Math 140
Introductory Statistics

An exercise for you

Utah's national parks

National Park	Area (sq mi)
Arches (A)	119
Bryce Canyon (B)	56
Canyonlands (C)	527
Capitol Reef (R)	378
Zion (Z)	229

Last time we created the sampling distribution for the total number of square miles in any 2 parks.

2 size sample

Sample of Two Parks	Total Area (sq mi)	Mean Area (sq mi)
A and B	175	87.5
A and C	646	323.0
A and R	497	248.5
A and Z	348	174.0
B and C	583	291.5
B and R	434	217.0
B and Z	285	142.5
C and R	905	452.5
C and Z	756	378.0
R and Z	607	303.5
	Mean	261.8

5 choose 2 possibilities =
10 combinations
Mean is 261.8
SD is 105.23

Try again for a sample of size 3

A and B and C

Total area

Mean Area from

the 3 sample

How many combinations are possible?

Mean is ?

SD is ?

Do for a sample size of 4 and 5

Sample size	Combination possibilities	Mean	SD (standard error)
N=2	5 choose 2 = 10	261.8	105.23
N=3	5 choose 3 = 10	261.8	?
N=4	5 choose 4 = 5	261.8	?
N=5	5 choose 5 = 1	261.8	0

You should see the standard error decreases, as n increases

7.2 Shape, center and sampling distributon of the mean

We want to estimate the total number
of children in US households

We will take a random sample of families
And compute the mean on those families

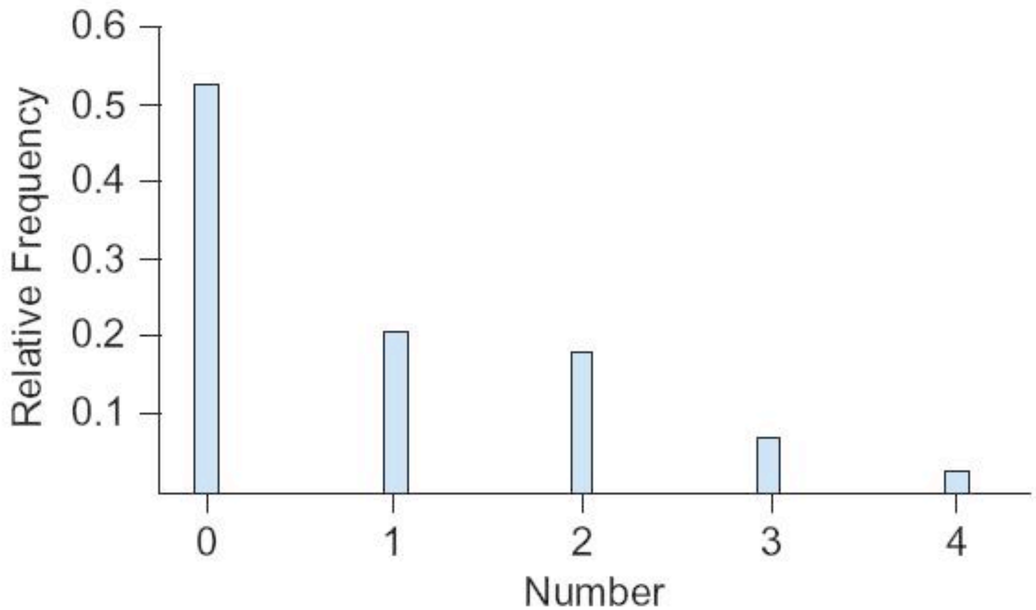
What is the best sample size to use?

Should I make groups of 4, 10, 20, 40 or more?
And average on those? How do decide sample size?

