

# Stat 411/511

## MORE ON THE RANDOM SAMPLING MODEL

Sep 29 2015

# Announcements

My office hours:

Mondays 11am 255 Weniger

Thursdays 3-5pm 3003 Cordley

# Help with Statistics Classes

## Kidder M111, *Fall Term 2015*

Students in ST 201, 314, 351, 352, and 411/511 may obtain help during any of the hours listed below (starting the second week). The class listed in parentheses below the Teaching Assistant's name has priority during that period.

Time	Mon	Tues	Wed	Thurs	Fri
0800-0900 (8:00-9:00)	Dane Skinner (ST 352)		Shaoshan Liao (ST 411/511)		
0900-1000 (9:00-10:00)	Dane Skinner (ST 352)	Chunxiao Wang (ST 351)	Shaoshan Liao (ST 411/511)	Joe Maurer (ST 351)	Chunxiao Wang (ST 351)
1000-1100 (10:00-11:00)		Caley Johns (ST 351)		Sam Engle (ST 351)	
1100-1200 (11:00-12:00)	Sam Engle (ST 351)	Caley Johns (ST 351)	Casey Bausell (ST 351)	Si Liu (ST 351)	
1200-1300 (12:00-1:00)	Spencer LeDoux (ST 411/511)	Matt Higham (ST 351)	Casey Bausell (ST 351)		Casey Stevens (ST 351)
1300-1400 (1:00-2:00)	Chris Comiskey (ST 411/511)	Ben Brintz (ST 411/511)	Laura Gamble (ST 201)		Dan Garmat (ST 351)
1400-1500 (2:00-3:00)	Trevor Ruiz (ST 201)	Chris Comiskey (ST 411/511)	Laura Gamble (ST 201)		
1500-1600 (3:00-4:00)	Trevor Ruiz (ST 201)	Peter Rise (ST 351)	Yiran Wang (ST 352)	Meng Mei (ST 314)	
1600-1700 (4:00-5:00)		Peter Rise (ST 351)	Yiran Wang (ST 352)	Meng Mei (ST 314)	
1700-1800 (5:00-6:00)	Matt Higham (ST 351)		Alyssa Pedersen (ST 201)		
1800-1900 (6:00-7:00)					

# Last time

population distribution

sample

population inference

histograms

probability distributions

sampling distributions

# Today

The sampling distribution for the sample average

Normal distribution

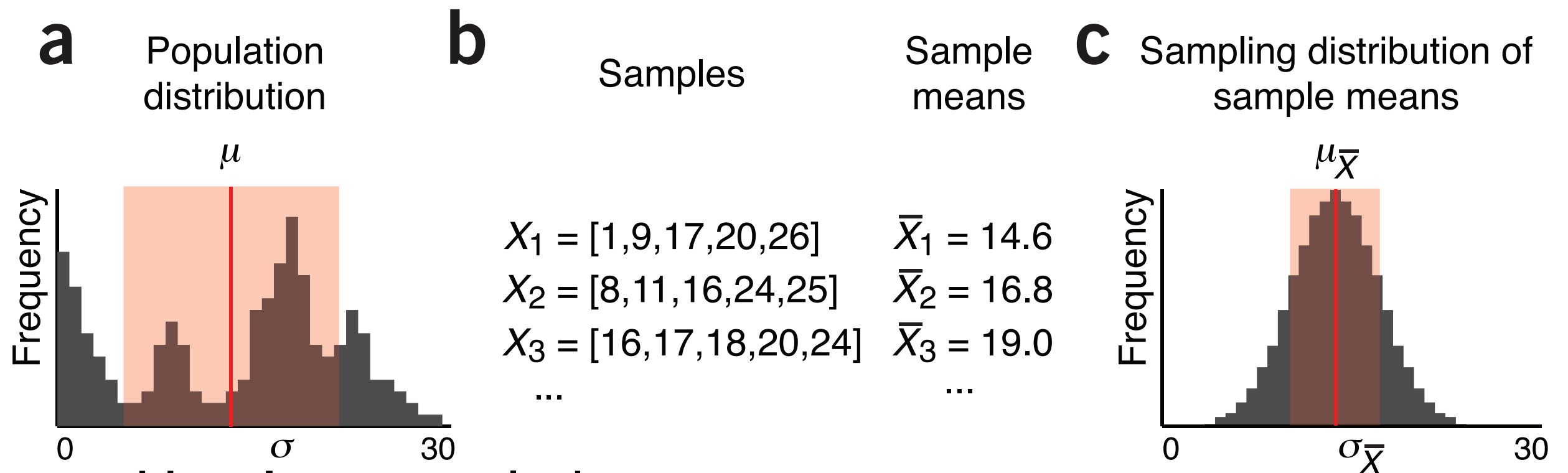
Using the sampling distribution for the sample average to construct a likely interval for the population mean

