Math 247: Sample <u>Means</u> of Random Samples (Section 9.1)

Suppose a medical researcher wanted to find out about how air pollution is related to children's health; specifically, she wants to look at lung health and development.

Could she design an experiment to find out whether pollution significantly impacts children's lungs? Why or why not?

No – not ethical! Observational study only – desirable LARGE sample to "control" for confounders.

Since setting up her study as an experiment isn't ethical, she'll have to do an observational study. She would need to gather data from a sample of kids in areas with high pollution and compare the results to either known values (One Sample) or to the values obtained from a sample of kids who live in low pollution areas (Two Samples).

She'll have to determine what she wants to measure on each of the kids in the sample that would tell her about lung health and development. Asthma, lung capacity

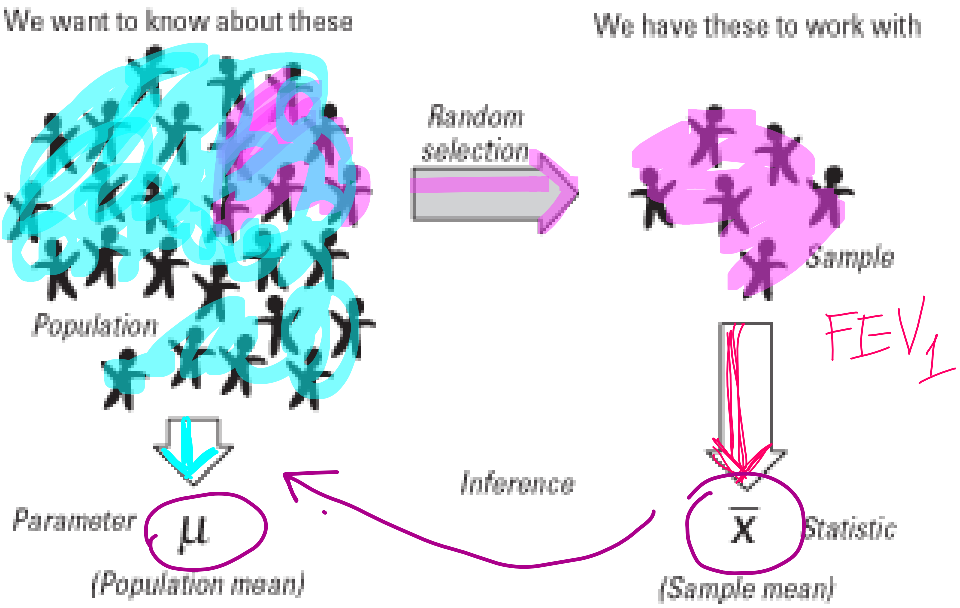
One of the measurements used to determine the health of a person's lungs is the <u>amount</u> of air a person can exhale under force in one second. This is called the "forced expiratory volume in one second". It's measured in liters (like soda bottles) and is abbreviated FEV_1.

Is FEV_1 a <u>qualitative</u> or <u>quantitative</u> variable? Quantitative, i.e numerical

Dealing with **quantitative** research questions: We can measure some value (the Variable of Interest) for each subject in a sample then combine those values into a **<u>sample mean</u>**.

This will tell us about what's going on for the subjects in the sample **<u>ON AVERAGE</u>**. – we are finding out about the group as a whole

Next, we'll want to find out what might be true of the Population of Interest, again, **ON AVERAGE**, based on what's true **ON AVERAGE** for our sample.



(no sample mentioned

population

Air Pollution and Children’s Health. Previous studies have established that the mean FEV$_1$ for all 10-year-old boys is 2.1 liters and that the population standard deviation is 0.3 liters. A random sample of 100 10-year-old boys who live in a community with high levels of ozone pollution is found to have a mean FEV$_1$ of 1.95 liters. Does this data show there is a significant* decrease in lung function in children living in high pollution areas? (*More on types of significance later!)

In the problem above, write down and label of all the quantities mentioned. Use the correct notation for each and state whether it is a statistic or a parameter.

mean (pop) $\mu = 2.1$ L $= 2.10$

S.D. (pop) $\sigma = 0.3$L

PARAMETERS

Sample size $= n = 100$

mean (sample) $= \bar{x} = 1.95$L

STATISTICS

observe the sample mean is lower than known pop mean

How will we determine statistical significance? We need to have a basis to judge how far away a sample mean is from the population mean and how unlikely that difference is just due to. . .