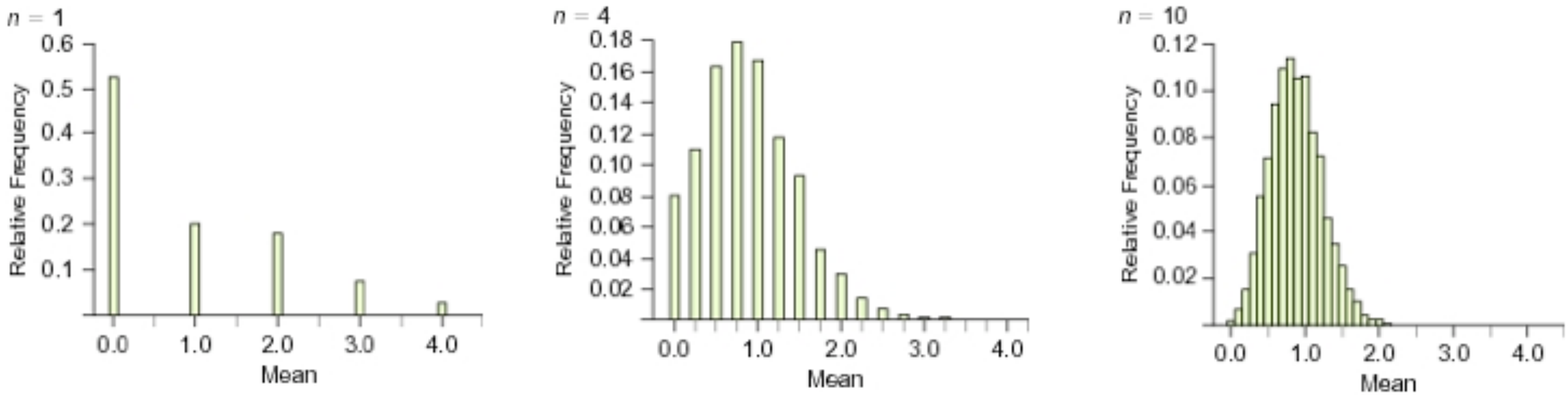
According to the Census Bureau

BUT WE DON'T KNOW THIS

Number of Children	Proportion of Families
0	0.524
1	0.201
2	0.179
3	0.070
4 (or more)	0.026



Try with different sampling sizes n

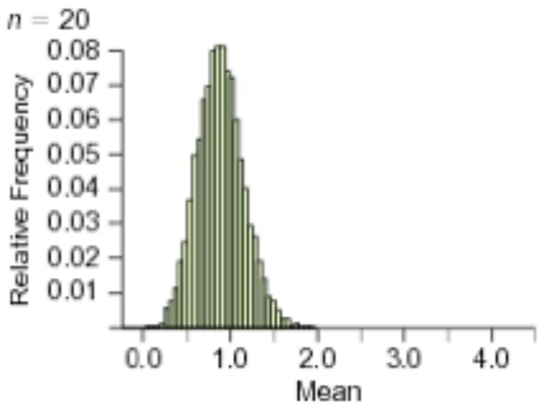


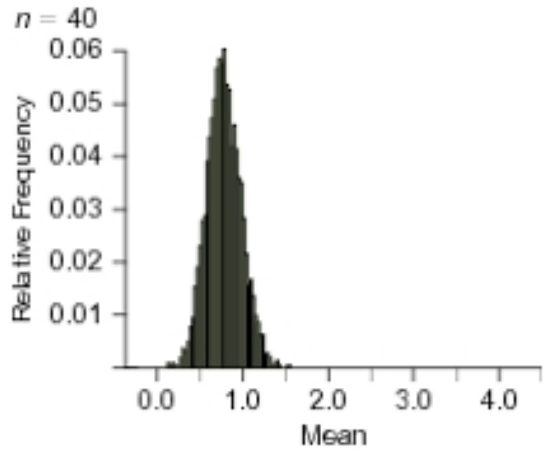
n=1 one family at a time

n=4, n=10

When we look at higher values of n we have more outcomes

Try with different sampling sizes n





Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1

n=20, n=10

The mean is the same and
the standard error becomes smaller as n increases

Noteworthy observations

For n=1 the sampling distribution is skewed
towards the right (more values close to zero)