# Sampling Distribution

**Sleuth:** histogram of all values for the statistic from all possible samples that can be drawn from a population

**Nature Article:** Sample statistics have their own distribution called the sampling distribution, which is constructed by considering all possible samples of a given size.

**OpenIntro:** distribution of the point estimates based on samples of a fixed size from a certain population.

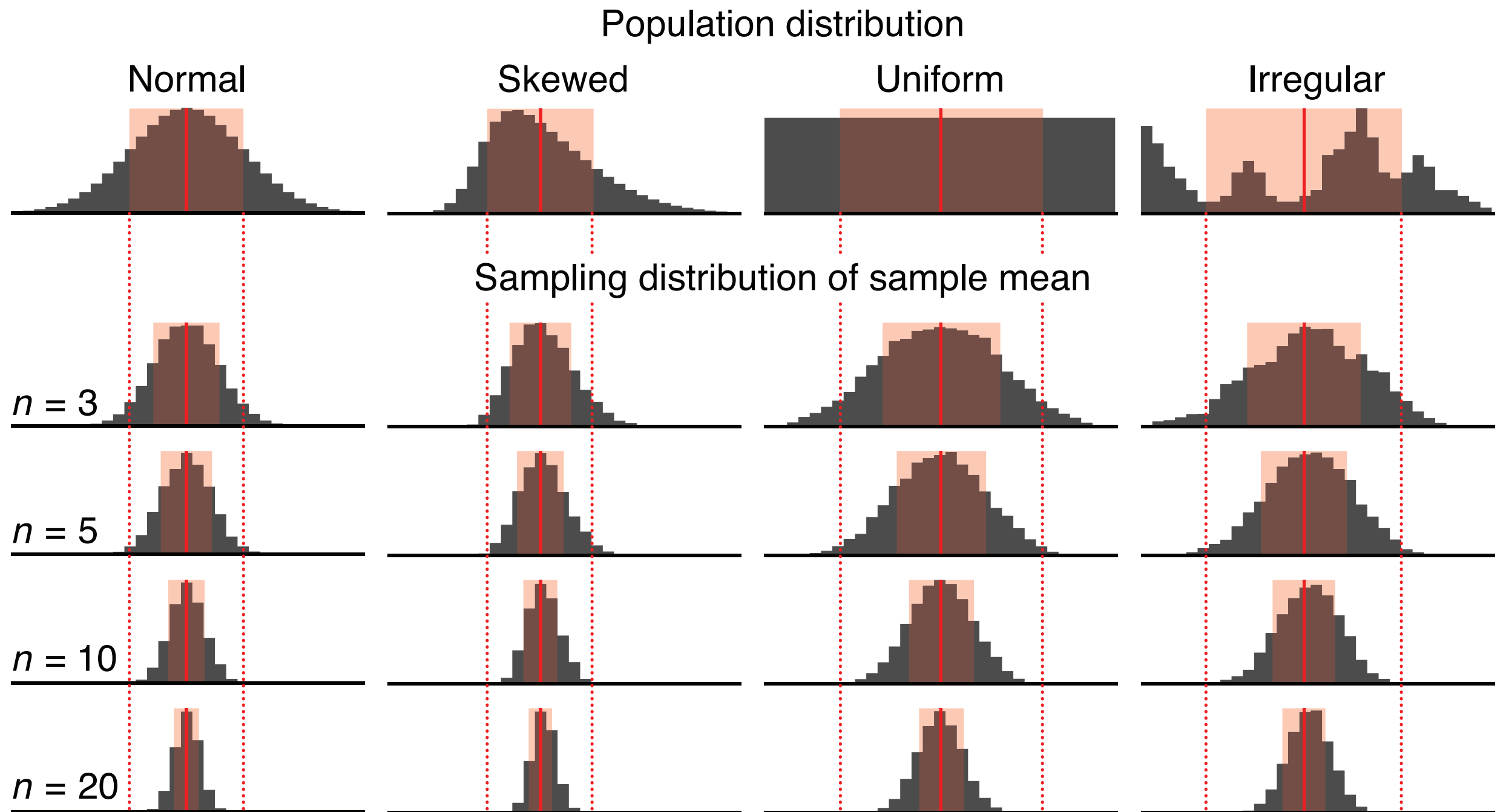


**a** Here's a population

**b** Imagine we take a sample of size  $n=5$  from this population. One example would be  $\{1, 9, 17, 20, 26\}$ , it's sample average is 14.6. But that is only one possible sample.

**c** Imagine all the other possible samples. For each sample find it's sample average and make a histogram of these sample averages. This is the sampling distribution of the sample average.

sample size