Sampling variability .

How have we measured how “far away” values are in the past?

How spread out means are . . . standard deviation

Z-scores = tell us how far away values are in a NORMAL distribution

We’ll have to investigate this issue before we can proceed with the Air Pollution/Children’s Health example (more later!)

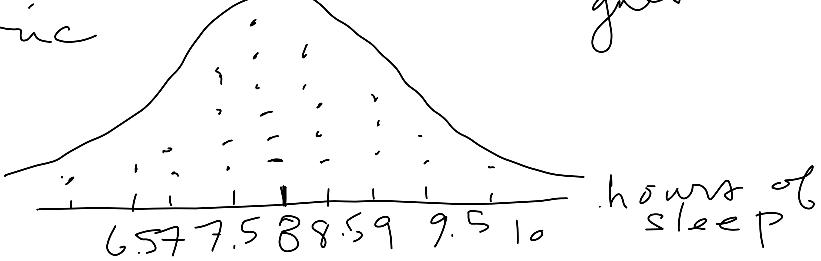
First, let’s look at the concept of a sampling distribution using simulations

of sample means

(Chapter 7 - we looked at the sampling distribution for proportions . . .)

Example: Imagine you had the all the data on how many hours adults sleep per night. What do you think the shape of the distribution of sleep hours would be; i.e., left skewed, right skewed, or symmetric? guess

Sketch your idea here: Symmetric



Using the Sleep Hours data in the Rossman/Chance Applet for Sampling Distribution of the mean, check the shape of the population distribution of Sleep Hours. The shape is Symmetric (18,000)

Now describe shape you see for the sampling distribution of sample means for each of the following and note what happens to the spread of the sampling distribution.

n = 3

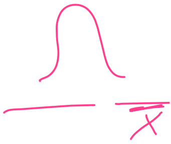
shape symmetric

Spread SD = .865

$SE = \frac{\sigma}{\sqrt{n}}$ $\frac{\sigma\,(\text{sleep all})}{\sqrt{3}}$

n = 10

shape symmetric

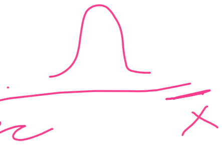


Spread SD = .470

$SE = \frac{\sigma}{\sqrt{10}}$

n = 25

shape symmetric



spread SD = .303

$SE = \frac{\sigma}{\sqrt{25}}$

Observations about shape: For each sample size, the Sampling distribution of the means ($\bar{x}$) was symmetric (NORMAL)