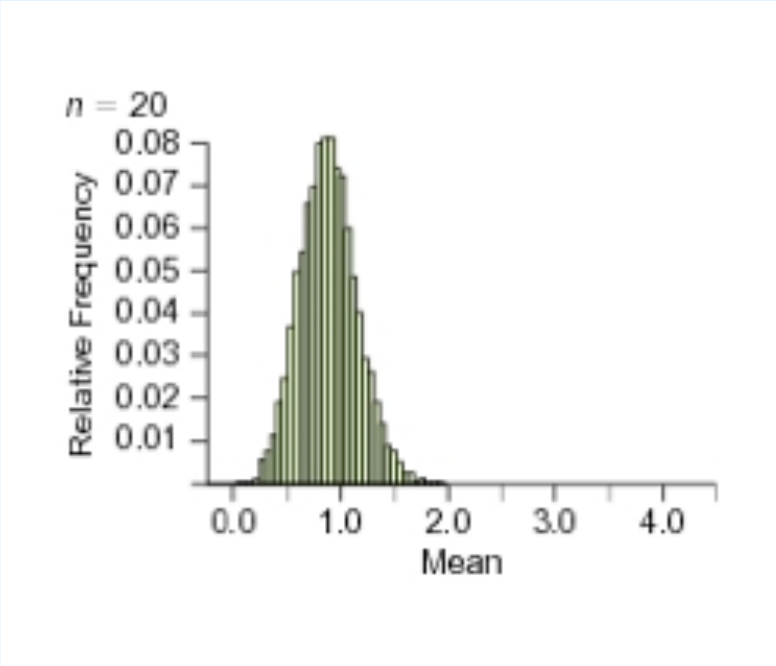
As n increases, the skew disappears

As n increases, the sampling distribution starts
Looking normal

The mean is the same 0.9 children per family

The standard error becomes smaller as n increases

Choose n=20

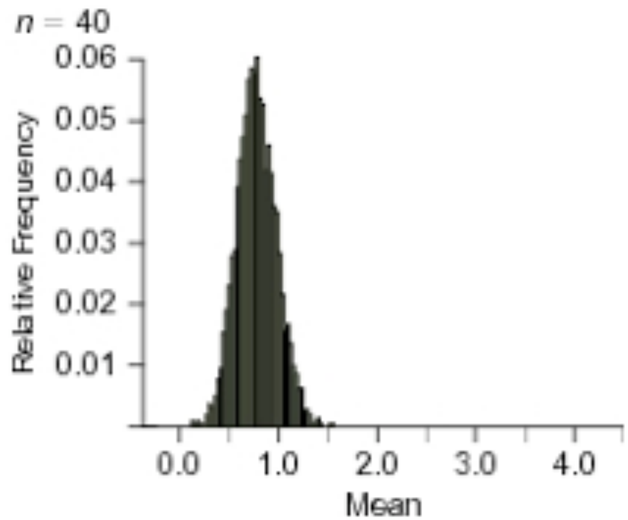


Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1

There is a 95% chance that all values between 0.9 - 1.96*0.25 and 0.9 + 1.96*0.25 are reasonably likely.

All values between 0.41 and 1.39 are likely.
This includes our estimate of 0.9

Choose n=40 - a tighter fit



Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1

There is a 95% chance that all values between 0.9 - 1.96*0.17 and 0.9 + 1.96*0.17 are reasonably likely.

All values between 0.67 and 1.23 are likely. This includes our estimate of 0.9

Calculating means from sampling distributions

Properties of the Sampling Distribution of the Sample Mean, $\bar{x}$

If a random sample of size n is selected from a population with mean μ and standard deviation σ, then the sampling distribution of $\bar{x}$ has these properties.

- The mean, $\mu_{\bar{x}}$, equals the mean of the population, μ:

$$\mu_{\bar{x}} = \mu$$

- The standard deviation, $\sigma_{\bar{x}}$, sometimes called the standard error of the mean, equals the standard deviation of the population, σ, divided by the square root of the sample size n:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

What does this mean?

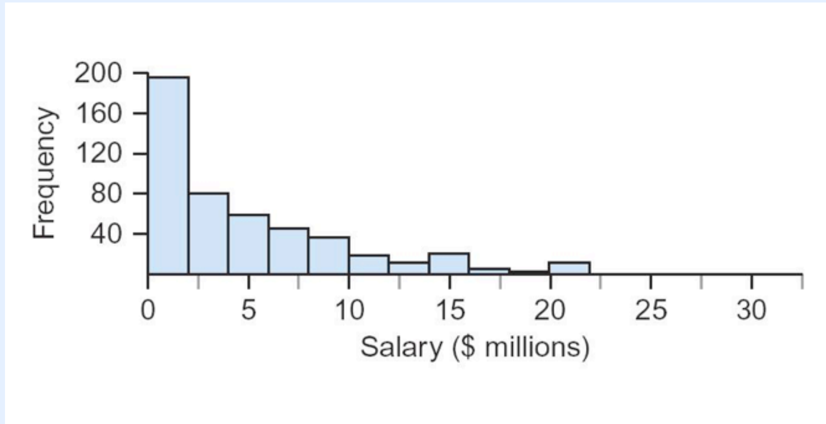
1. It does not matter what the underlying distribution looks like, the mean calculated from all random sampling

IS THE SAME AS THAT OF THE UNDERLYING POPULATION

This is true for symmetric and non-symmetric distributions

We saw this in the case of the NBA salaries

What does this mean?



This was a skewed distribution with average
$4.6 million

When we did the random sampling, the average
was still $4.6 million

True always if I use all possible samples

What does this mean?

2. It turns out that the standard error
(calculated from your sampling distribution)
is the same as the standard deviation of the population
divided by the square root of n

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

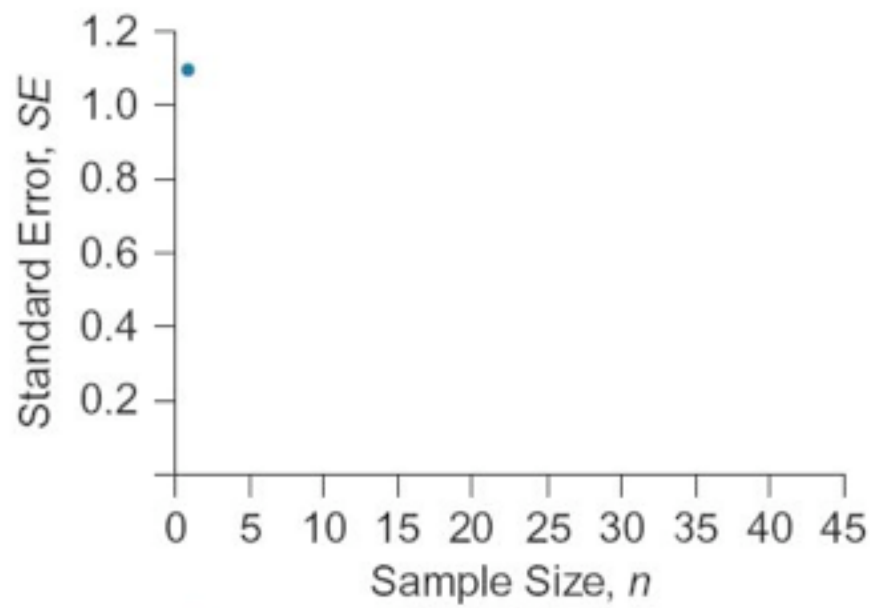
This tells us that indeed the standard
error gets smaller as n gets
larger

Finally

- The shape will be approximately normal if the population is approximately normal. For other populations, the sampling distribution becomes more normal as n increases. (This property is called the **Central Limit Theorem**.)

