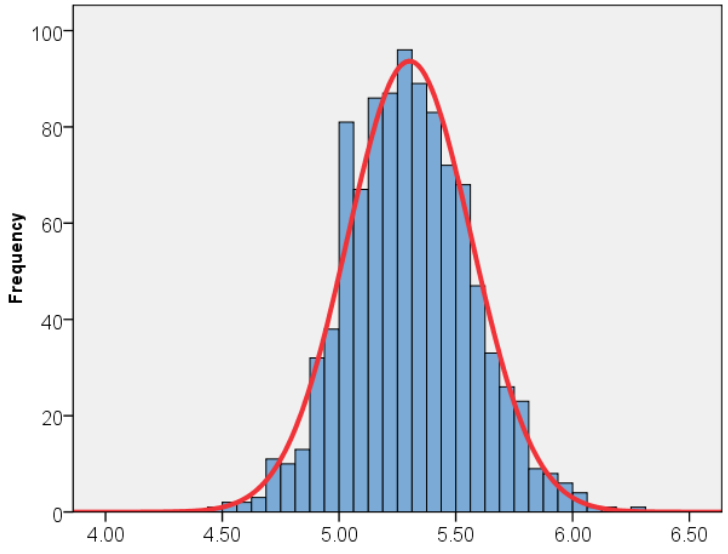
Number of People in the Household

1,000 Sample means, each w/ n=100

Sampling Distributions



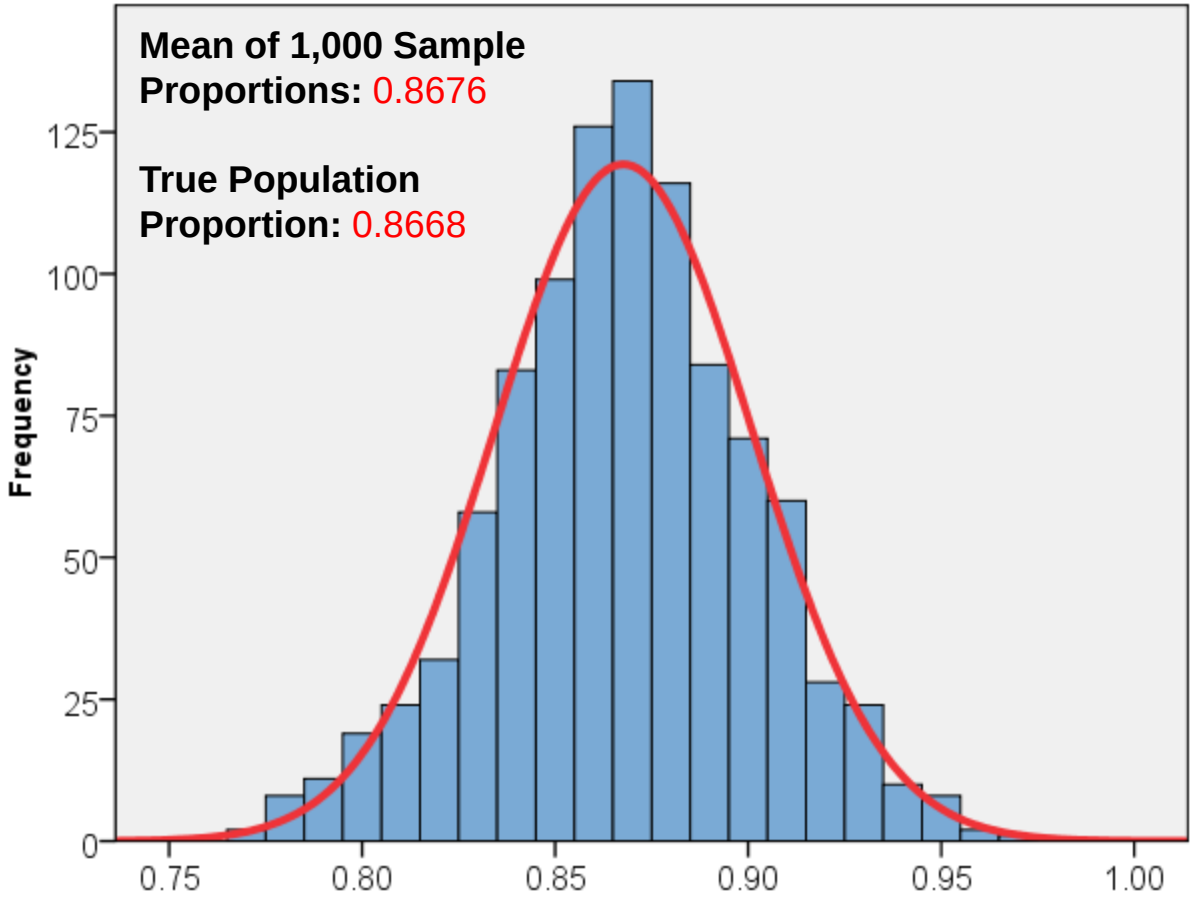
Distribution of # of People



Sampling Distribution

Number of People in the Household

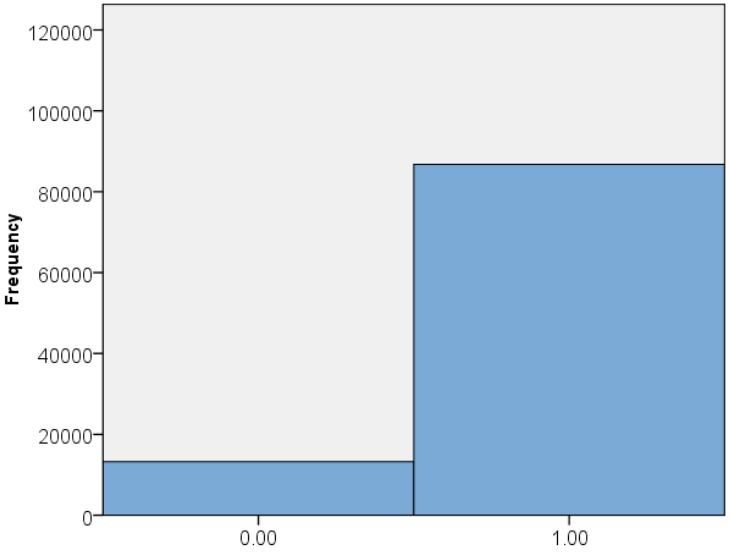
Sampling Distributions



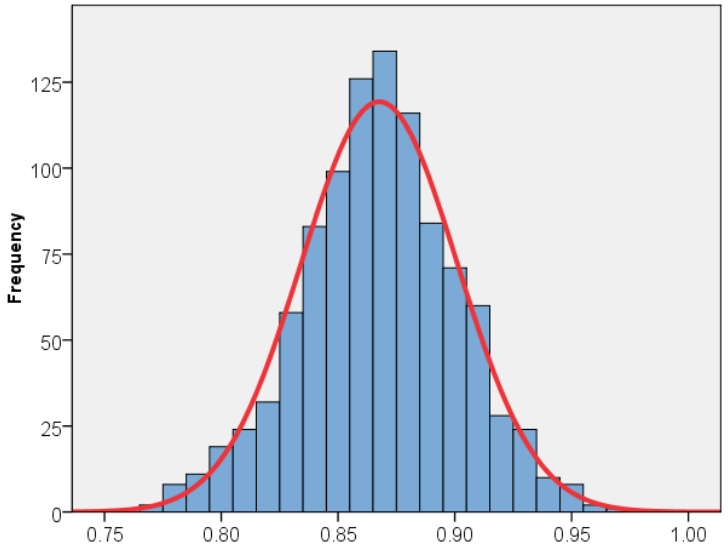
Proportion of People Born in the US

1,000 Sample proportions, each w/ n=100

Sampling Distributions



Distribution of Birthplace



Sampling Distribution

