Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:

- Types of Sampling
- Sampling Distribution
- Central Limit Theorem

A Probability Question

You flip a coin 300 times. What's the probability you get 173 or more heads?

A Probability Question

You flip a coin 300 times. What's the probability you get 173 or more heads?

Let $X = Binom(300, 0.5)$. We want $P(X \geq 173)$.

A Probability Question

You flip a coin 300 times. What's the probability you get 173 or more heads?

Let $X = Binom(300, 0.5)$. We want $P(X \geq 173)$.

Well... we could think of doing

$$P(X \geq 173) = P(X = 173) + P(X = 174) + \cdots + P(X = 300),$$

but that's hopeless because the sum is too big.

A Probability Question

You flip a coin 300 times. What's the probability you get 173 or more heads?

Let $X = Binom(300, 0.5)$. We want $P(X \geq 173)$.

Well... we could think of doing

$$P(X \geq 173) = P(X = 173) + P(X = 174) + \cdots + P(X = 300),$$

but that's hopeless because the sum is too big.

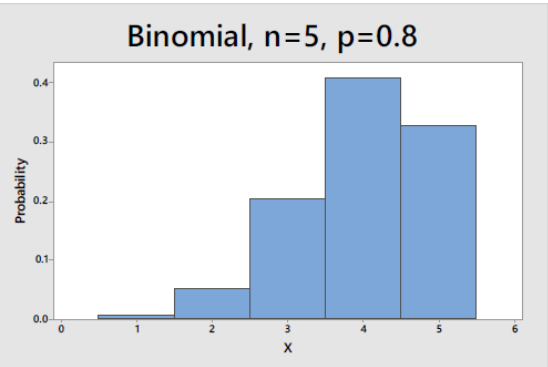
Also, complementary probability

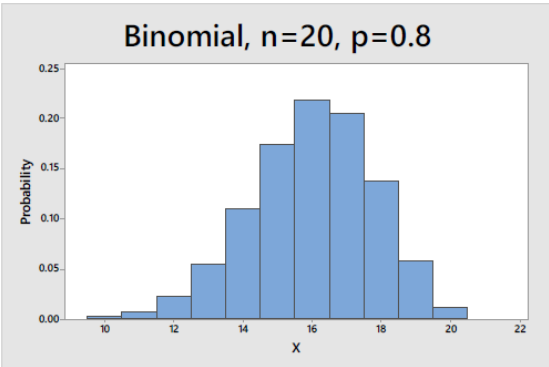
$$\begin{aligned} P(X \geq 173) &= 1 - P(X \leq 172) \\ &= 1 - P(X = 172) - P(X = 171) - \cdots - P(X = 0) \end{aligned}$$

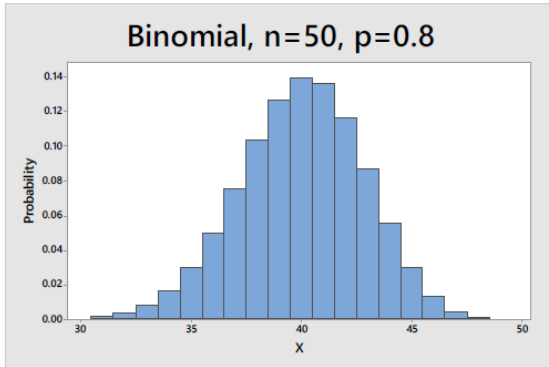
won't work: doing that is just as difficult as the given problem.

The Normal Model to the Rescue

Cool observation: As n gets large, the Binomial model begins to look like the Normal model.