You try

Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1

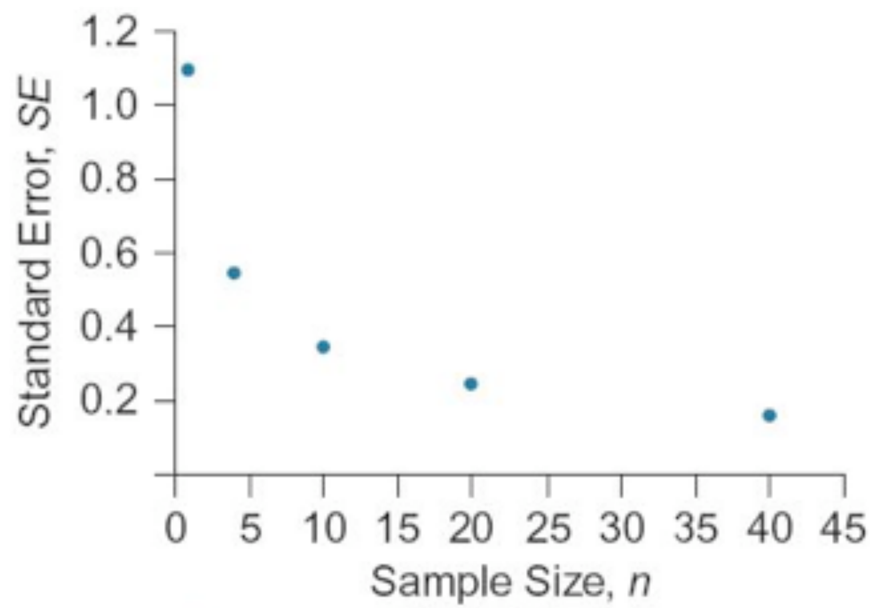


Estimate what the standard error should be when n=30
And the 95% confidence interval