Born in the United States?

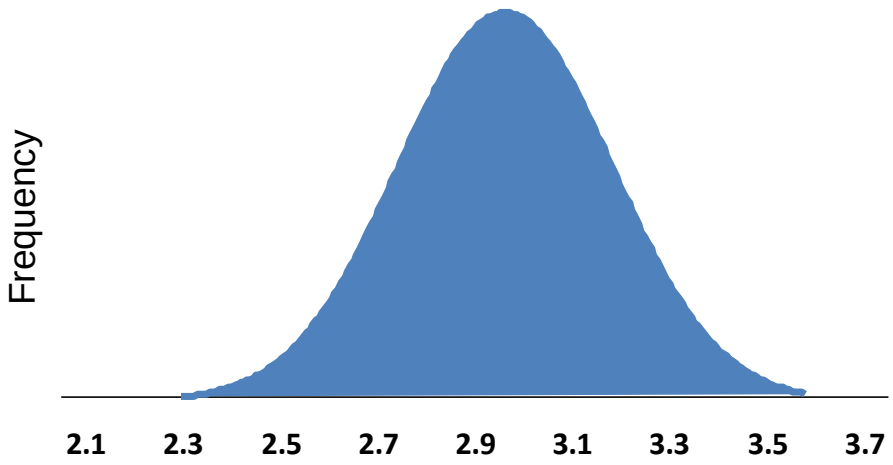
Sampling Distributions

You select **one** sample of n=10 college students

What is the probability that your sample mean GPA ($\bar{x}$) is greater than 3.5?

Infinity

Random Samples of n=10 college students



Sampling Distribution for Mean GPA

$\mu = 3.0$ $\sigma = 0.2$

Sampling Distributions

You select **one** sample of n=10 college students

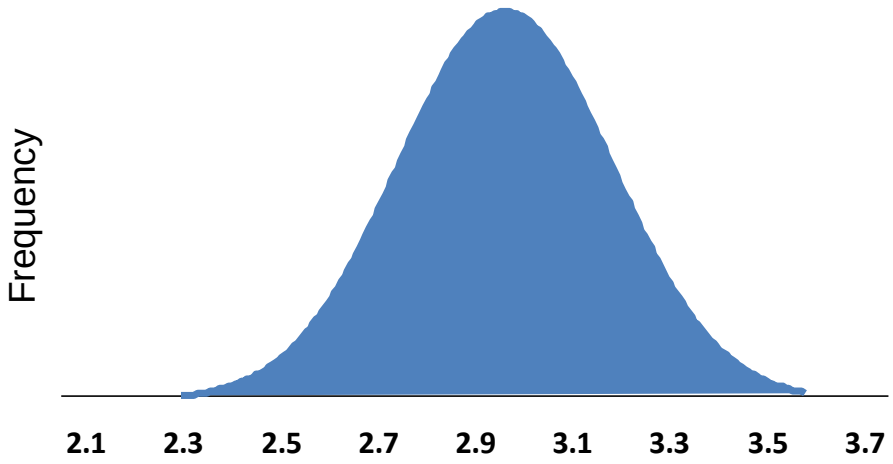
What is the probability that your sample mean GPA ($\bar{x}$) is greater than 3.5?

