

# Math 140

## Introductory Statistics

# Let's make our own sampling!

If we use a random sample  
(a survey)

or if we randomly assign treatments to subjects  
(an experiment)

we can come up with proper, unbiased conclusions

We should work with randomized data  
to avoid bias

But HOW to produce, collect and analyze data?

# 7.1 Generating sampling distributions

Generate sampling distributions and study:

The sample mean

The sample shape, the center, the spread

How to draw proper conclusions about what is likely and what is rare.

How to relate this to the entire population?

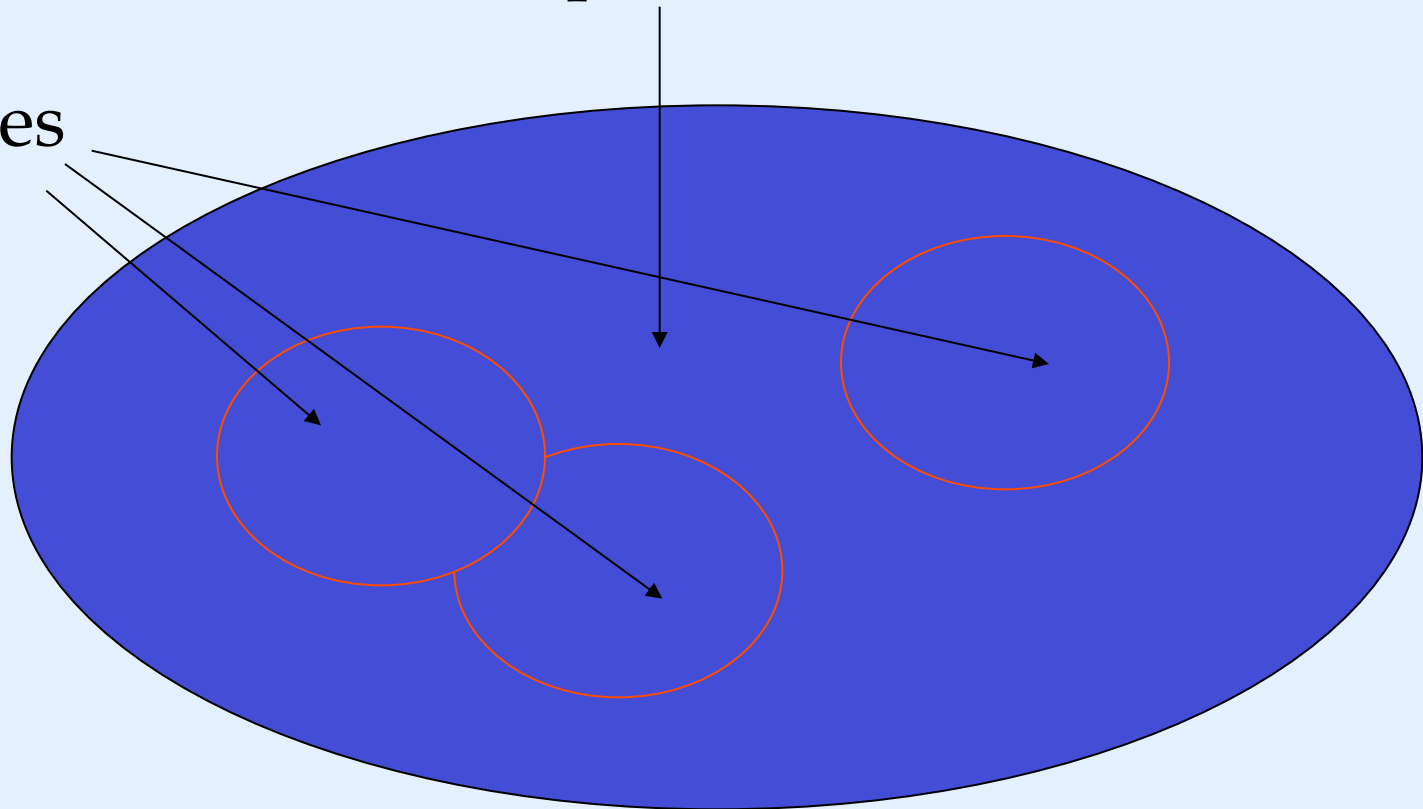
How to do this in the easiest way?