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

You try

Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1

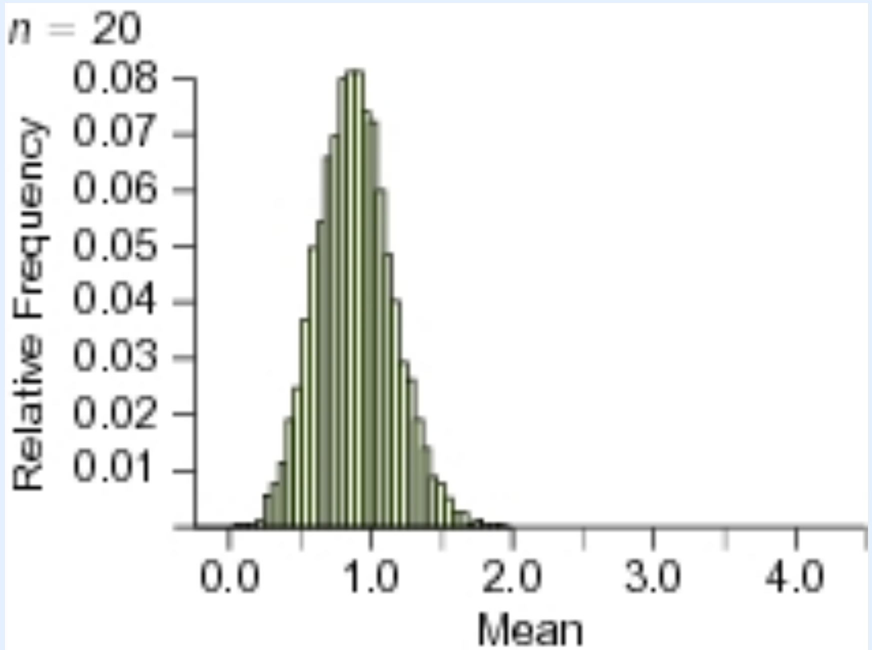


$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.1}{\sqrt{30}} = 0.20$$

For n=30 the standard error is about 0.20

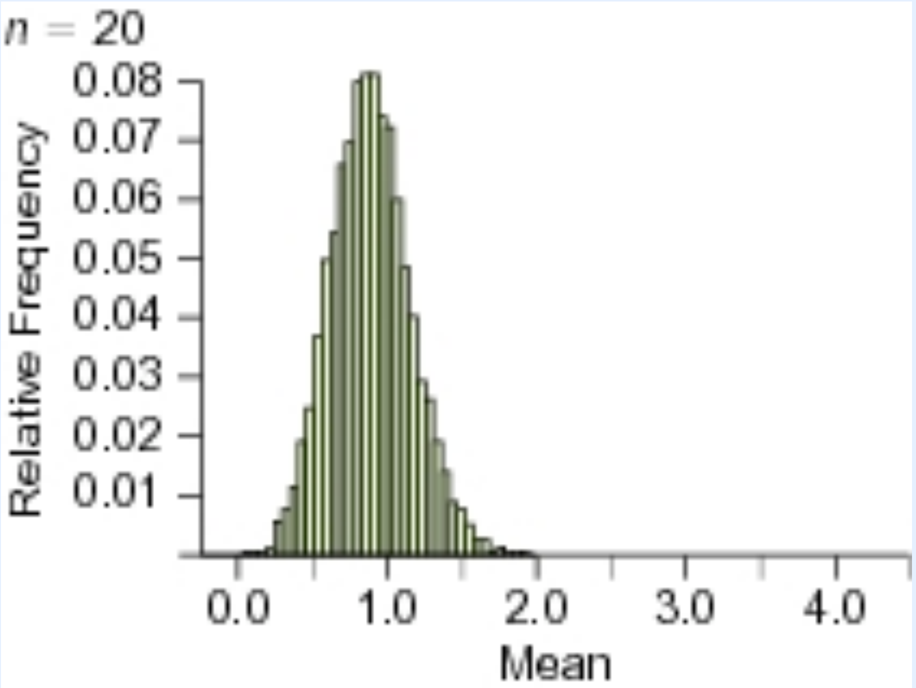
The 95% confidence interval is between 0.52 and 1.49

Let’s go back to the kids



What is the probability that a random sample of 20 families will have an average of 1.5 kids or less?

Let’s go back to the kids

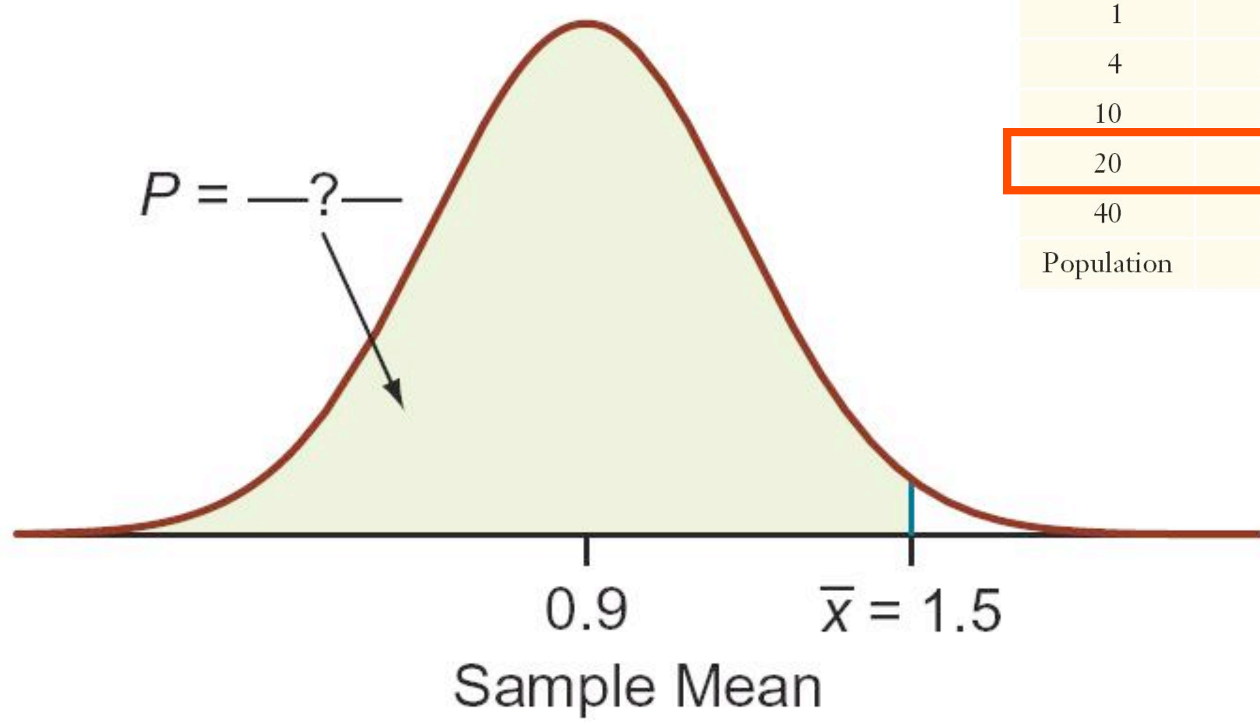


Well, it looks normal, so let’s use what we know about normal distributions!