Your Turn: What was the point of this figure?



## **Facts** about the sampling distribution for the **sample average**

Regardless of the shape of the population distribution, the sampling distribution:

- 1** will have the same mean as the population distribution  $\mu_{\bar{X}} = \mu$
- 2** have a smaller standard deviation  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- 3** and it's shape will be closer to a Normal distribution than the population distribution  
(how close depends on the sample size and how close the population distribution was to Normal).

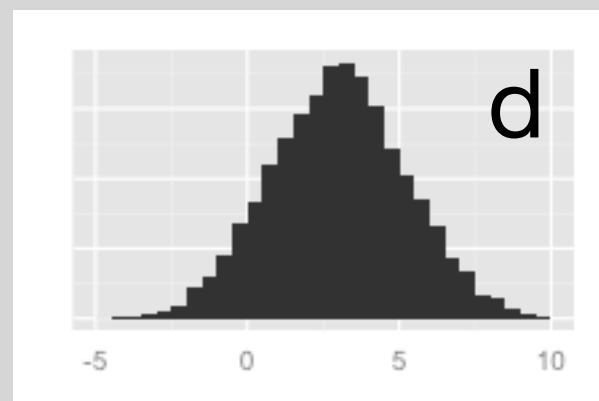
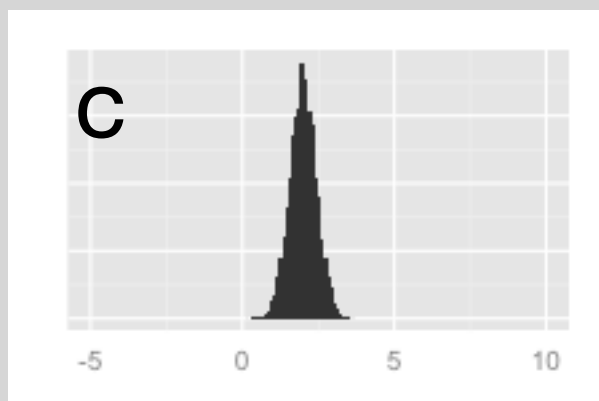
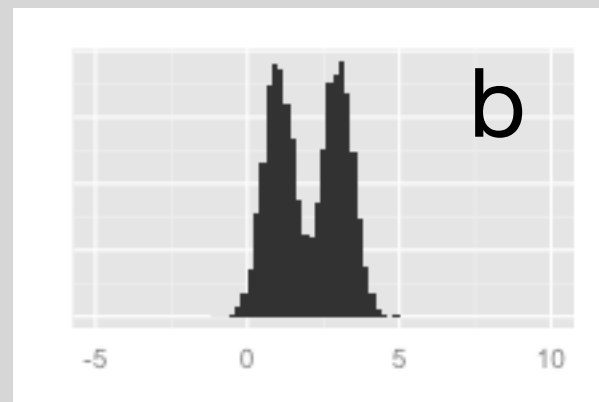
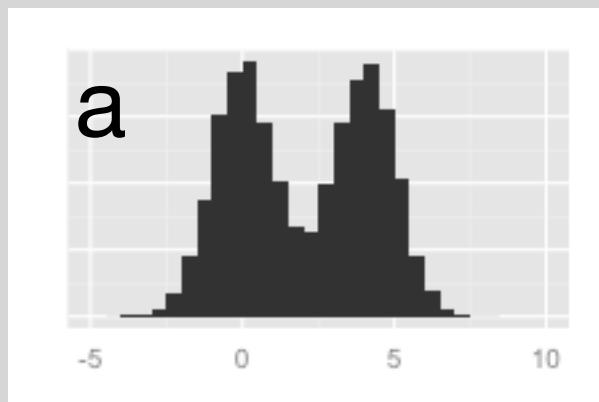
## **Central Limit Theorem**