$$P(\bar{X} > 3.5) = P(Z > 2.5)$$

$$= 0.0062$$

Infinity

Random Samples of n=10 college students



Sampling Distribution for Mean GPA

$\mu = 3.0$ $\sigma = 0.2$

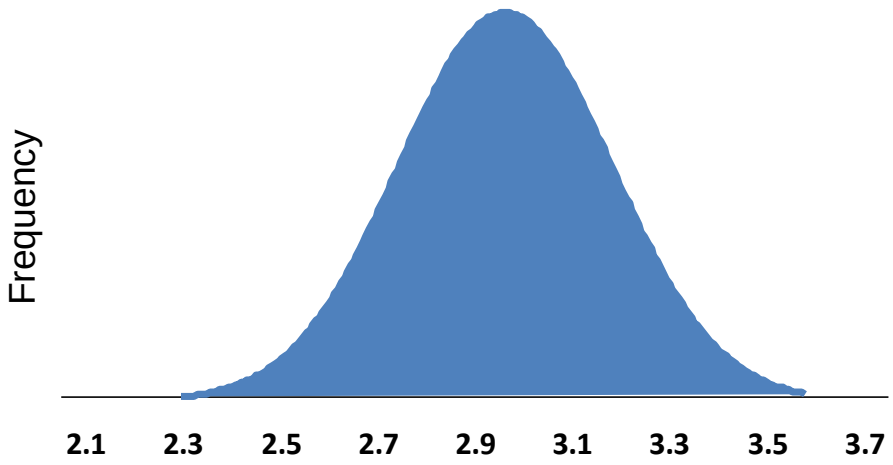
Worksheet

You select **one** sample of n=10 college students

What is the probability that your sample mean GPA ($\bar{x}$) differs from the population mean by more than ±0.4?