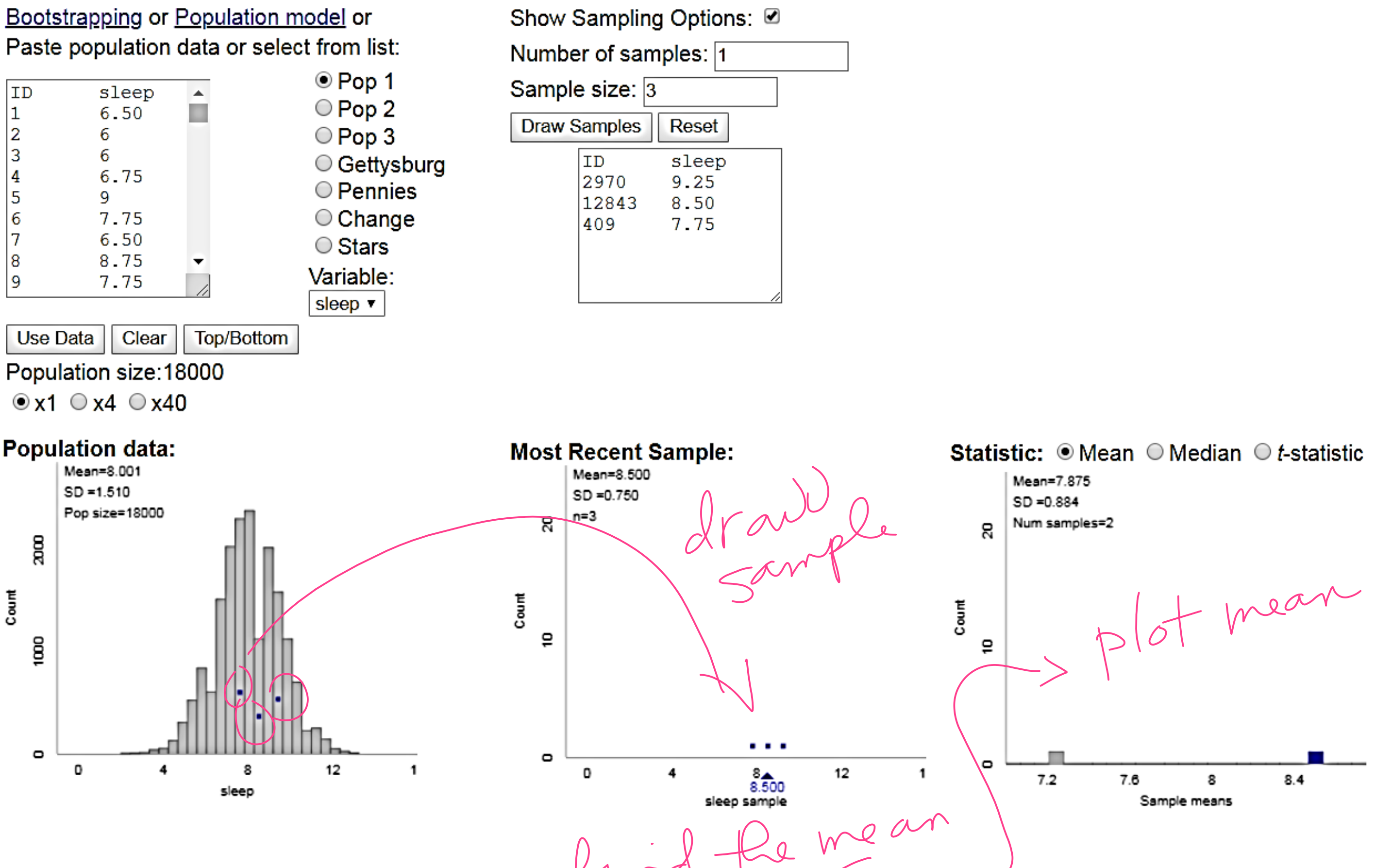
Observations about spread: As sample size got bigger, the S.D. got smaller → less spread → less variability

Summary:

Shape: If the Variable of Interest (quantitative) in the population has a symmetric distribution then the sampling distribution of the sample means will also be approximately symmetric.

Spread: The spread of the Sampling Distribution of the Mean is called the Standard Error, SE.

- The Standard Error gets smaller as the sample size increases
- The Standard Error is given by the formula $\boxed{SE = \frac{\sigma}{\sqrt{n}}}$



Bootstrapping or Population model or
Paste population data or select from list:

```
ID      sleep
1       6.50
2       6
3       6
4       6.75
5       9
6       7.75
7       6.50
8       8.75
9       7.75
```

● Pop 1
○ Pop 2
○ Pop 3
○ Gettysburg
○ Pennies
○ Change
○ Stars

Variable: sleep ▾

Use Data | Clear | Top/Bottom

Population size:18000

● x1 ○ x4 ○ x40

Show Sampling Options: ☑

Number of samples: 1

Sample size: 3

Draw Samples | Reset

```
ID      sleep
2970    9.25
12843   8.50
409     7.75
```

Population data:

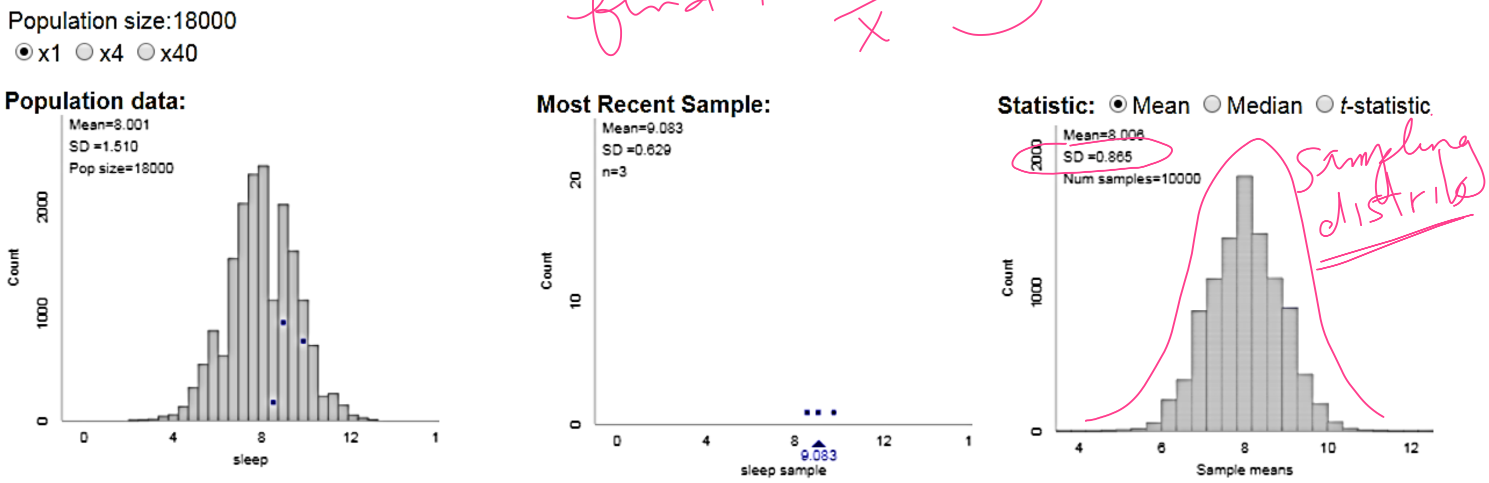
Mean=8.001
SD =1.510
Pop size=18000

Most Recent Sample:

Mean=8.500
SD =0.750
n=3

Statistic: ● Mean ○ Median ○ *t*-statistic

Mean=7.875
SD =0.884
Num samples=2



Population size:18000

● x1 ○ x4 ○ x40

Population data:

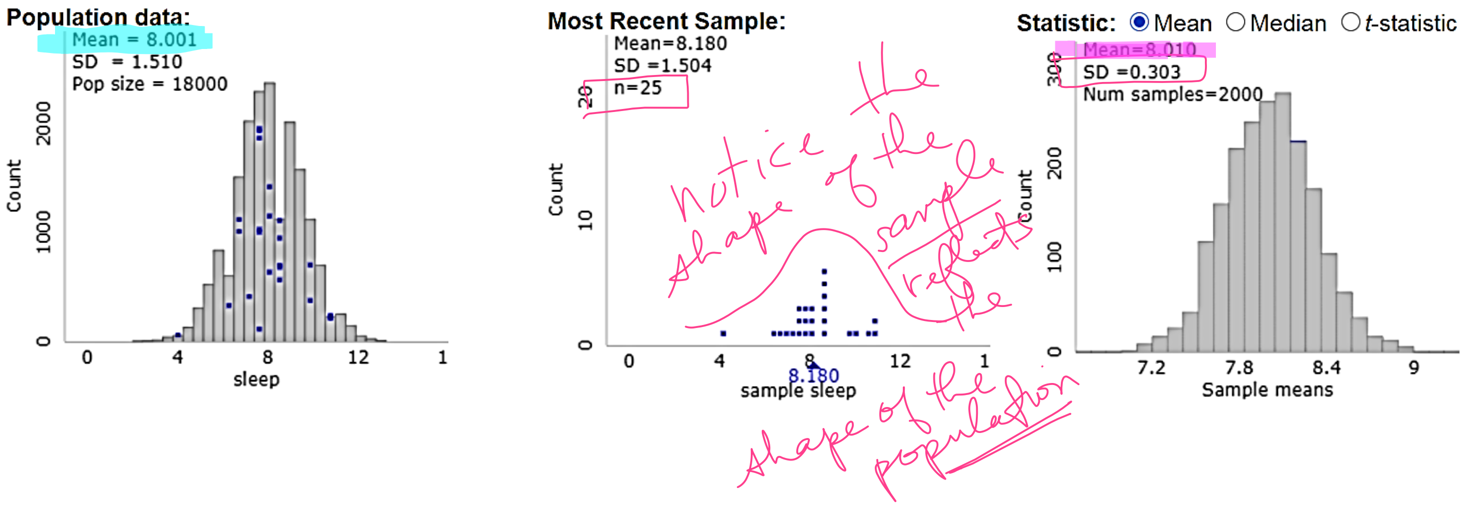
Mean=8.001
SD =1.510
Pop size=18000

Most Recent Sample:

Mean=9.083
SD =0.629
n=3

Statistic: ● Mean ○ Median ○ *t*-statistic

Mean=8.006
SD =0.865
Num samples=10000



Population data:

Mean = 8.001
SD = 1.510
Pop size = 18000

Most Recent Sample:

Mean=8.180
SD =1.504
n=25

Statistic: ● Mean ○ Median ○ *t*-statistic

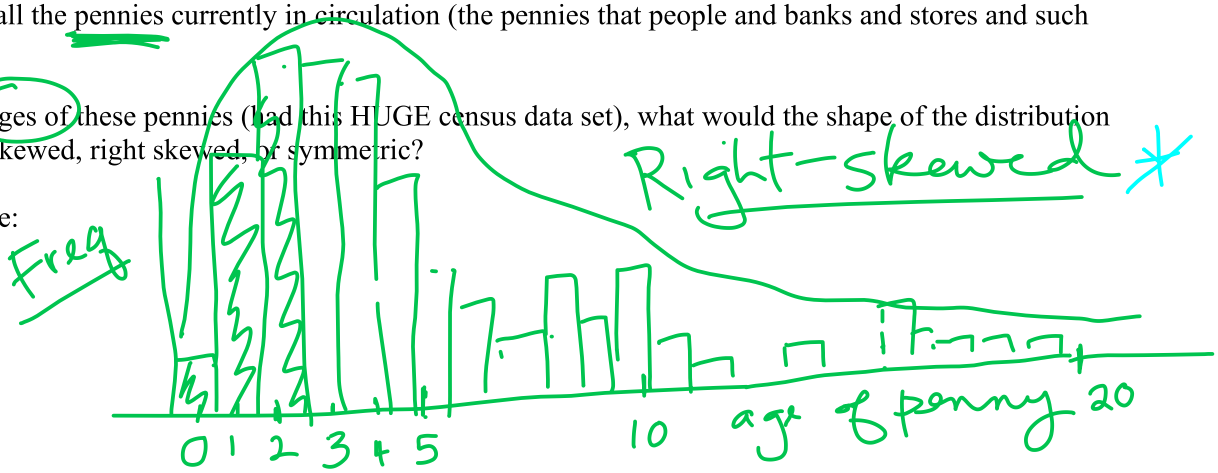
Mean=8.010
SD =0.303
Num samples=2000

Example: Imagine all the pennies currently in circulation (the pennies that people and banks and stores and such actually have).

If you knew all the ages of these pennies (had this HUGE census data set), what would the shape of the distribution of ages be; i.e., left skewed, right skewed, or symmetric?

Right-skewed *

Sketch your idea here:



Using the Penny Age data in the Rossman/Chance Applet for Sampling Distribution of the mean, check the shape of the population distribution of Penny Ages. The shape is Right-skewed

Now describe shape you see for the sampling distribution of sample means for each of the following and note what happens to the spread of the sampling distribution.

n = 3

PROBLEM

shape: right skewed

spread:

SD = ignore

n = 10

PROBLEM

shape: ALMOST symmetric (a little right skewed)

SD = ignore

n = 25

shape: pretty darn symmetric!

mean of $\bar{x}$'s $= \mu$

SD = 1.900

$= \frac{\sigma}{\sqrt{n}} = \frac{9.613}{\sqrt{25}}$

= 1.9226 (not perfect but close - simulation

Sampling Distrib

Observations about shape:

Shape showed some skewing until n=25

For symmetric distrib well need a large sample!

Observations about spread:

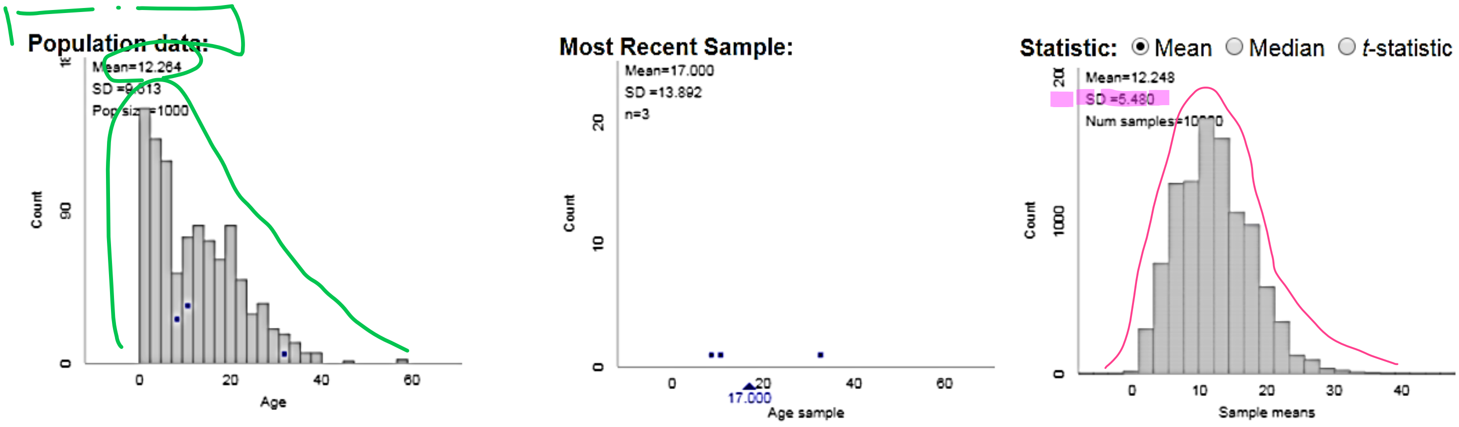
Not relevant until sample size is large - same behavior as before

Summary:

Shape: If the Variable of Interest (quantitative) in the population does NOT have a symmetric distribution then the sampling distribution of the sample means will not be symmetric unless the sample size is LARGE ($n \geq 25$)

Spread: The spread of the Sampling Distribution of the Mean is called the Standard Error, SE.

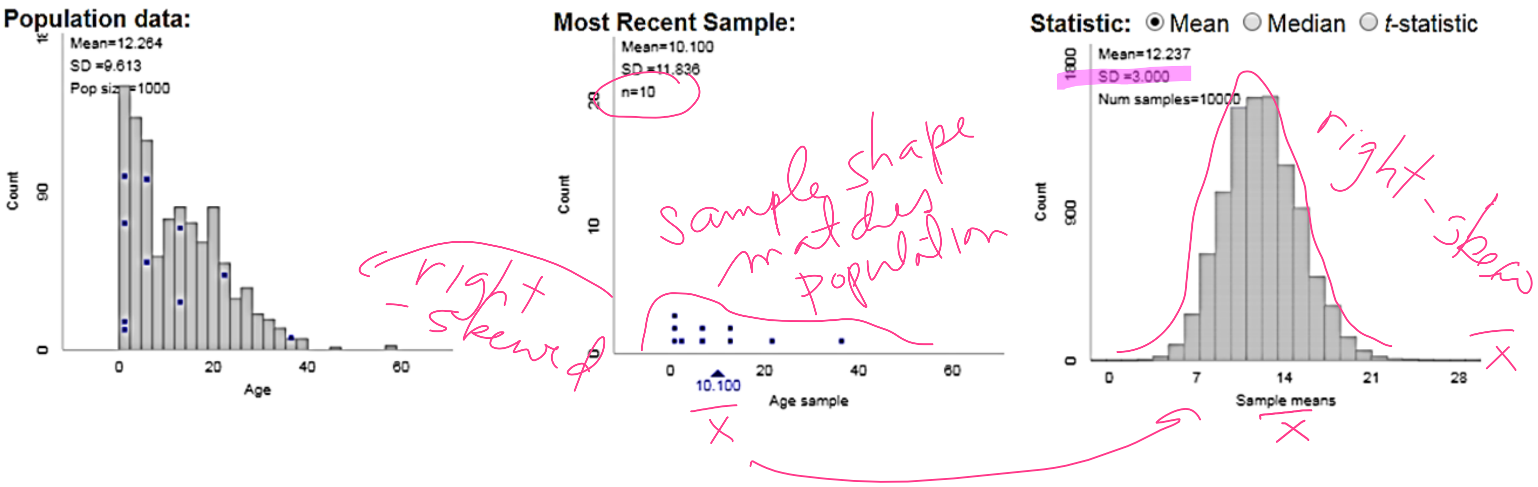
- The Standard Error gets smaller as the sample size increases
- The Standard Error is given by the formula $SE = \frac{\sigma}{\sqrt{n}}$