population

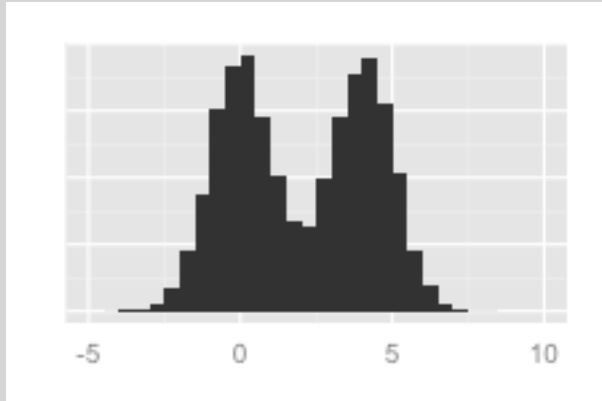
# Your turn

look carefully

Which histogram is most likely a histogram of the **sampling distribution** of the **sample average** for a sample of size 30?



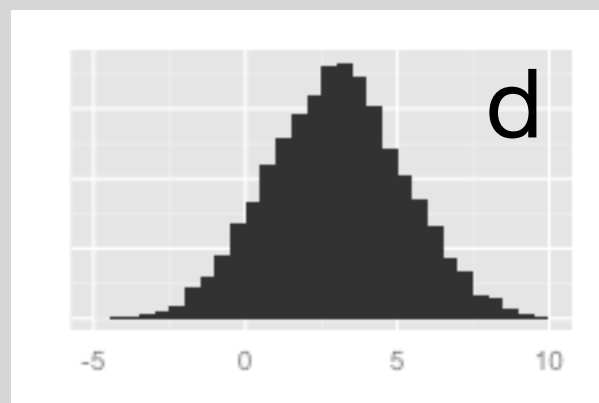
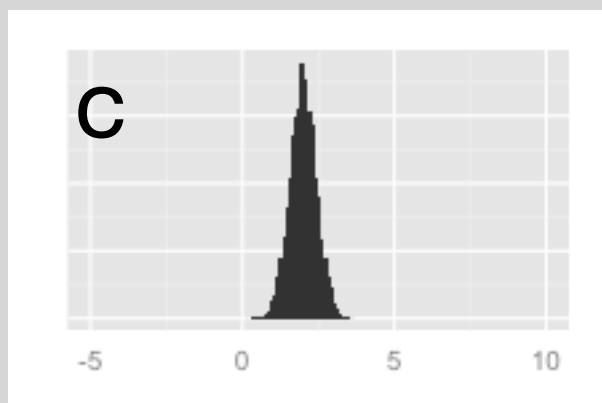
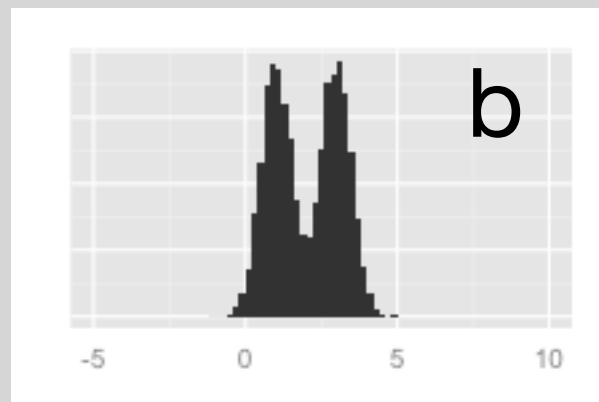
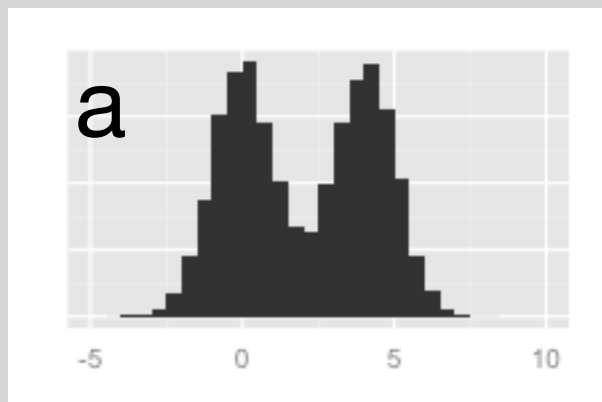
population