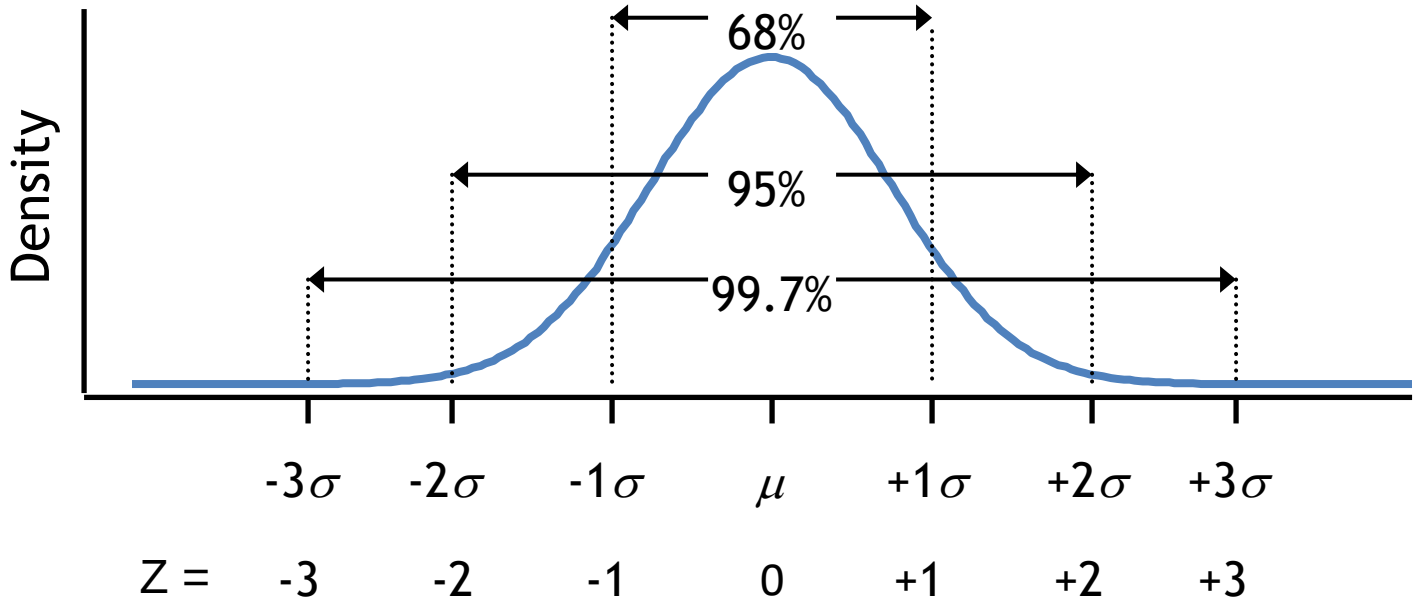
Infinity

Random Samples of n=10 college students



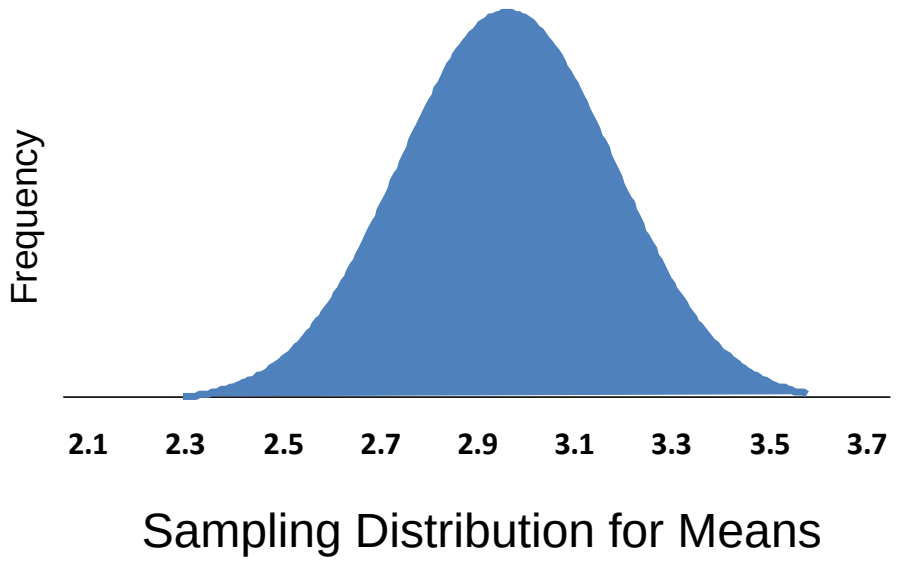
Sampling Distribution for Mean GPA

$\mu = 3.0$ $\sigma = 0.2$


Density
68%
95%
99.7%
-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
Z =
-3
-2
-1
0
+1
+2
+3

Sampling Distributions

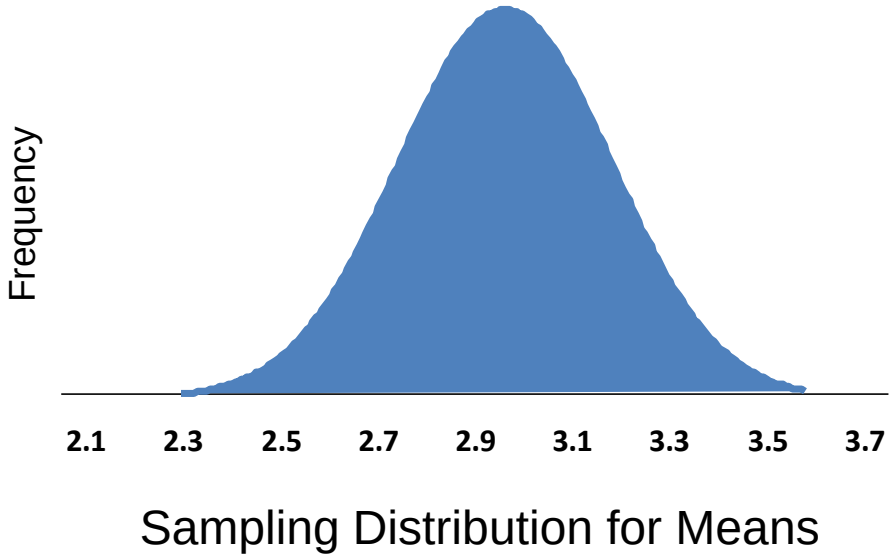
We are 95% certain that a randomly selected sample mean ($\bar{x}$) will fall within +/- 1.96 standard deviations (σ) of the true population mean (μ)



Sampling Distributions

So…

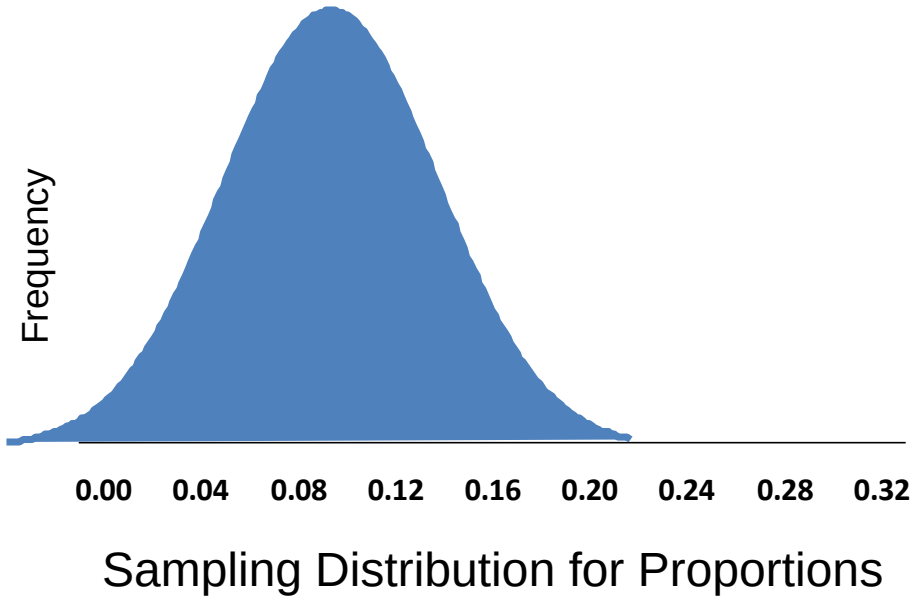
For **one** observed sample mean ($\bar{x}$), we are 95% certain that the true population mean (μ) falls within +/- 1.96 standard deviations (σ) of $\bar{x}$



Sampling Distributions

So…

For **one** observed sample proportion ($\hat{p}$), we are 95% certain that the true population proportion (p) falls within +/- 1.96 standard deviations (σ) of $\hat{p}$



Sampling Distributions

Knowing what we know about…

sampling,
Z scores,
random variables, and
sampling distributions…

…we can make inferences about population **means** (μ) and **proportions** (p) using sample means ($\bar{x}$) and proportions ($\hat{p}$)