# Sample vs. Population

Sampling size  $n$

Population

Samples



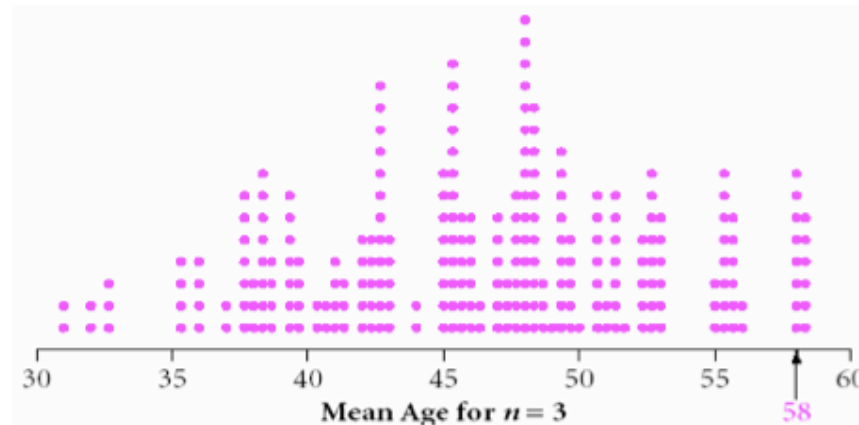
# Our friends at Westvaco

Recall, the people laid off were 55, 55 and 64  
is this discrimination or not?

We need to compare with **RANDOM**  
layoffs of three people

Randomly select three workers from the group of 10  
with ages above, and calculate the mean age of the  
three selected.

