Stat 411/511

Inference in the Sampling Model

Sep 28 2015

Charlotte Wickham

stat511.cwick.co.nz

Announcements

Today

A recap of statistical inference in the random sampling model

Wednesday: paired t-test

Your turn

What terms did you highlight?

An example of the statistical process

in the random sampling model

We have a question.

Do textbooks cost more at the bookstore than on Amazon?

We translate this to a question about a **population** distribution.

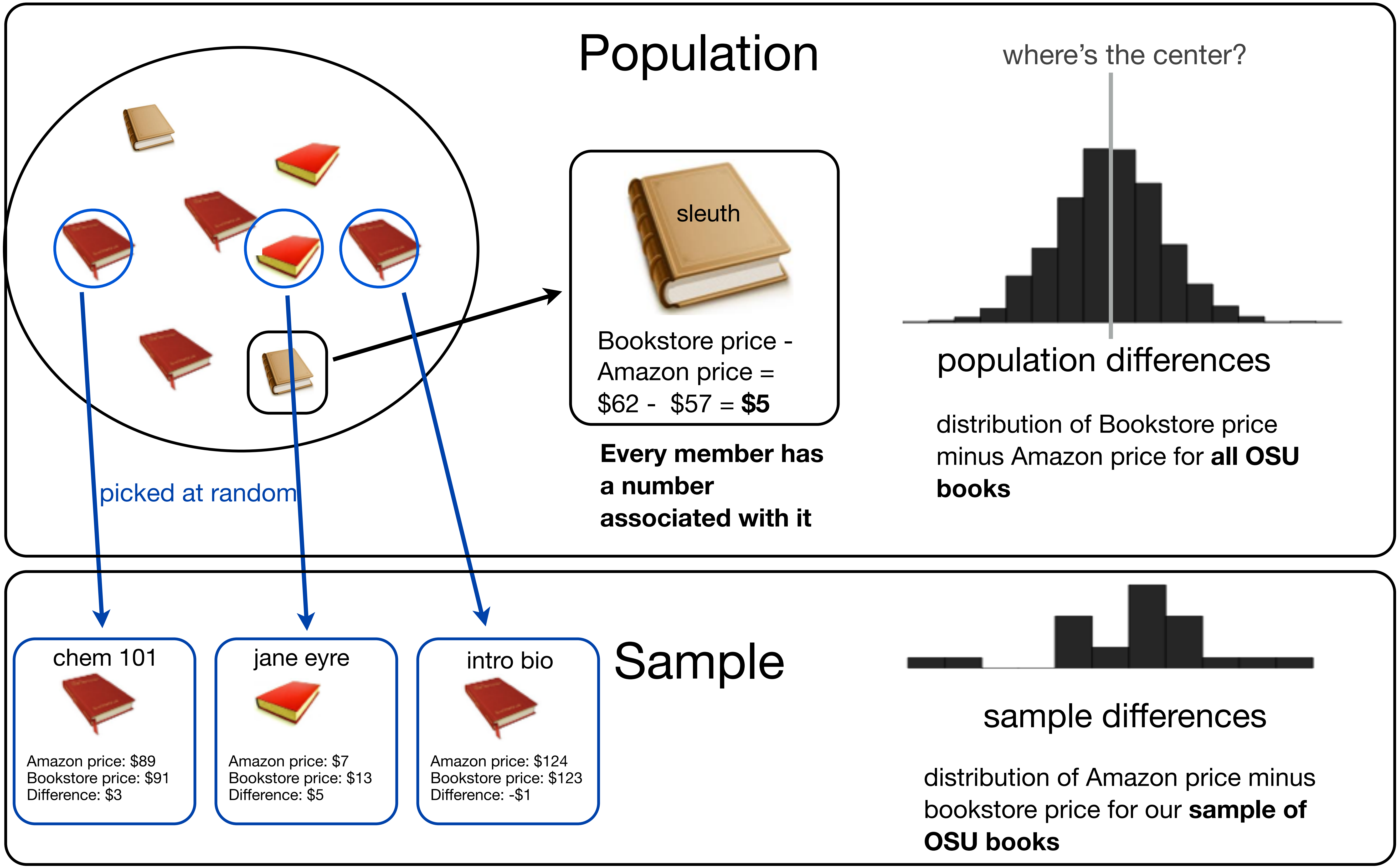
What is the mean difference in price between Amazon and the bookstore, for **all textbooks required in OSU classes**? Is this mean bigger than zero?