# Your turn

look closely

Which histogram is most likely a **histogram** of a **sample** of size 300 from the population?



The key to making inferences in the random sampling model is the **relationship** between the *population distribution* and the *sampling distribution*.

Ok, but we don't know  $\mu$ ,  $\sigma$  or the shape of the population distribution, so we don't know exactly what the sampling distribution is.

If we did, we wouldn't be asking a question about the population.

A common way to proceed is to **assume** the sampling distribution is Normal.

# The Normal distribution

A particular distribution shape.

Defined by a mathematical function.

Completely specified by its mean (center) and standard deviation (spread).

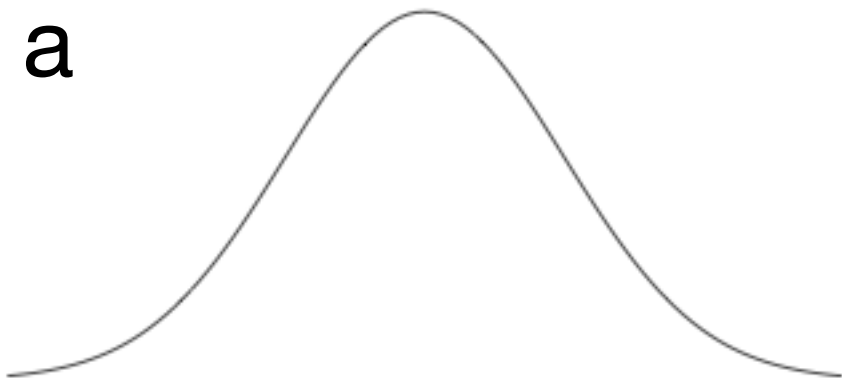
Useful approximation to many distributions, but, very few things are exactly Normal.

***68-95-99.7% rule:***

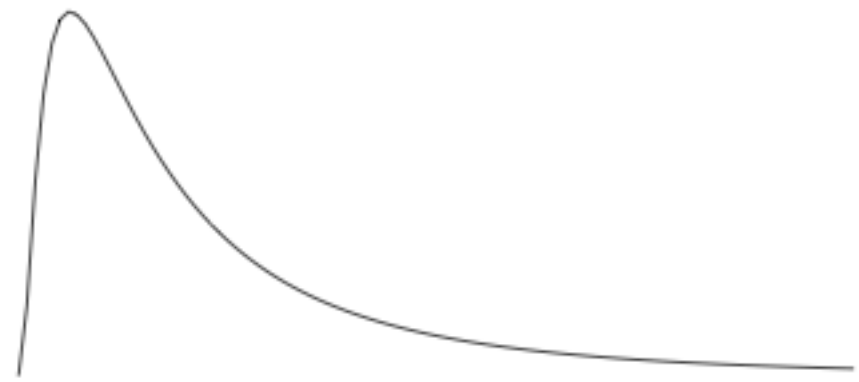
If data is Normally distributed, 68% of observations will be within 1 standard deviation of the mean, 95% within 2 sds, 99.7% within 3 sds.