25 33 35 38 48 55 55 55 56 64

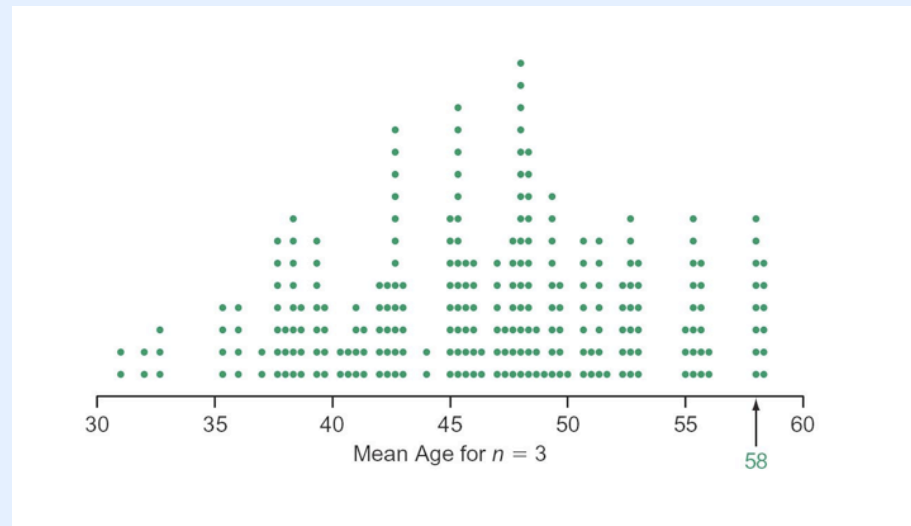


# Perform a random ‘simulation’

All possible sets of 3 people chosen from 10

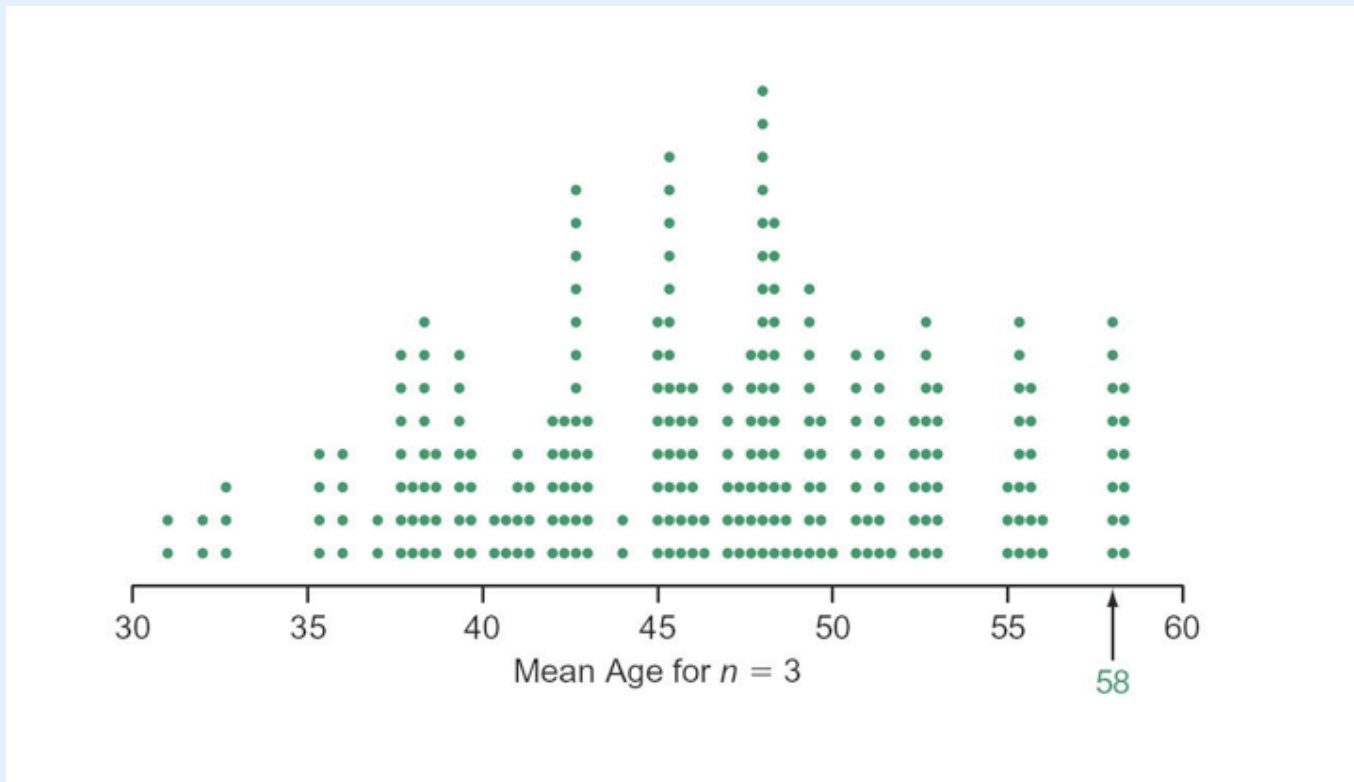
$$\binom{10}{3} = 120$$

For each of these groups calculate average age and create a dot plot - this is your **sampling distribution**



# Conclusions from Westvaco

The average age of the people that were actually laid off was 58



Common sense: It is rather hard for this to happen by chance - Westvaco has some explaining to do

# Generate a sample distribution

## Simulated sampling distribution:

distribution of summary statistics obtained from taking repeated random samples.