We can't/won't/don't collect data on the whole population, but instead get a sample from the population and use properties of the sample to estimate properties of the population.

The average price difference in our sample of size, n=100, is $10 with a sample standard deviation of $5. With 95% confidence we estimate that OSU textbooks on Amazon cost between $9 and $11 less than at the bookstore.

Random sampling model

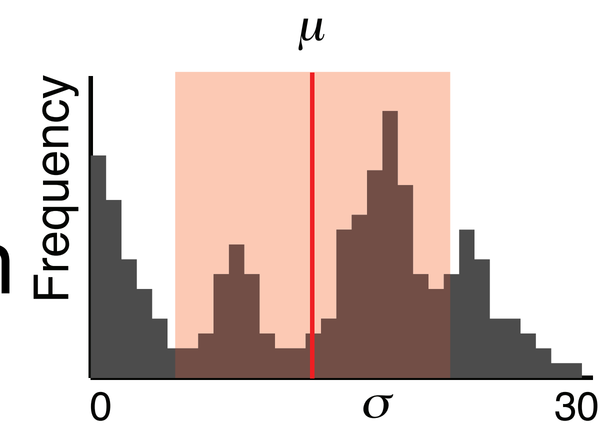
Single population



Histograms and distribution functions

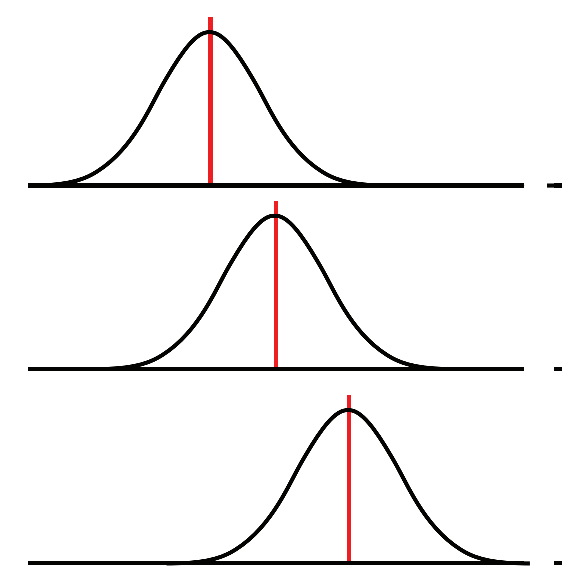
A histogram is a graphical representation of the distribution of a **finite** set of numbers.

To find the number of observations that were in a given range we add the heights of the bars.