Proportions

Proportions

Call p the population proportion

Call p-hat ($\hat{p}$) the sample proportion

Under conditions described below, if we generate many random sample of the same size, then the distribution of the several p-hats will have a mean of p and variance of

$$\frac{\sigma^2}{\sqrt{n}} = \frac{\sum_{i=1}^{k}(Y_i - p)^2 p_i}{\sqrt{n}} = \frac{(0-p)^2 p_0 + (1-p)^2 p_1}{\sqrt{n}} = \ldots = \frac{p(1-p)}{\sqrt{n}}$$

and standard deviation:

$$\sqrt{\frac{p(1-p)}{n}}$$

Proportions

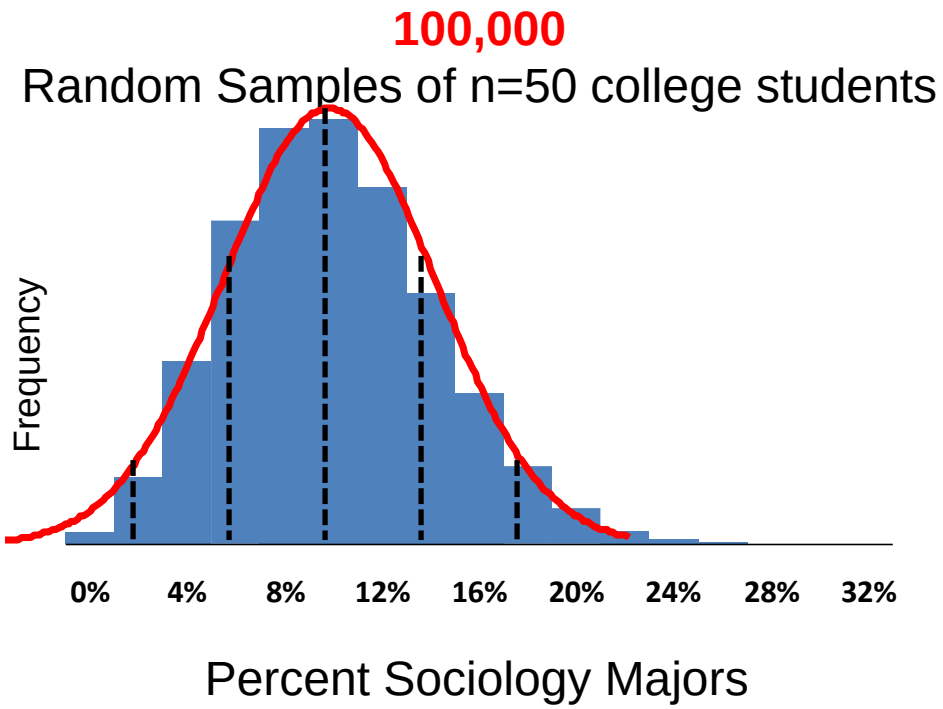
To make inferences about a population proportion based on sample data, the following things have to be true:

1. There is a population with a fixed proportion who have a certain attribute
2. The sample is randomly selected
3. The size of the sample, n, is large … generally, such that both np and n(1-p) equal at least 5 … so how big “relatively large” is depends in part on the population proportion being estimated

Proportions

1. There is a population with a fixed proportion who have a certain attribute
2. The sample is randomly selected
3. p equals about 0.10, so np=(50)(0.10)=5

   and n(1-p)=(50)(0.9)=45



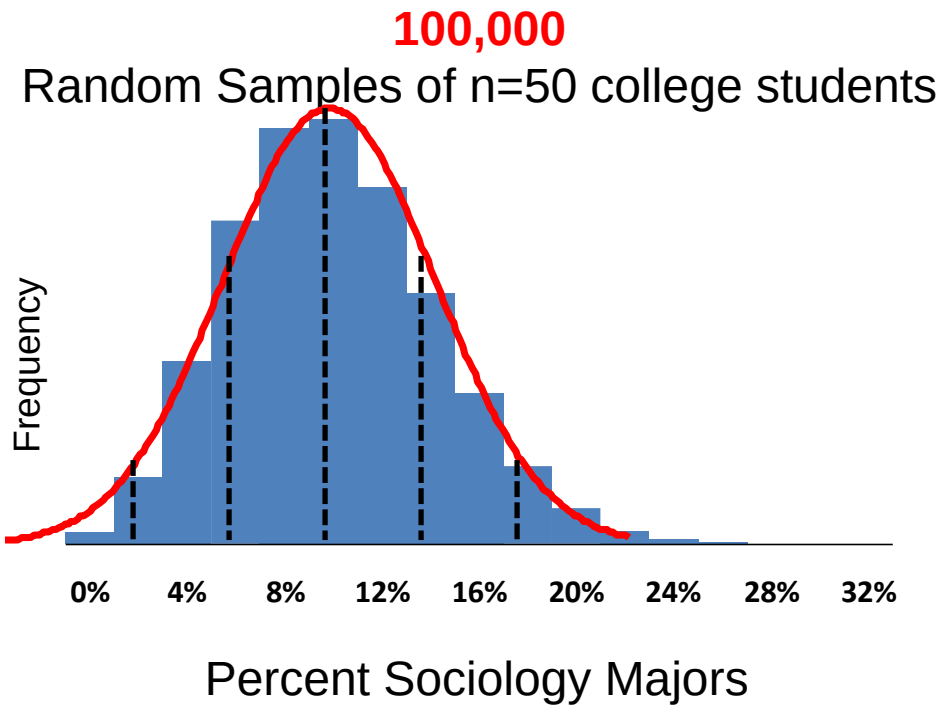
Proportions

If you have many $\hat{p}$, their distribution will be centered over p and will have standard deviation

$$\sqrt{\frac{p(1-p)}{n}}$$

In our example, the center is 0.10 and the standard deviation is

$$\sqrt{\frac{0.1(0.9)}{50}} = 0.042$$



Proportions

But what if I select only **one** random sample, with one $\hat{p}$?