Population data:
Mean=12.264
SD =9.613
Pop size=1000

Most Recent Sample:
Mean=17.000
SD =13.892
n=3

Statistic: ● Mean ○ Median ○ *t*-statistic
Mean=12.248
SD =5.480
Num samples=10000



Population data:
Mean=12.264
SD =9.613
Pop size=1000

Most Recent Sample:
Mean=10.100
SD =11.836
n=10

Statistic: ● Mean ○ Median ○ *t*-statistic
Mean=12.237
SD =3.000
Num samples=10000

right-skewed

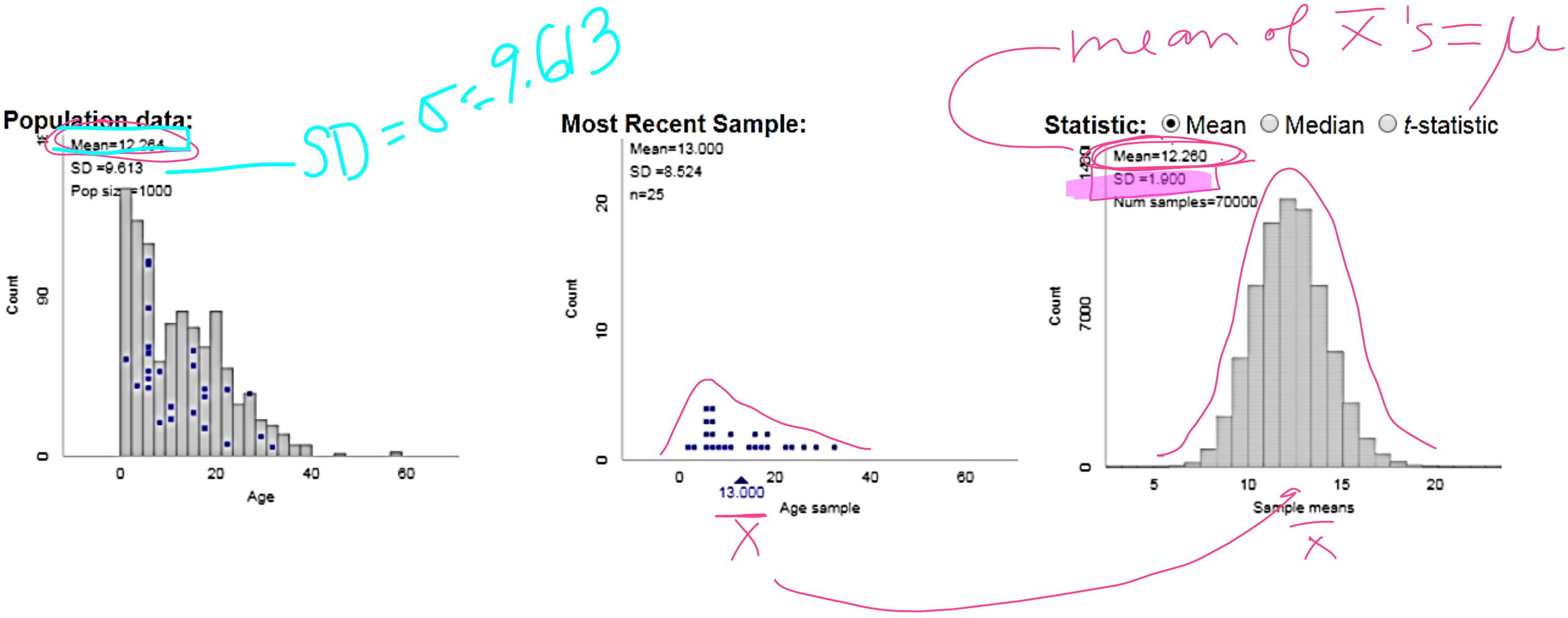
sample shape matches population

$\bar{x}$

right-skewed

$\bar{x}$

$\bar{x}$



$SD = \sigma = 9.613$

mean of $\bar{x}$'s $= \mu$

Population data:
Mean=12.264
SD =9.613
Pop size=1000

Most Recent Sample:
Mean=13.000
SD =8.524
n=25

Statistic: ● Mean ○ Median ○ *t*-statistic
Mean=12.260
SD =1.900
Num samples=70000

$\bar{x}$

$\bar{x}$