# Your turn

a



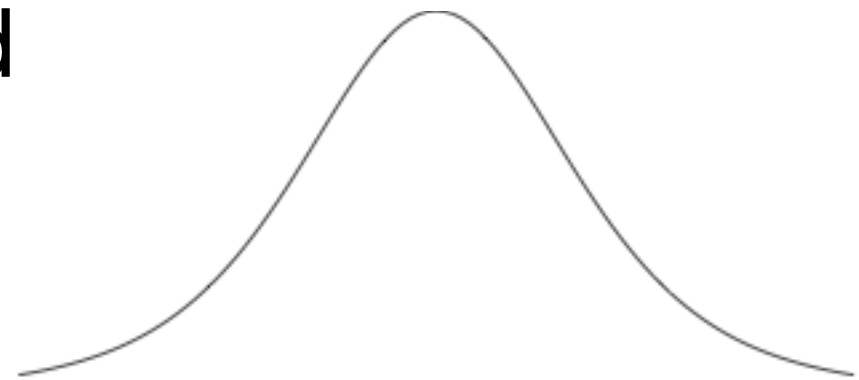
b



c



d

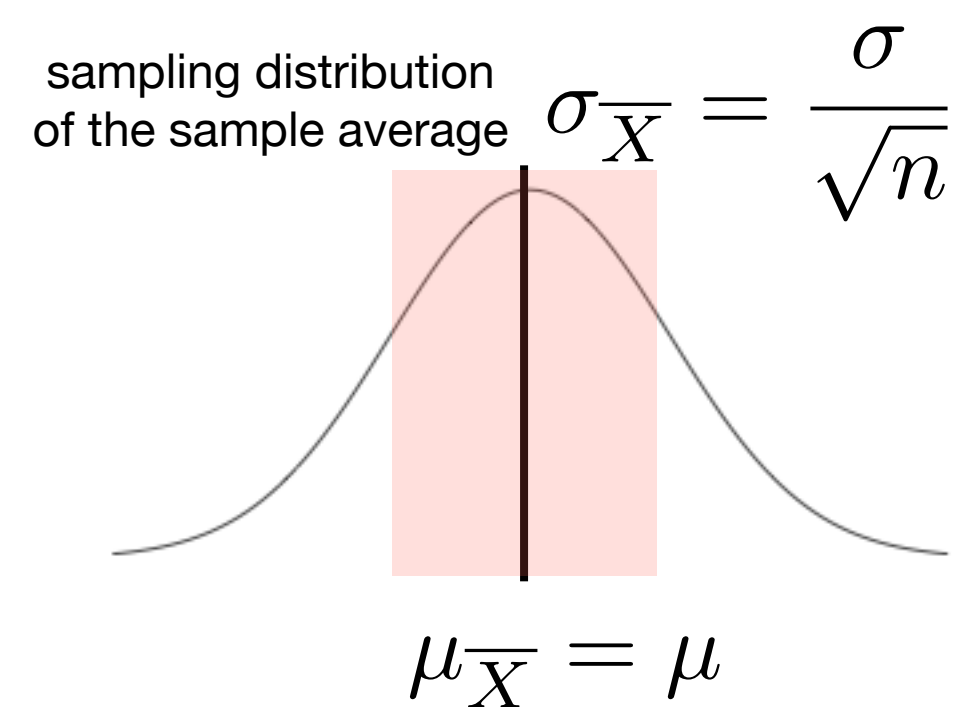


Which of these are Normal distributions?

# Next up..

Use the facts about the sampling distribution for the **sample average**, to construct a range of likely values for the **population mean**.

Let's get started by assuming the sampling distribution for the sample average **is** Normal.



We know it would be unusual to see a value two standard deviations from the mean.

How unusual?

In about 95% of possible samples, our sample average would fall **within** two standard deviations of the mean of the sampling distribution.