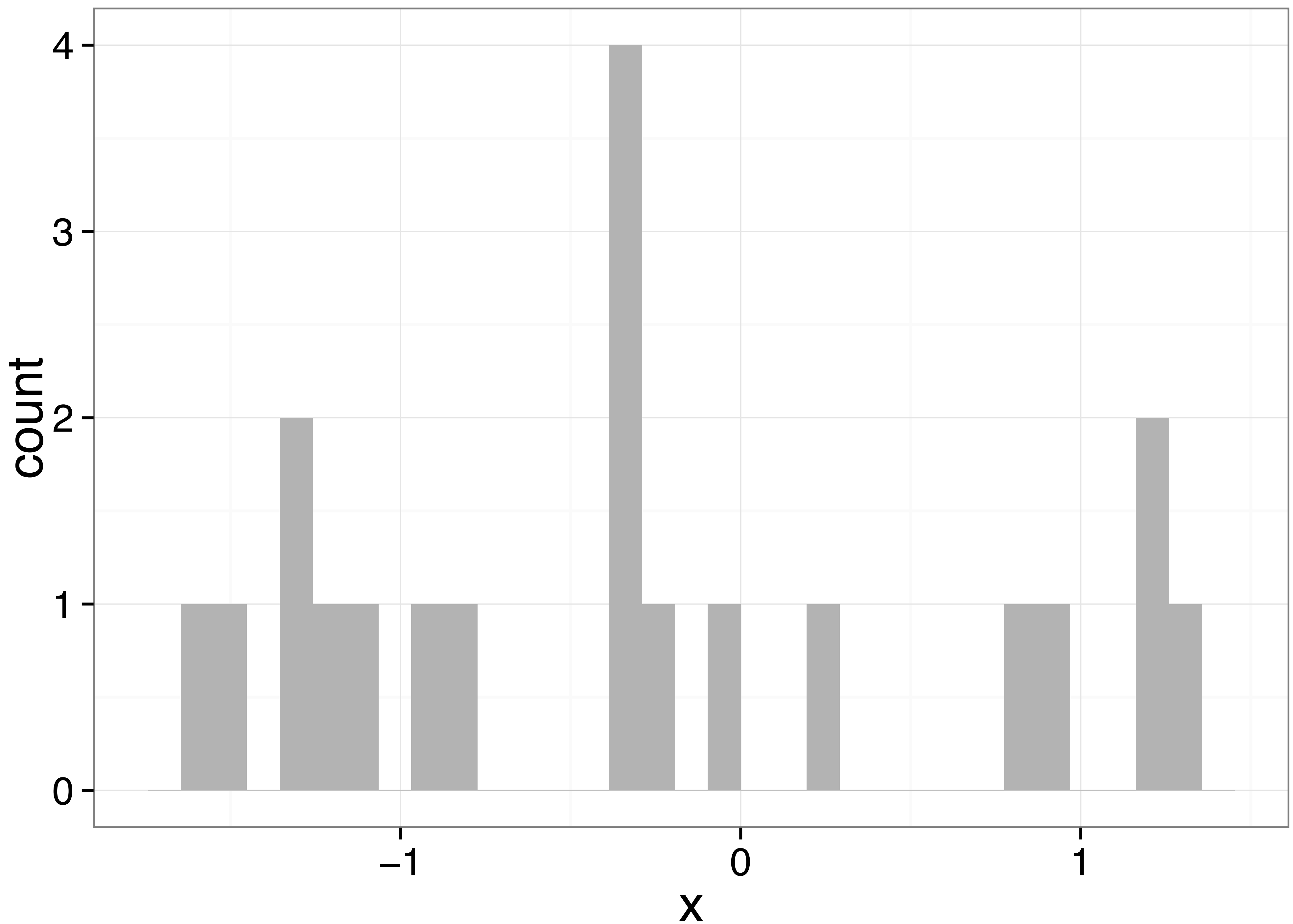


Sometimes you'll see a smooth curve as a representation of a population distribution. Think of it like a histogram with infinitely small bin width, and infinitely many observations.

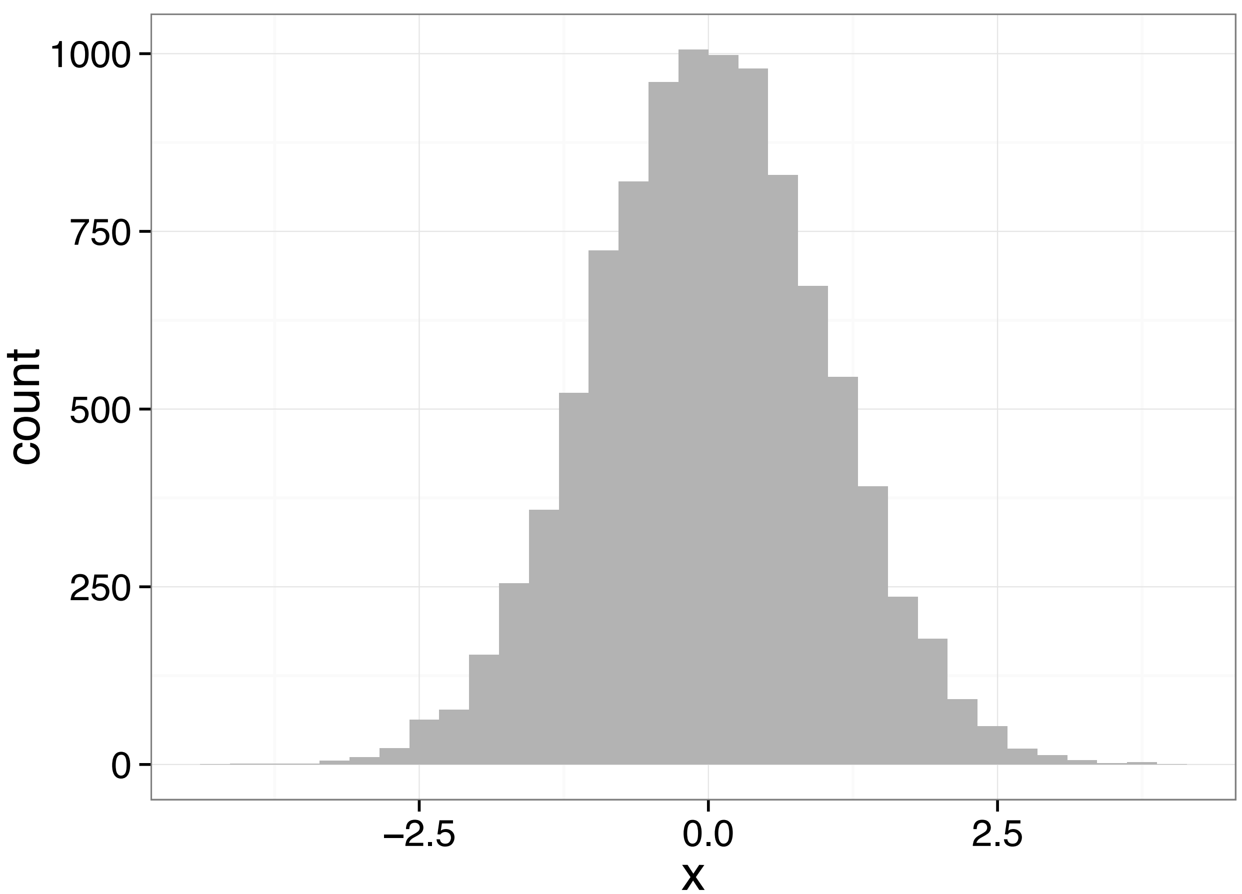
Areas under the curve represent probabilities.