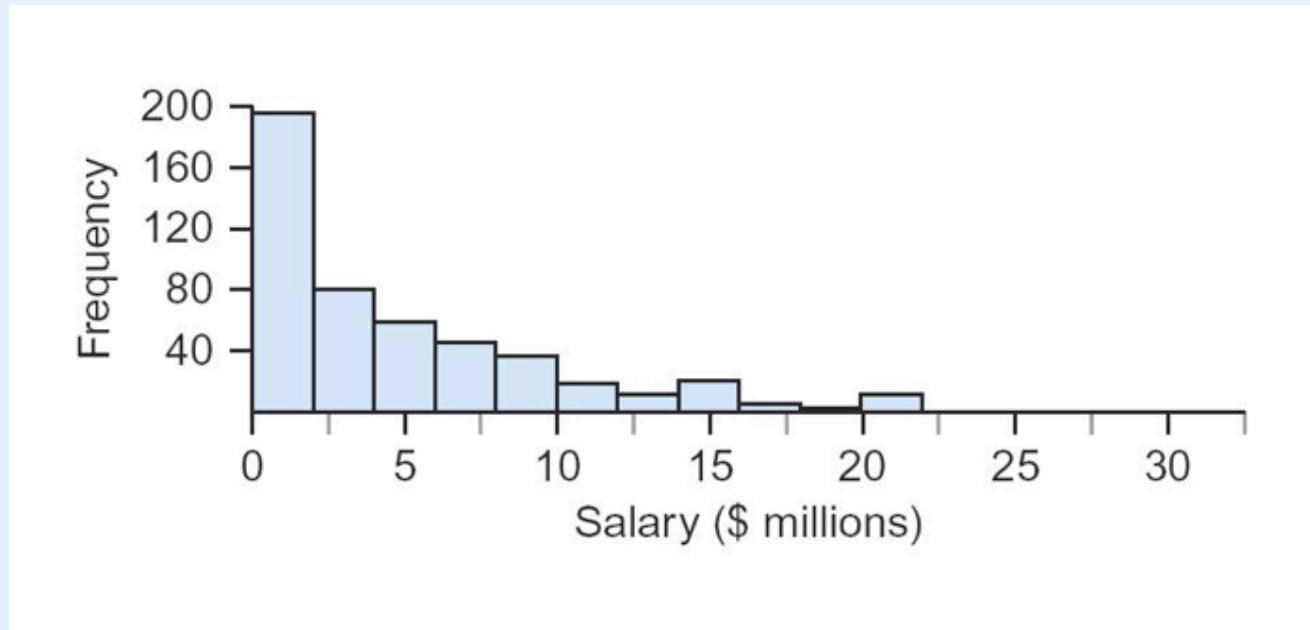
- I. Take a random sample of a fixed size  $n$  from a population.
- II. Compute a summary statistic for this sample.
- III. Repeat steps I and II many times.
- IV. Display the distribution of the summary statistic.



We will often have access to samples,  
but not necessarily to the entire  
population

(too big or inaccessible)

# Our friends at the NBA



These are the salaries of NBA players.  
The mean is \$4.6 million and the SD is \$4.7 million.  
Highly skewed

**THESE ARE POPULATION STATISTICS  
(EVERYBODY)**

# Our friends at the NBA

Suppose this data was not public and I am an NBA player who wants to know the average salary of my colleagues.

I can only access 10 people at random.

How is the average I find different from the true average?

Since the distribution is skewed, should I be concerned?

# Lets' 'simulate' a sampling distribution

Select random samples of 10 from our distribution

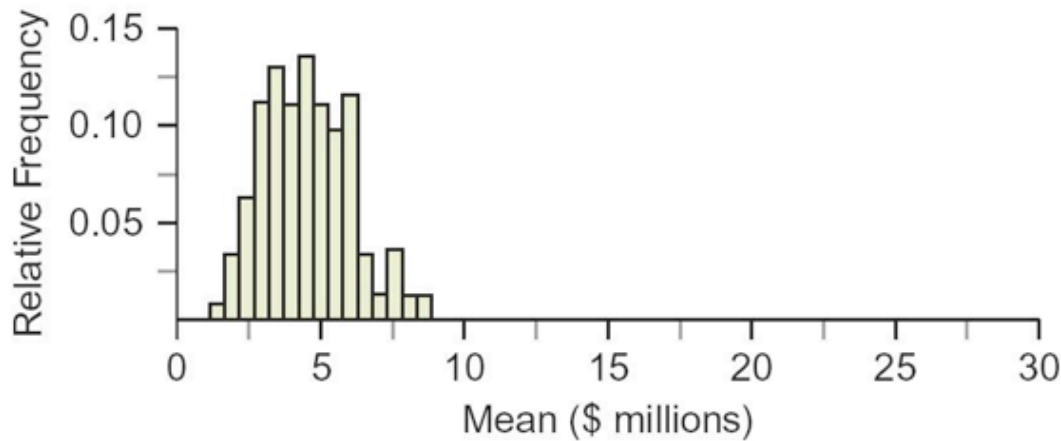
Calculate average salary

Repeat many times (200?)