In about 95% of possible samples,

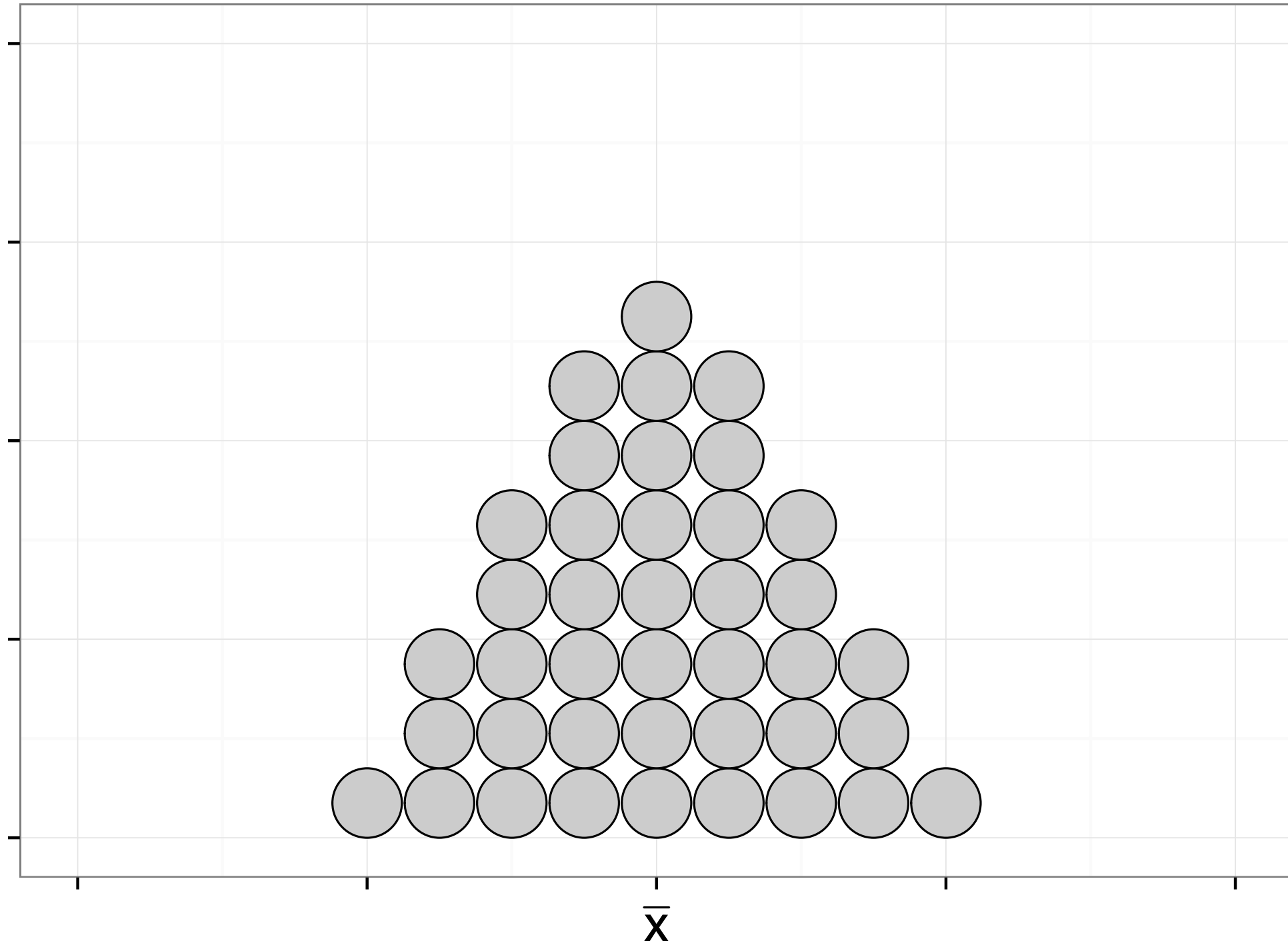
$$\bar{X} > \mu - 2\sigma / \sqrt{n} \quad \text{and} \quad \bar{X} < \mu + 2\sigma / \sqrt{n}$$