Center = 0.9, standard error = 0.25

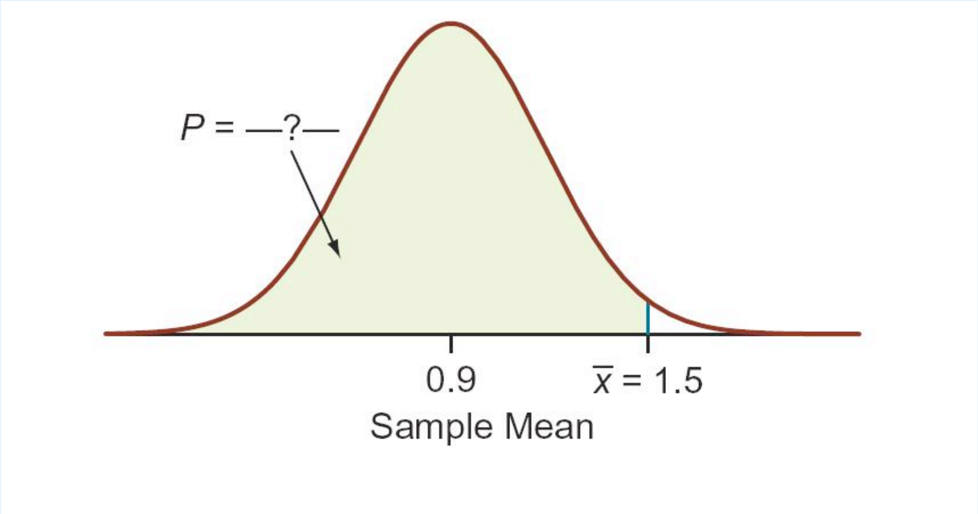
Let's go back to the kids

Sample Size, n	Mean	Standard Error, $\sigma_{\bar{x}}$
1	0.9	1.1
4	0.9	0.55
10	0.9	0.35
20	0.9	0.25
40	0.9	0.17
Population	0.9	1.1



What is the probability that a random sample of 20 families will have an average of 1.5 kids or less?

Recall rescaling and recentering?

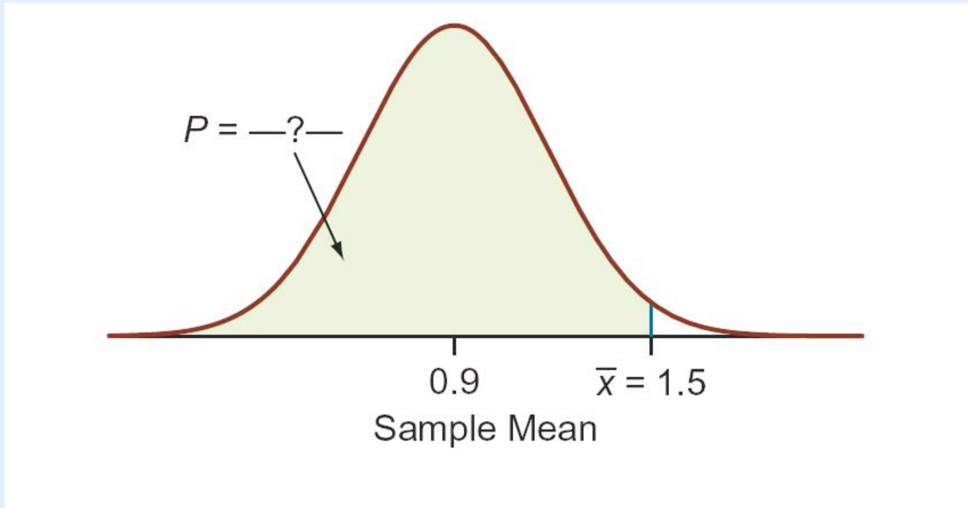


We need to calculate the corresponding z-score!

$$z = \frac{x - mean}{std\ deviation} = \frac{1.5 - 0.9}{0.25} = 2.4$$

The x we care for is 1.5. We need to find the corresponding z so we can check on page 759

Recall rescaling and recentering?



From Table A

The probability is 0.9918

For you

What is the probability that we find 30 kids total
From a random sample of 20 families in the US?

For you

What is the probability that we find 30 kids total
From a random sample of 20 families in the US?

A little bit of a trick question.
This means that PER FAMILY
we find 30/20 kids=1.5

The answer is the problem we just did!

Corn yields in Indiana

Let's select random plots of the size 1/1000 of an acre.
This is about 18 feet of 1 row of corn.

On average each plot has a yield of about 15,000 kernels
With SD of 2000. This is the POPULATION DATA

Suppose we decide to randomly sample n=25 plots

What is the approximate probability that the mean number of kernels will exceed 16,000?

What are reasonably likely outcomes for the mean yield of these n=25 sample plots?

Corn yields in Indiana

We expect the mean of the sample to be?

We expect the standard error of the sample to be?

Then take this info to Table A

Corn yields in Indiana

We expect the mean of the sample to be 15,000

We expect the standard error of the sample to be

$$\sigma = \frac{2000}{\sqrt{25}} = \frac{2000}{5} = 400$$

The z-score is?