Place them in a chart

**THESE ARE SAMPLE STATISTICS**

# Average simulated salaries

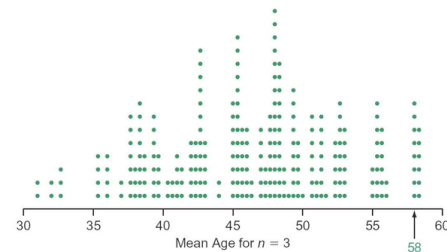


200 simulations  
The distribution is  
approximately  
Normal

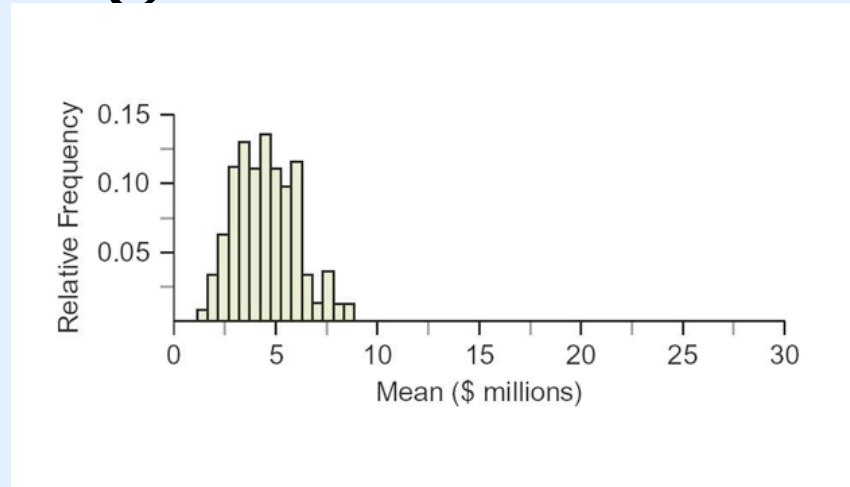
Centered at about  
\$4.6 million

SD is about  
\$1.5 million

Equivalent of what  
we did for  
Westvaco!



# Average simulated salaries



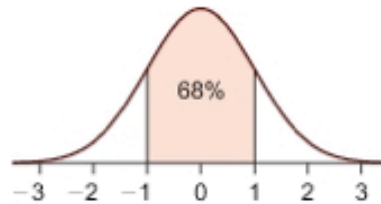
From our 200 simulations  
The distribution is approximately normal and  
centered at about **\$4.6 million**,  
the SD is about **\$1.5 million**

The mean of the entire population  
was **\$4.6 million** and the SD was **\$4.7 million**.

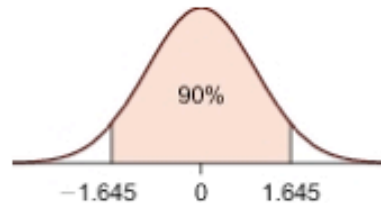
# Recall properties of the normal distribution

## Central Intervals for Normal

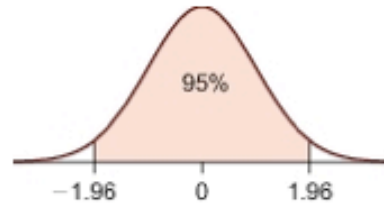
Distributions 68% of the values lie within 1 standard deviation of the mean.



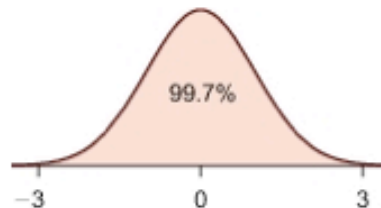
90% of the values lie within 1.645 standard deviations of the mean.