What is my best guess about p? It is $\hat{p}$

What is my best guess about the standard deviation of the sampling distribution of sample proportions, $\hat{p}$?

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is called the **standard *error*** of the sampling distribution of $\hat{p}$

Proportions

I selected one random sample of 487 likely voters in Minnesota and got a sample proportion $\hat{p}$ of 0.52 who will vote for Kanye West.

What is the probability that my single $\hat{p}$ differs from the population proportion p by more than 0.05?

Proportions

I selected one random sample of 487 likely voters in Minnesota and got a sample proportion $\hat{p}$ of 0.52 who will vote for Kanye West.

What is the probability that my single $\hat{p}$ differs from the population proportion p by more than 0.05?

The standard error of the sampling distribution of $\hat{p}$ for our example is:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.52(1-0.52)}{487}} = 0.023$$

Proportions

I selected one random sample of 487 likely voters in Minnesota and got a sample proportion $\hat{p}$ of 0.52 who will vote for Kanye West.

What is the probability that my single $\hat{p}$ differs from the population proportion p by more than 0.05?

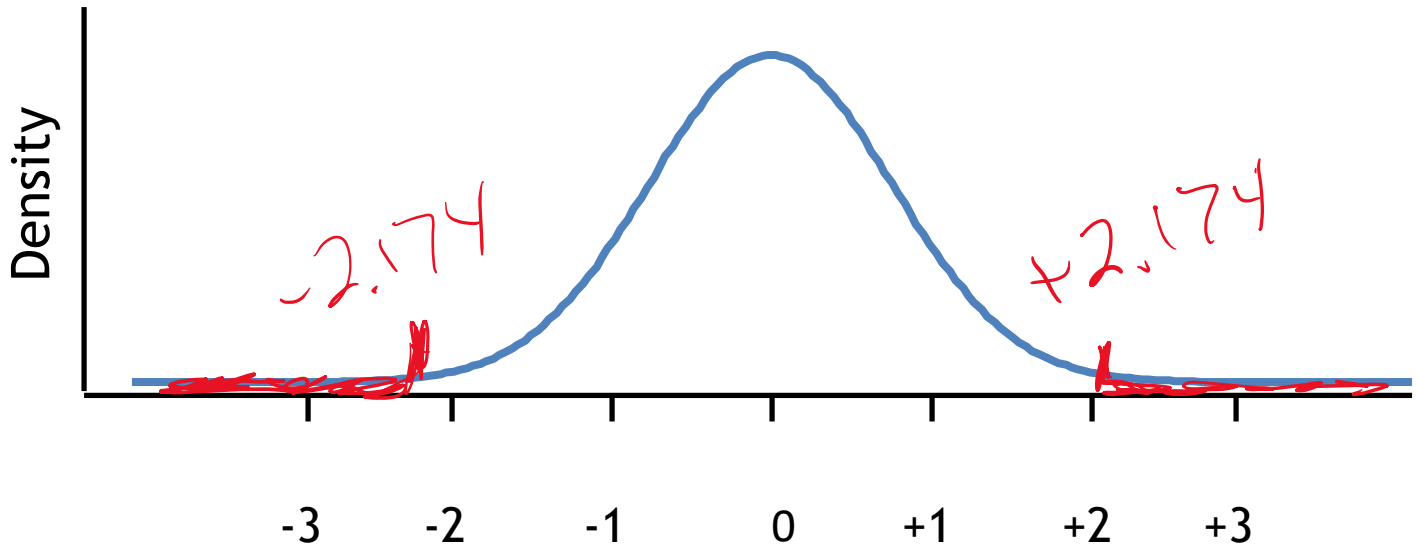
Thus the sampling distribution of $\hat{p}$ for our example has an expected mean of 0.52 and a standard error of 0.023:

Proportions

I selected one random sample of 487 likely voters in Minnesota and got a sample proportion $\hat{p}$ of 0.52 who will vote for Kanye West.

What is the probability that my single $\hat{p}$ differs from the population proportion p by more than 0.05?

The probability of a difference from p of 0.05 or more is the same as the probability of getting a Z score that differs from 0 by 0.05/0.023 = 2.174 standard errors


Density
-2.174
+2.174
-3
-2
-1
0
+1
+2
+3

Proportions

I selected one random sample of 487 likely voters in Minnesota and got a sample proportion $\hat{p}$ of 0.52 who will vote for Kanye West.

What is the probability that my single $\hat{p}$ differs from the population proportion p by more than 0.05?

The probability that our single sample proportion differs from the population proportion by 0.05 or more is

$P(Z>2.174) + P(Z<-2.174) = 0.015 + 0.015 =$ **0.03**

Worksheet

I selected one random sample of 200 people and got a sample proportion $\hat{p}$ of 0.4.

What is the probability that my $\hat{p}$ differs from p by more than 0.035?