Corn yields in Indiana

Mean = 15,000

$$\sigma = \frac{2000}{\sqrt{25}} = \frac{2000}{5} = 400$$

The z-score is

$$z = \frac{x - 15,000}{400} = \frac{16,000 - 15,000}{400} = 2.5$$

From table A the probability is 1-0.9938 = 0.0062

Corn yields in Indiana

Reasonably likely values are between

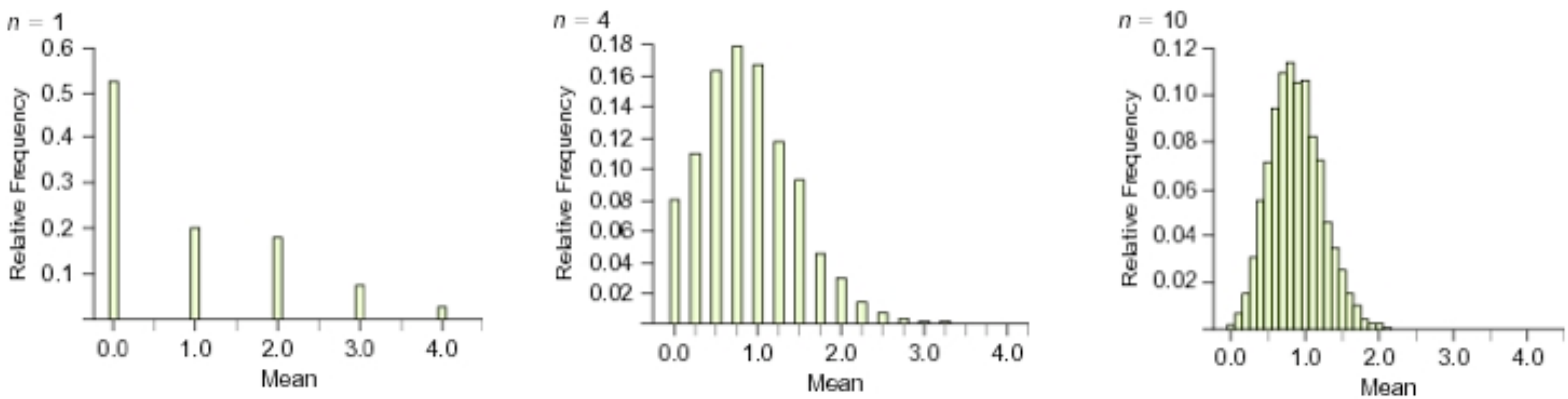
15,000 - 1.96*400 = 15,000 - 794 = 14,216

And

15,000 + 1.96*400 = 15,000 + 794 = 15,784

Can this always be done?

We need to make sure the sampling distribution is normally distributed



For example, n=1, n=4 are not really normal
n=10 begins to look normal

1. We can use z-scores only if the random sample distribution is normal

So for the case of the NBA
players we should be careful,
because the distribution was skewed
but if the underlying population is itself normally
distributed then we can use z-scores.

In general

When Can the Sampling Distribution of the Mean Be Considered Approximately Normal?

- If the population is approximately normally distributed, you can assume that the sampling distribution of $\bar{x}$ is approximately normal too, no matter what the sample size.
- If the sample size is 40 or more, it's pretty safe to assume that the sampling distribution of $\bar{x}$ is approximately normal.
- Using the normal approximation may be reasonably accurate with smaller sample sizes if the population isn't too badly skewed.

Also

2. The mean of the random sampling distribution is the same as the mean of the population

ALWAYS!

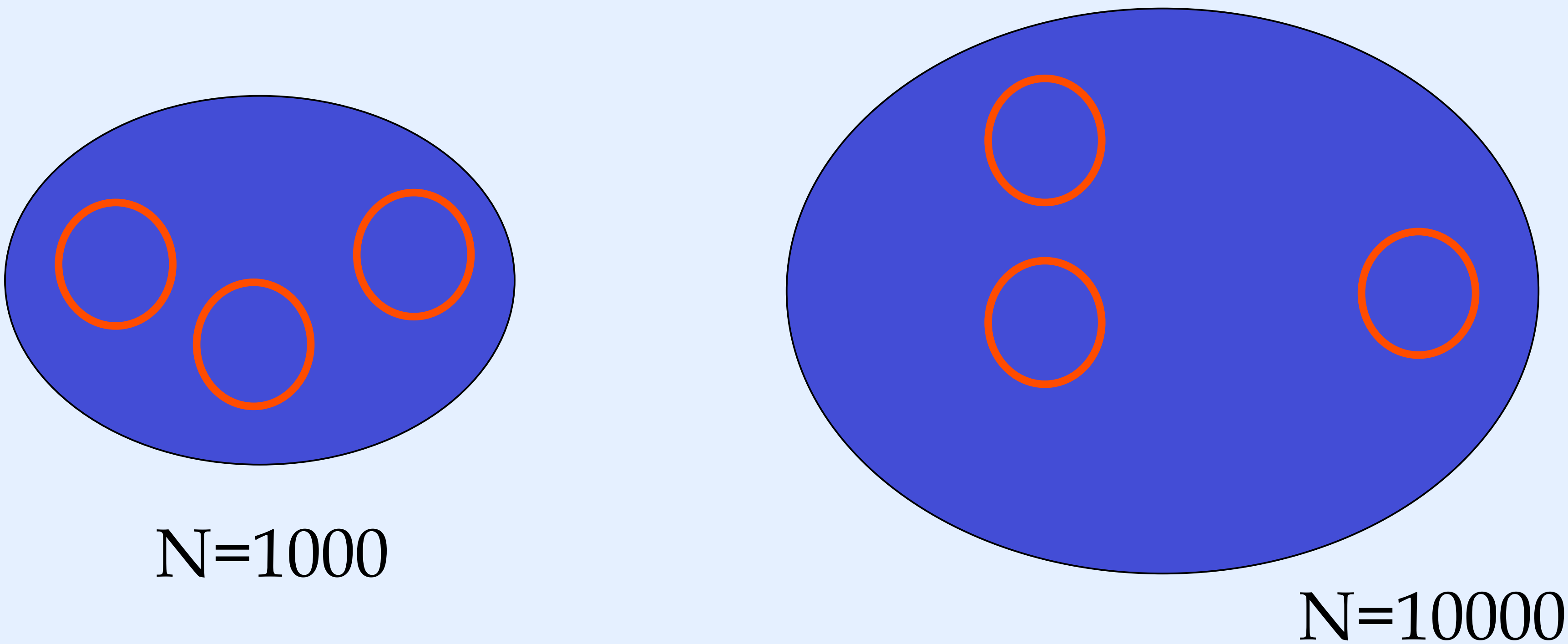
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. This too we can use almost always, only very rare exceptional cases when not appropriate