95% of the values lie within 1.96 (or about 2) standard deviations of the mean.



99.7% (or almost all) of the values lie within 3 standard deviations of the mean.

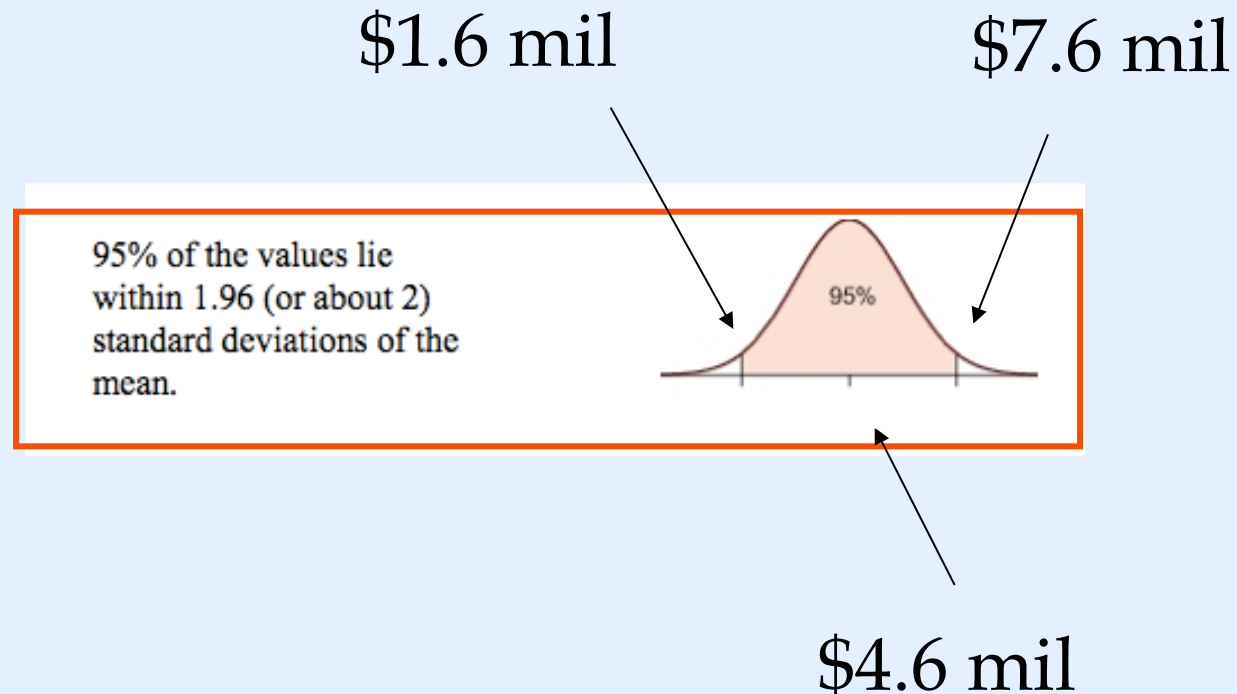


For us the mean is \$4.6 million and the SD is \$1.5 million

We can be 95% sure that our sample mean is within 3 million from the population mean

# Normal distribution

We can be 95% sure ANY  
mean of 10 people we pick falls between  
\$1.6 and \$7.6 million and centered about \$4.6 million





# Average simulated salaries

We can be pretty confident that the selection of 10 people will give us a good idea about the average salary of NBA players

We did not need to sample the entire population!

The SD from our SAMPLING DISTRIBUTION is \$1.5 million.

The SD from our POPULATION DISTRIBUTION Is \$4.7 million

# Average simulated salaries

The SD from our SAMPLING DISTRIBUTION  
is \$1.5 million.