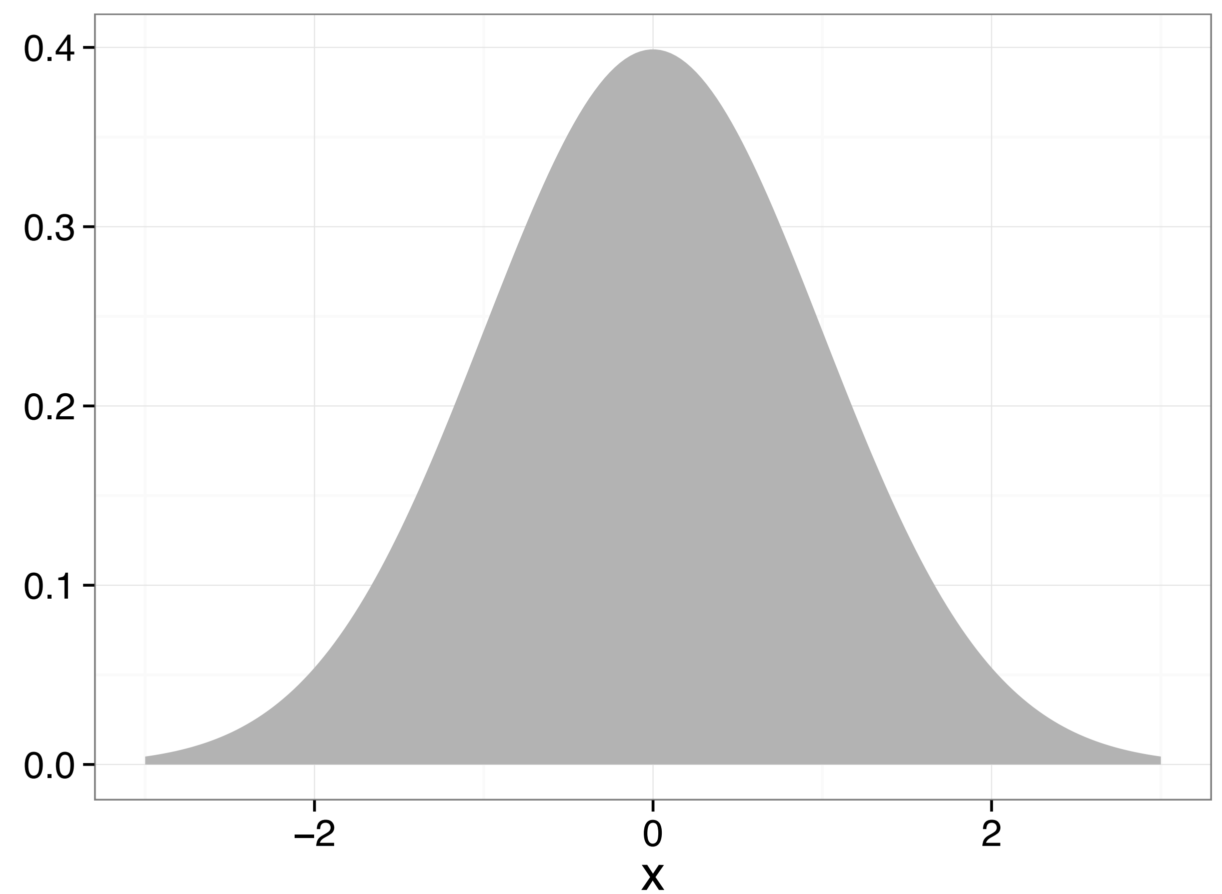
technically, this is called a probability density function, ST521