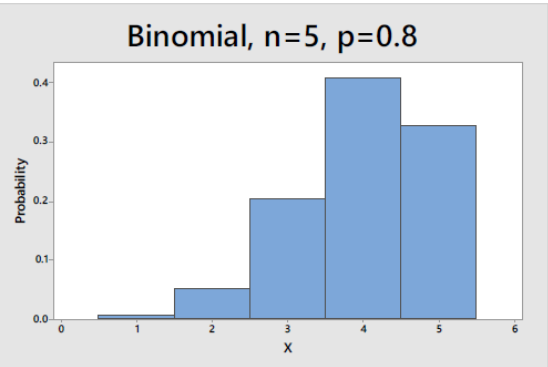


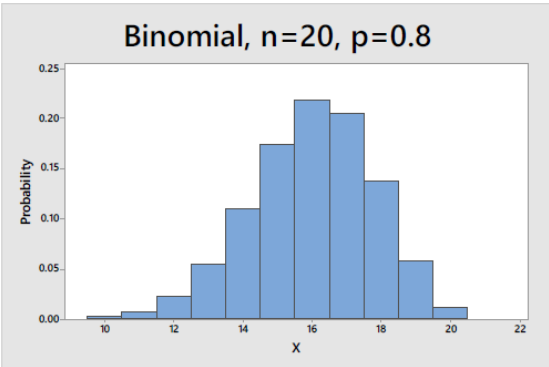
Recall that the mean and standard deviation of $X = Binom(n, p)$ are

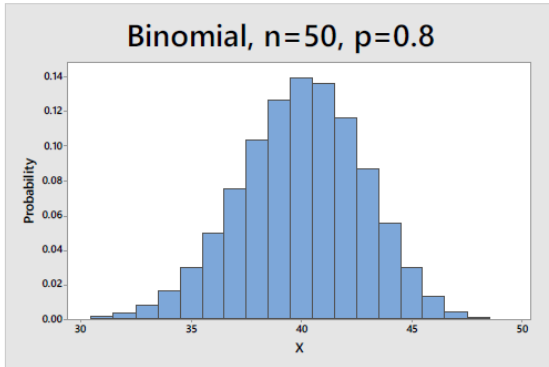
$$E(X) = np \quad \text{and} \quad SD(X) = \sqrt{npq}.$$

The Normal Model to the Rescue

Cool observation: As n gets large, the Binomial model begins to look like the Normal model.







Recall that the mean and standard deviation of $X = Binom(n, p)$ are

$$E(X) = np \quad \text{and} \quad SD(X) = \sqrt{npq}.$$

Cool idea: we can imitate a Binomial model using a Normal model $Y = N(\mu, \sigma)$ by setting

$$\mu = E(X) = np \quad \text{and} \quad \sigma = SD(X) = \sqrt{npq}.$$

You flip a coin 300 times. What's the probability you get 173 or more heads?

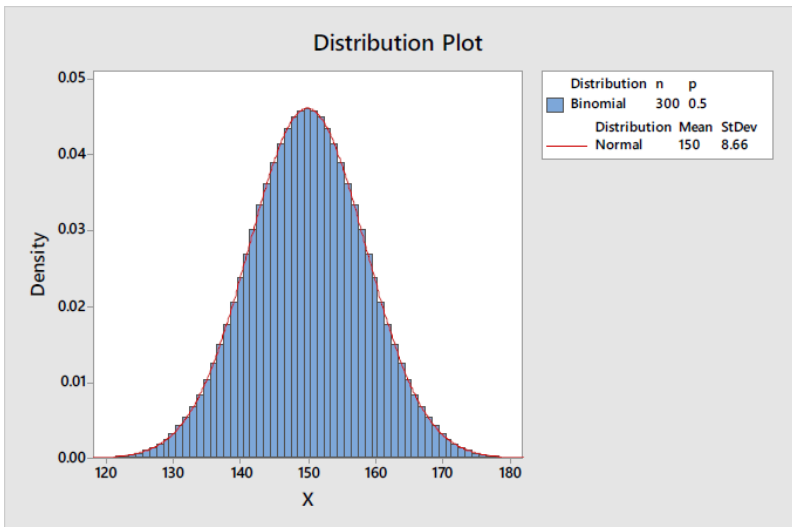
Set up a Normal model $Y = N(\mu, \sigma)$ with