Means

Means

The same sort of logic can be applied to the sampling distribution of sample means

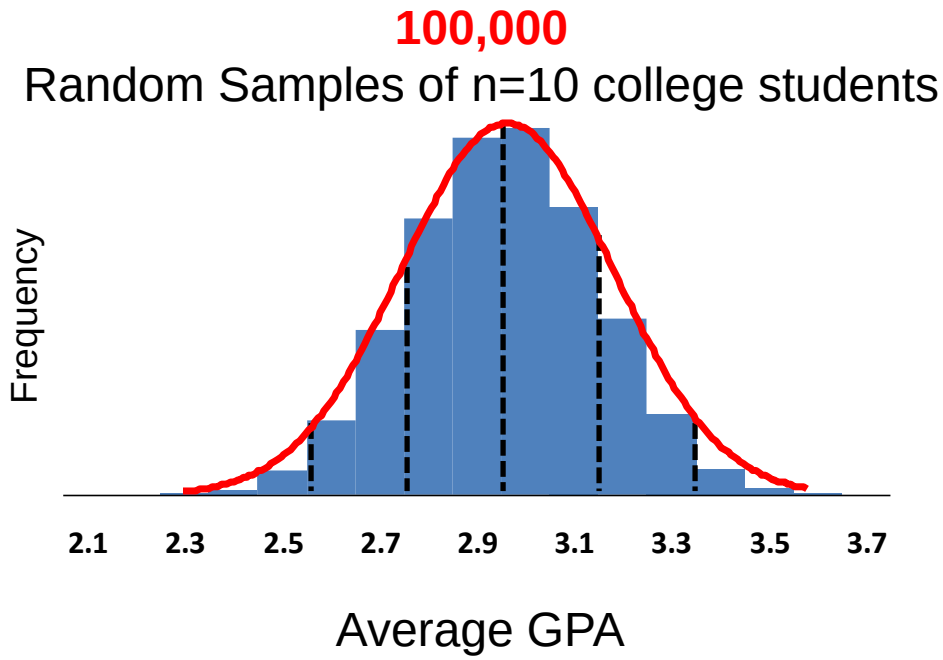
In the population, the mean is μ and the standard deviation is σ

In any sample, the mean is $\bar{x}$ and the standard deviation is s

From the central limit theorem, the sampling distribution of means is centered over μ and has variance $\sigma^2 / \sqrt{n}$

Means

For our 100,000 sample means of GPA … each based on a random sample of n=10 college students … the mean value is 3.0 with a standard deviation of 0.2



Means

But what if I select only **one** random sample, with one $\bar{x}$?

What is my best guess about μ? It is $\bar{x}$

What is my best guess about the standard deviation of the sampling distribution of sample means, $\bar{x}$?

$$\sqrt{s^2/n}$$

This is called the **standard *error*** of the sampling distribution of $\bar{x}$

Means

I selected one random sample of 500 people and got a sample mean $\bar{x}$ of 180 with a standard deviation, s, of 30

What is the probability that my $\bar{x}$ differs from μ by more than 2?

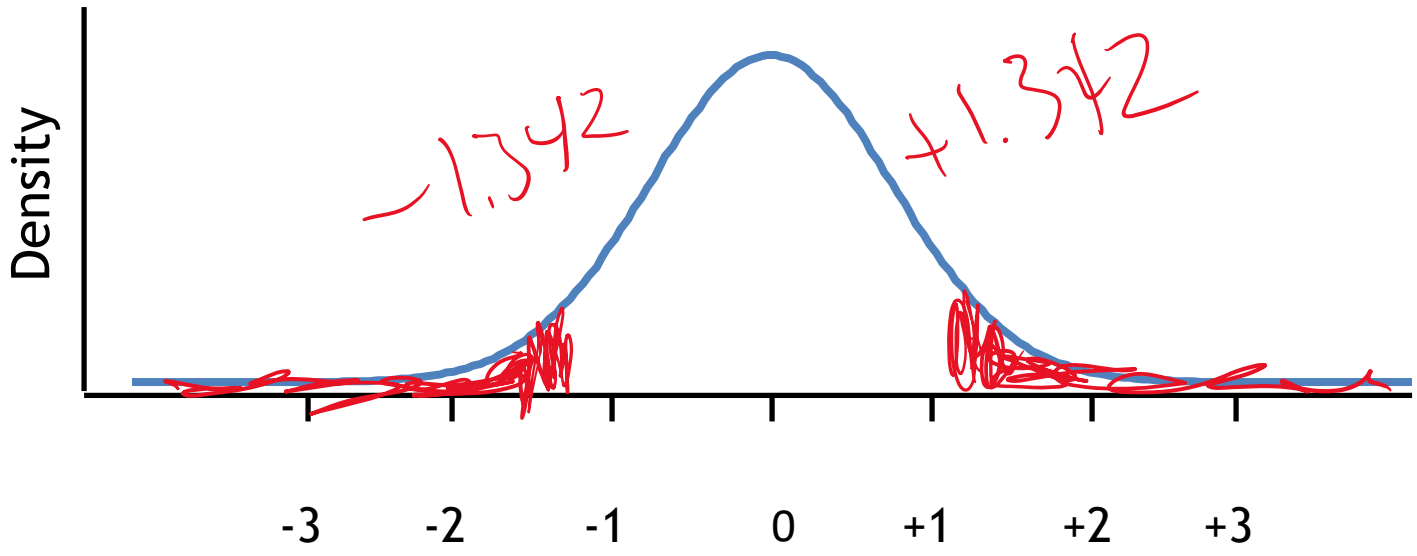
Means

I selected one random sample of 500 people and got a sample mean $\bar{x}$ of 180 with a standard deviation, s, of 30

What is the probability that my $\bar{x}$ differs from μ by more than 2?

The standard error of the sampling distribution of $\bar{x}$ for our example is:

$$\sqrt{s^2/n} =$$
$$\sqrt{30^2/500} =$$
$$1.342$$


Density
-1.342
+1.342
-3
-2
-1
0
+1
+2
+3

Means

I selected one random sample of 500 people and got a sample mean $\bar{x}$ of 180 with a standard deviation, s, of 30

What is the probability that my $\bar{x}$ differs from μ by more than 2?