count
4
3
2
1
0
−1
0
1
x



1000
750
500
250
0
count
−2.5
0.0
2.5
x



0.4
0.3
0.2
0.1
0.0
−2
0
2
x

Statistical Inference

Population inference is using a sample to infer properties of the population.

For the textbooks: using the sample average to infer the population mean

This is **statistically justified** as long as:

observations are **sampled at random** from the **population of interest**.

Chance enters the study through the act of randomly taking a sample.

This is one of two “mechanisms of chance” we will cover.

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

What is a *sampling distribution*?

Parameters, Statistics and Estimates

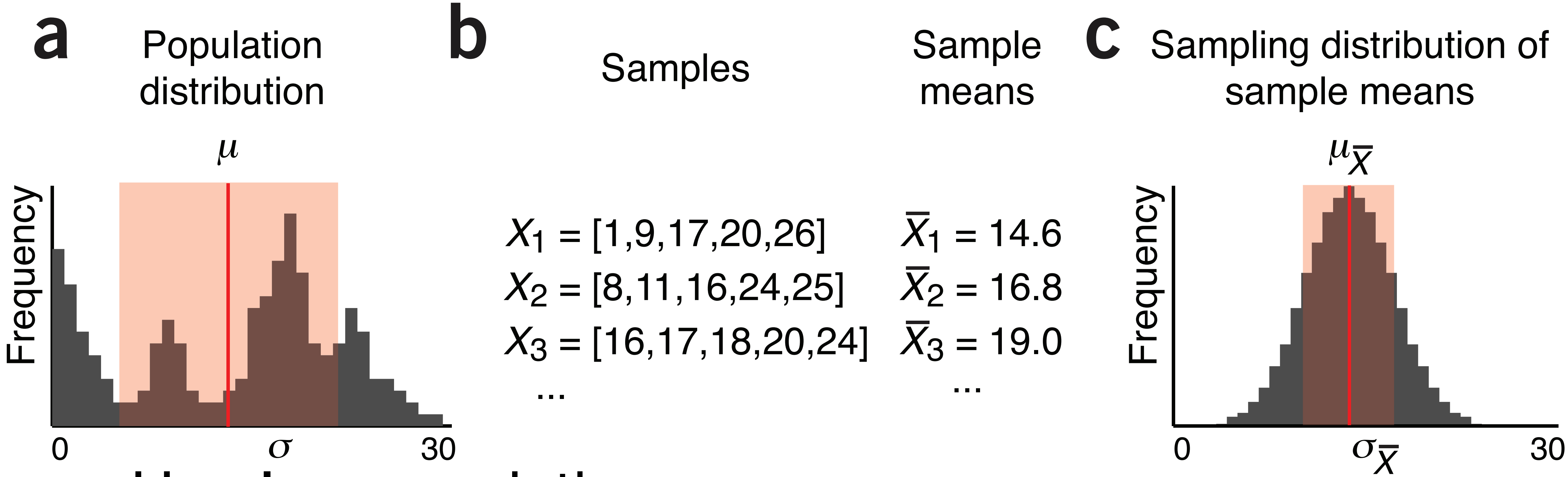
Name	Definition	Examples
Parameter (of a population, or of a model)	An **unknown** value in a probability model	Population mean, μ (mu) Population standard deviation, σ (sigma)
Statistic (of a dataset)	Something you can calculate from data	Sample average, $\overline{Y}$ (y bar) or $\overline{X}$ Sample standard deviation, SD or s
Estimate (of a parameter)	A statistic used as a guess for a parameter	The sample average is an estimate of 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 parameters have their own distribution called the sampling distribution, which is constructed by considering all possible samples of a given size.