Questions

Using the Properties of the Sampling Distribution of the Mean

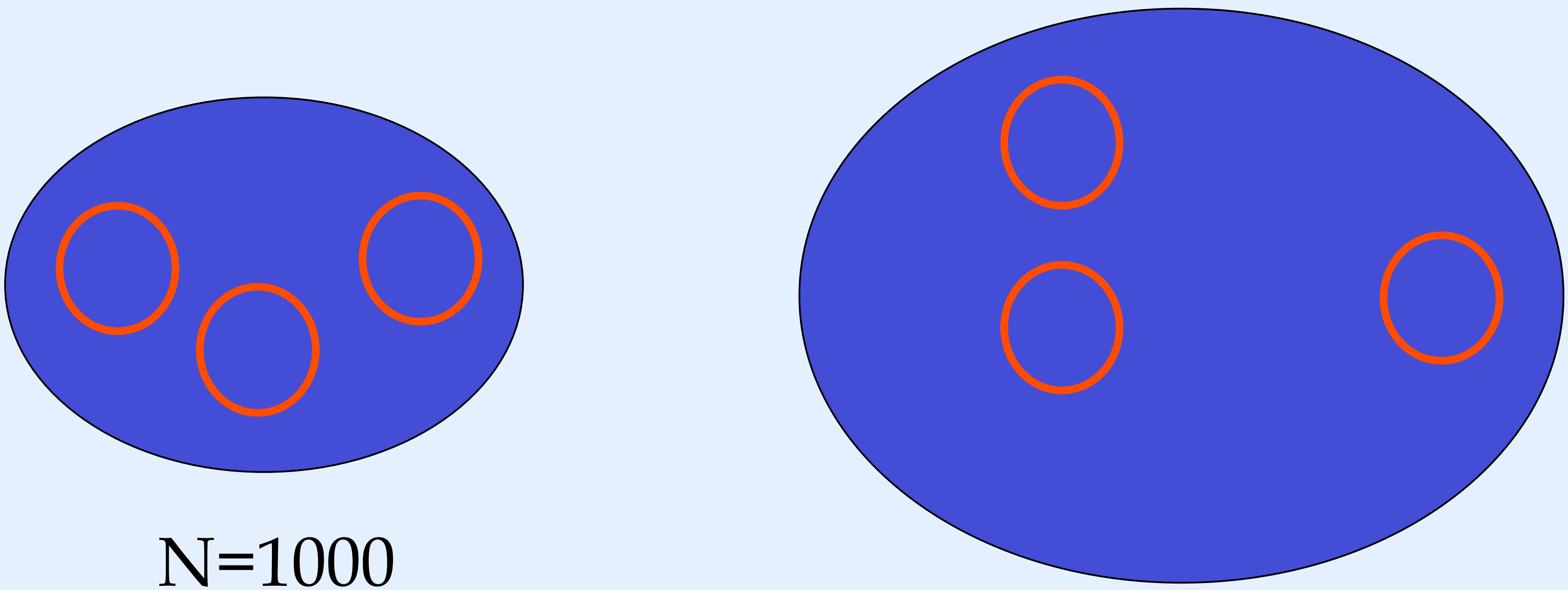
If you select 100 households at random, would the standard error of the sampling distribution of the mean number of children be larger if the population size, *N*, is 1000 or if it is 10,000?



Questions

Using the Properties of the Sampling Distribution of the Mean

If you select 100 households at random, would the standard error of the sampling distribution of the mean number of children be larger if the population size, *N*, is 1000 or if it is 10,000?



N=1000

N=10000

Homework

Page 337 P7, E15, E16, E17, E18, E19, E20, E22, E27, E28