This is called the **STANDARD ERROR**

The SD from our POPULATION DISTRIBUTION  
Is \$4.7 million

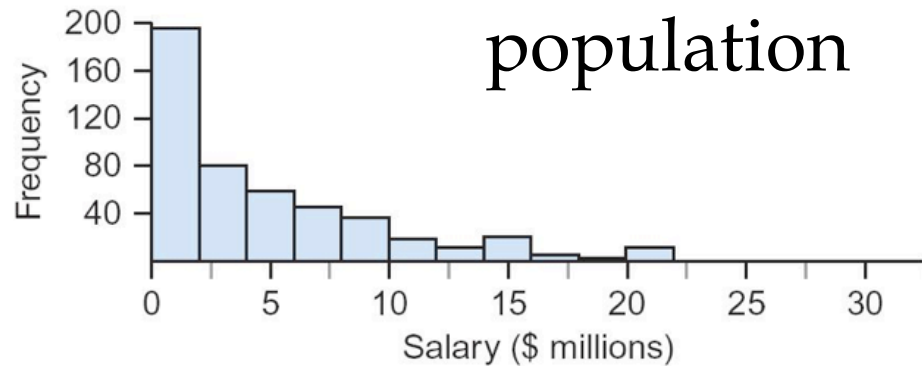
This is called the **POPULATION STANDARD  
DEVIATION**

# Definitions

Values that lie in the middle 95% of a sampling distribution are called **REASONABLY LIKELY EVENTS**

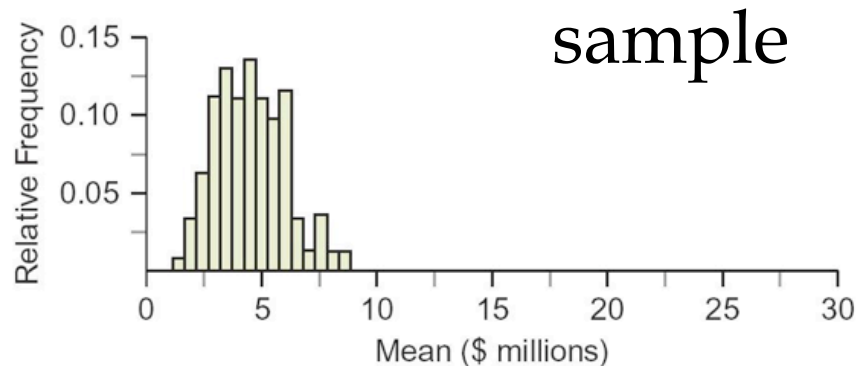
Values that lie in the left 2.5% and in the right 2.5% Sides of a sampling distribution are called **RARE EVENTS**

# Let's compare



Would we be surprised to draw a player with an \$3 million salary?

What about \$8 million salary?



Would be surprised to draw 10 players with an average salary of \$8 million?

# 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

Create the sampling distribution for the total number of square miles in any 2 parks.  
Use all possible samples of 2 parks.

# 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
----------	-----

Sample of Two Parks	Total Area (sq mi)
---------------------	--------------------

A and B	175
---------	-----

# 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

How many possible ways of selecting 2 parks?

We can only survey 600 square miles a year.

What is the probability that we DO NOT finish the survey within the first year?

# Utah's national parks

We can use all possible combinations

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

Sample of Two Parks	Total Area (sq mi)
---------------------	--------------------

A and B	175
A and C	646
A and R	497
A and Z	348
B and C	583
B and R	434
B and Z	285
C and R	905
C and Z	756
R and Z	607





# Utah's national parks

Probability we don't finish survey is  $4/10$

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

Sample of Two Parks	Total Area (sq mi)
---------------------	--------------------

A and B	175
A and C	646
A and R	497
A and Z	348
B and C	583
B and R	434
B and Z	285
C and R	905
C and Z	756
R and Z	607



# Sample and population means

Any sample mean  $\bar{x}$

Population mean  $\mu$

Usually they are different, but  
**OVER MANY SAMPLES**  
they tend to be the same

Also,  
**THE LARGER THE SAMPLE SIZE**  
the closer they will be

# Estimator points

Any sample mean  $\bar{x}$

Population mean  $\mu$

When we use a summary statistic derived from the sample, (such as the sample mean) as an estimate of the population statistic (such as the population mean) we call it an **estimator point**.