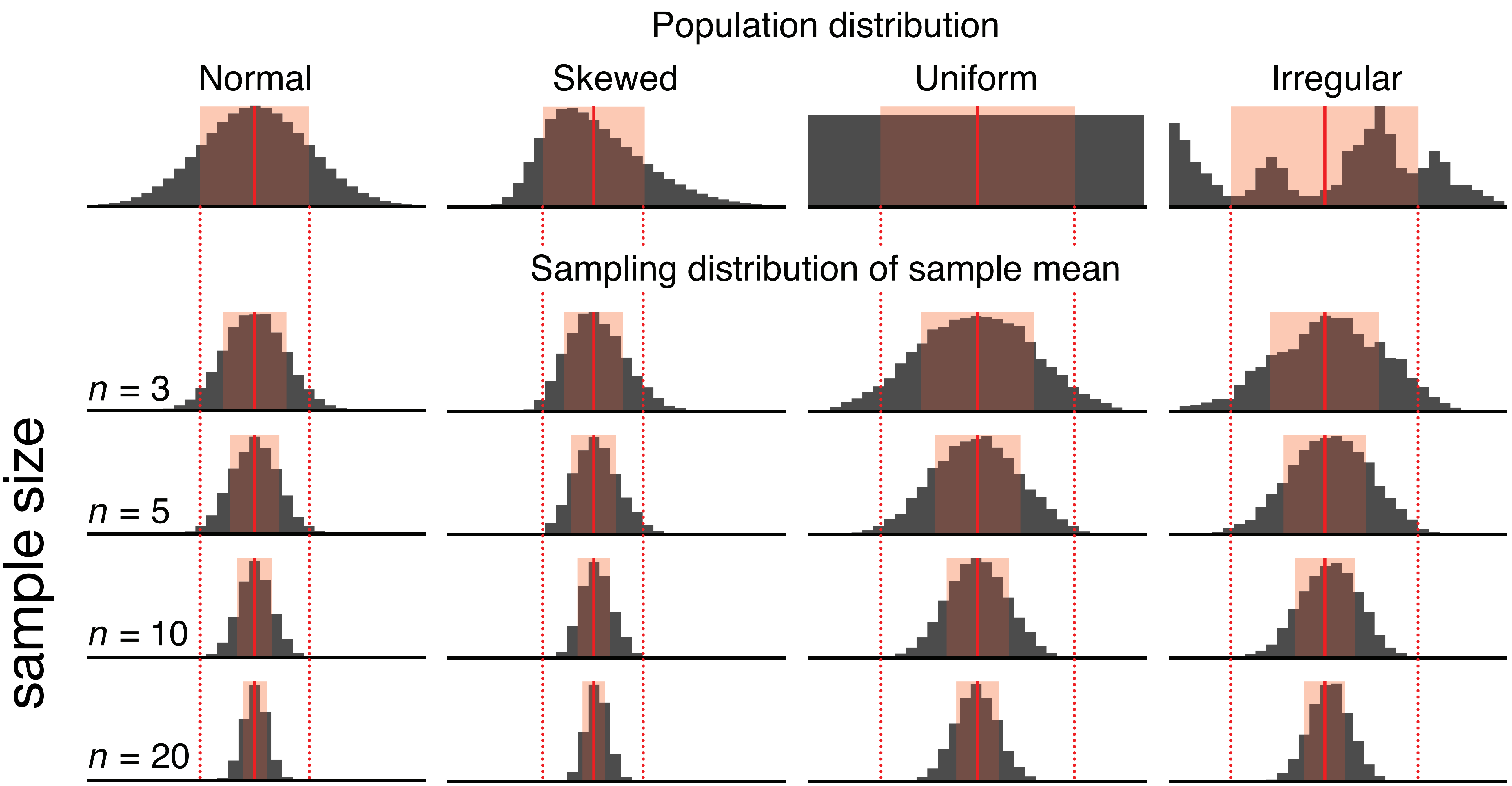
OpenIntro: distribution of the point estimates based on samples of a fixed size from a certain population.



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.