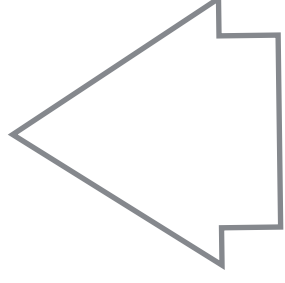
Rearrange: in about 95% of possible samples,

$$\mu > \bar{X} - 2\sigma / \sqrt{n} \quad \text{and} \quad \mu < \bar{X} + 2\sigma / \sqrt{n}$$

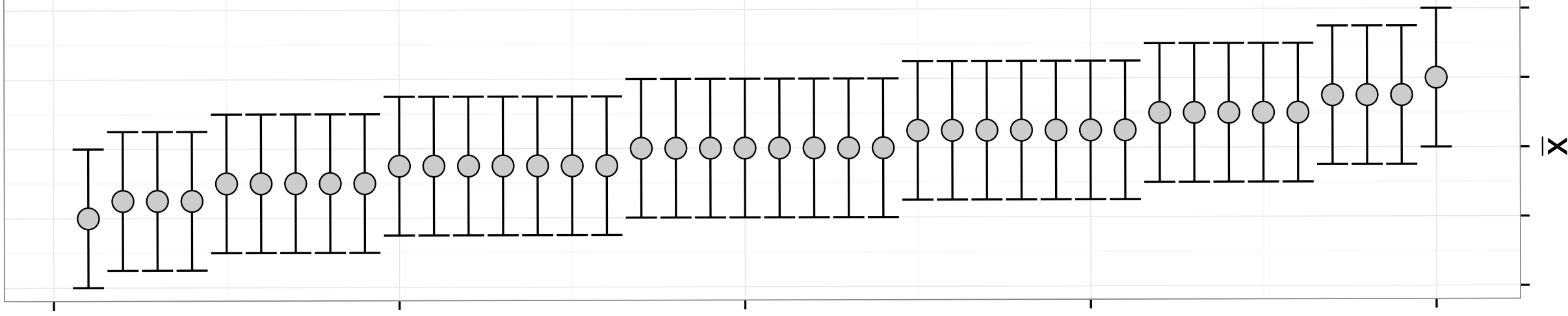


In a picture...





this way up



## Z-based confidence interval

With 95% confidence, the population mean is between  $\bar{X} - 2\sigma/\sqrt{n}$  and  $\bar{X} + 2\sigma/\sqrt{n}$

a 95% confidence interval

in 95% of possible samples, this interval covers the true population mean.  
**correct**

there is 95% chance the population mean is in this interval.

**incorrect**

We still don't know  $\sigma$ , but we could estimate it with the sample standard deviation,  $s$ .

Of course, this all hinges on our **assumption** that “the sampling distribution is Normal” is true.

# Recap

Population inference is using a sample to learn about a population.

This process relies on knowing how the sampling distribution of our statistic relates to the population distribution and our parameters of interest.

If we are interested in the population mean, assuming the sampling distribution of the sample average is Normal, leads us to a 95% confidence interval for the mean of the population,

$$\bar{X} \pm 2 \frac{\sigma}{\sqrt{n}} \quad \text{know } \sigma, \text{ one sample Z-based CI}$$

# Standard deviation of the mean

sample mean  
sample average  $\bar{X}$

The standard deviation of the sampling distribution of the sample average for a sample of size  $n$ , is the population standard deviation divided by the square root of the sample size.

$$SD_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This tells us how much the sample average varies from its mean across different possible samples.

But we usually don't know  $\sigma$

Often we estimate the population standard deviation with the sample standard deviation.

I.e. we estimate  $\sigma$  with  $s$ .

# Standard error of the mean

sample mean  
sample average  $\bar{X}$

If we plug in  $s$  for  $\sigma$  in the standard deviation of the sampling average, we called it the standard error.

The **standard error** of the sample average is an **estimate** of the **standard deviation of the sampling distribution** of the sample average.

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

It's an **estimate** of how much the sample average varies from it's mean across different samples.

# Next time...

...assuming the sampling distribution of the sample average is **Normal**...

**When is this true?**

**What is the effect of using  $s$ , instead of  $\sigma$ ?**

**The t-based confidence interval**