# Desired estimator points

The mean of the sampling distribution should be the same if you calculated the mean of the entire population  
unbiased

Also it is desirable that as the sample size increases,  
The SD should decrease  
So that we have the most precision possible  
And the least standard error

# Back to Utah

Calculate mean and SD for all parks

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

Then do the same for all 10 samples of 2 parks

# Back to Utah

Sample of Two Parks	Total Area (sq mi)	Mean Area (sq mi)
A and B	175	87.5

At the end calculate the mean area for all your samples.

Is this mean the same as for the initial distribution?

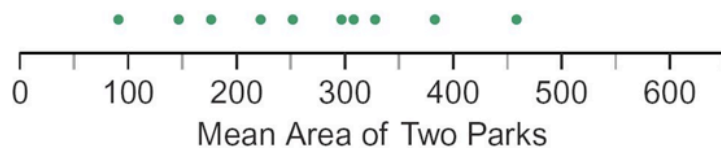
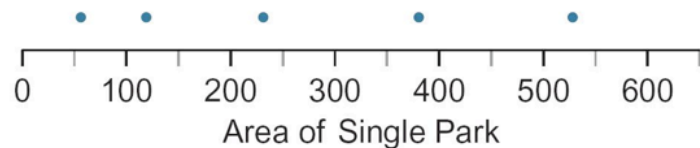
If so, our sample mean an unbiased estimator.

Now calculate the SD of the sampling distribution.

Compare with the previous SD.

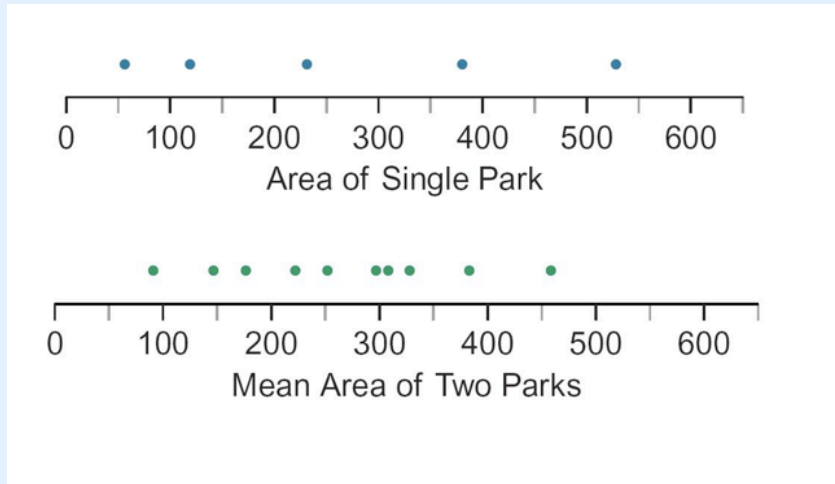
# Back to Utah

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



The SD should be smaller here (105.23) than for the entire population (171.85)

# Back to Utah



Sample size 1

Sample size 2

This means that the spread we have is less if we use Sample sizes of 2 than if we use sample sizes of 1.

The mean is the same, no bias

The spread is different



# Concepts

A simulated sample distribution is the distribution of a sample statistic (the mean) for a large number of repeated samples

The sample distributions are best described by shape, center and spread

Sampling distributions DO NOT necessarily have the same shape as the population from which they were taken

# Concepts

The SD of the sampling distribution is called the standard error

If the sampling distribution is normal, reasonably likely outcomes are those that lie within 2 SD of the mean (95% of data)

P5 page 319 Estimate the range of Utah's national parks

Range = Largest Area - Smallest Area

Select 3 parks at random and calculate the range

1) What is the range of the entire POPULATION?

2) Make a table for the range of groups of 3

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

3) Place your values on a dot plot

4) What is the mean of the sample?

5) Is the sample range biased or unbiased?

# Practice

Page 321 P3, P4, P5, E1, E2, E3, E5, E6, E7, E10,