Population distribution
Normal
Skewed
Uniform
Irregular
Sampling distribution of sample mean
sample size
n = 3
n = 5
n = 10
n = 20

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 $\mu_{\overline{X}} = \mu$

2 have a smaller standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

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

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 μ, σ 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.

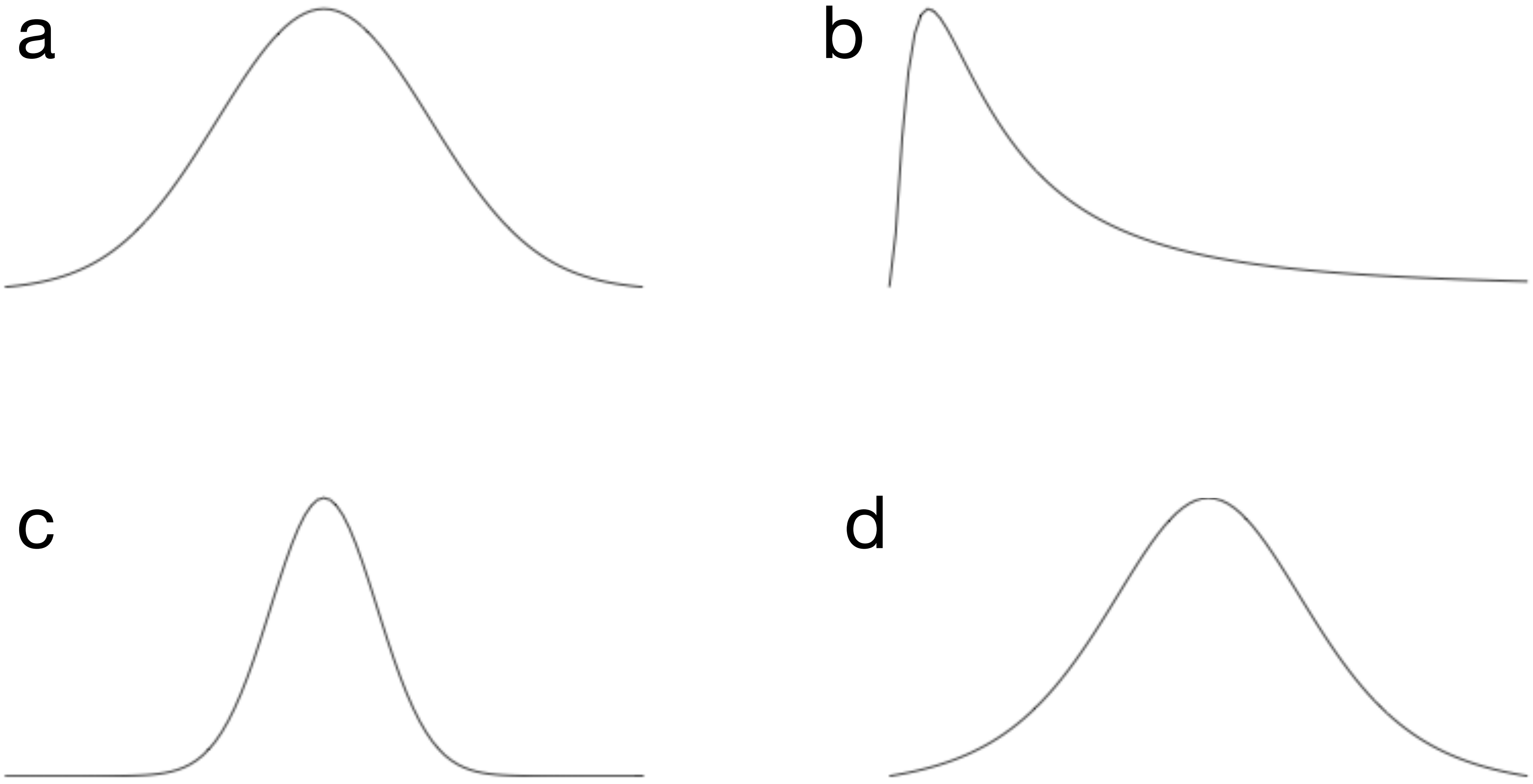
Completely specified by it's mean (center) and standard deviation (spread).

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



Which of these are Normal distributions?

Next time

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

Did today's material feel foreign?

Read Chapter 4 in OpenIntro:

http://www.openintro.org/stat/down/OpenIntroStatSecond.pdf