The probability that our single sample mean differs from the population mean by 2 or more is

$P(Z>1.342) + P(Z<-1.342) = 0.0901 + 0.0901 =$ **0.1802**

Worksheet

I selected one random sample of 1,000 people and got a sample mean $\bar{x}$ of 200 with a standard deviation, s, of 50

What is the probability that my $\bar{x}$ differs from μ by more than 3?

t distribution

When n is large … say, more than 50 … sampling distributions of means follow the Z distribution

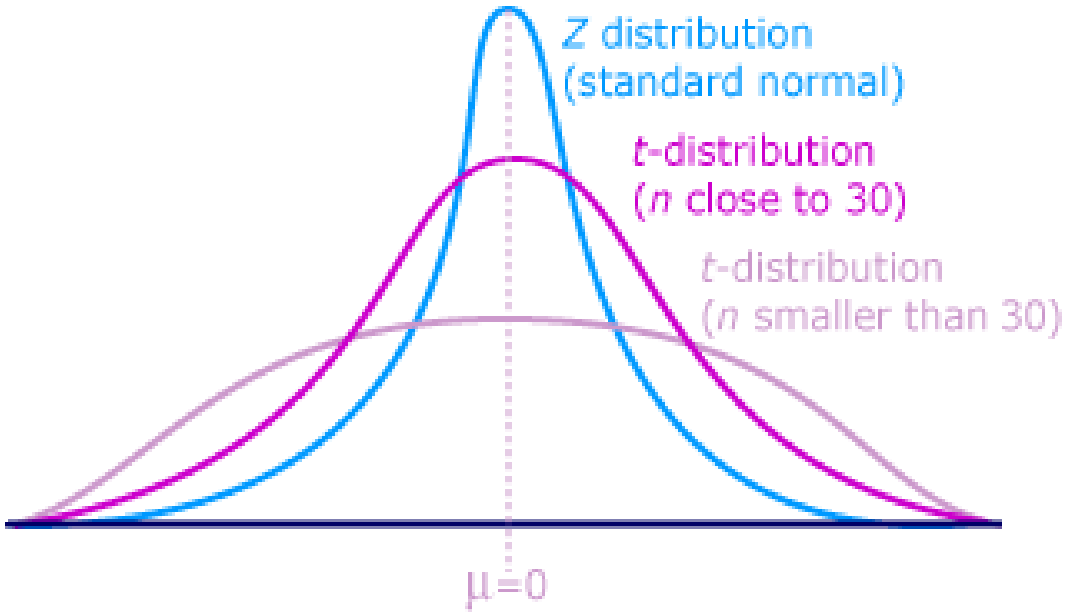
When n is smaller … say, less than 50 … sampling distributions of means follow **t distributions**, not Z distributions

There is a different t distribution for each value of n; each t distribution is defined by its degrees of freedom, df, where df equal n-1

t distribution

t scores are computed the same way as Z scores

$$Z = \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \qquad t = \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}$$



t distribution

When n is large, the t and the Z distributions are (approximately) the same and so the area under the curve within any given range of Z or t scores is the same

When n is small, use the t distribution with n-1 degrees of freedom

To be safe, in practice most people always use the t distribution for sampling distributions of means

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

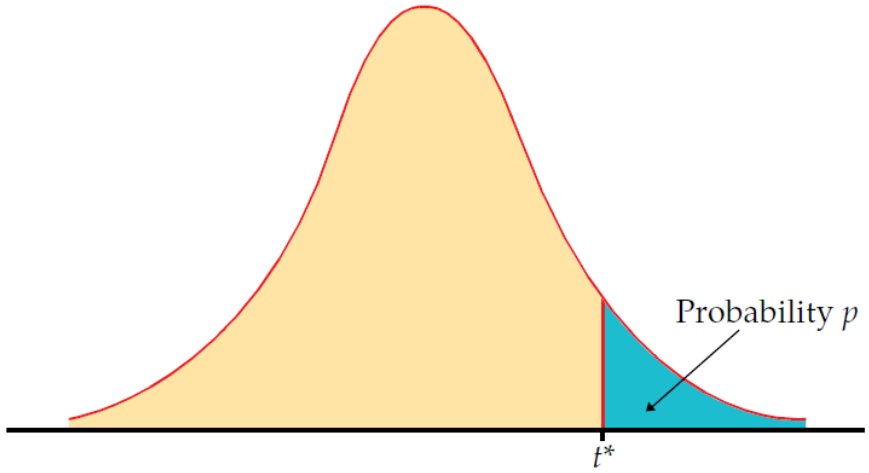


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965

Worksheet

I selected one random sample of 20 people and got a sample mean $\bar{x}$ of 200 with a standard deviation, s, of 50

What is the probability that my $\bar{x}$ differs from μ by more than 18?

Why is this Useful?

1. Confidence Intervals

Based on the distribution of Y in sample data, we are confident that the distribution of Y in the population has particular qualities
(e.g., that its mean is within a certain range of values)

"With 95% certainty, I conclude based on my sample data that between 25% and 35% of everyone in the population has been arrested"

Why is this Useful?

2. Hypothesis Tests

Based on the distribution of Y in the sample data, we can evaluate the likely truth of theoretically-informed hypotheses about the distribution of Y in the population (e.g., that the mean of X is above some value)

"With 95% certainty, I reject the claim that fewer than 20% of everyone in the population has ever been arrested"

Want More?

Parts A through E of David Lane's book

http://onlinestatbook.com/2/sampling_distributions/sampling_distributions.html

Chapter 6 of Lowry's book

http://vassarstats.net/textbook/

This section of Jerry Dallal's book

http://www.jerrydallal.com/LHSP/meandist.htm

Stat Trek's discussion

http://stattrek.com/sampling/sampling-distribution.aspx