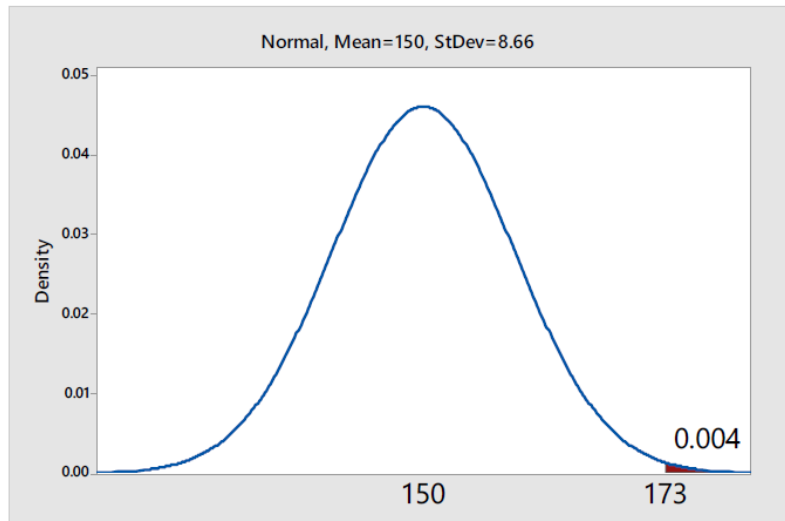
- $\mu = np = 300 \times 0.5 = 150$
- $\sigma = \sqrt{npq} = \sqrt{300 \times 0.5 \times 0.5} \simeq 8.66$

You flip a coin 300 times. What's the probability you get 173 or more heads?

Set up a Normal model $Y = N(\mu, \sigma)$ with

- $\mu = np = 300 \times 0.5 = 150$
- $\sigma = \sqrt{npq} = \sqrt{300 \times 0.5 \times 0.5} \simeq 8.66$





$$P(X \geq 173) \simeq P(Y \geq 173) \simeq 0.004.$$

When is it Safe to Approximate a Binomial via a Normal?

In general, statisticians use this practice as long as both $np \geq 10$ and $nq \geq 10$ (at least 10 successes and 10 failures)

When is it Safe to Approximate a Binomial via a Normal?

In general, statisticians use this practice as long as both $np \geq 10$ and $nq \geq 10$ (at least 10 successes and 10 failures)

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

When is it Safe to Approximate a Binomial via a Normal?

In general, statisticians use this practice as long as both $np \geq 10$ and $nq \geq 10$ (at least 10 successes and 10 failures)

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

Option 1: Let $X = Binom(80, 1/6)$, and you can compute

$$P(X \leq 17 = P(X = 0) + P(X = 1) + \cdots + P(X = 17).$$

(YOU can compute this... but I personally won't!)

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

Option 2: See if we can use a Normal model to approximate this Binomial situation.

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

Option 2: See if we can use a Normal model to approximate this Binomial situation.

Note that $np = 80/6 \simeq 13.3 \geq 10$ and $nq = 80(5/6) \simeq 66.6 \geq 10$.

Let $Y = N(\mu = np, \sigma = \sqrt{npq}) = N(40/3, 10/3)$.

Since we expect more than 10 Successes/Failures,

$$\text{So } P(X \leq 17) \simeq P(Y \leq 17).$$

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

Option 2: See if we can use a Normal model to approximate this Binomial situation.

Note that $np = 80/6 \simeq 13.3 \geq 10$ and $nq = 80(5/6) \simeq 66.6 \geq 10$.

Let $Y = N(\mu = np, \sigma = \sqrt{npq}) = N(40/3, 10/3)$.

Since we expect more than 10 Successes/Failures,

$$\text{So } P(X \leq 17) \simeq P(Y \leq 17).$$

• If you want to use a table and z-scores, note that $\frac{17-40/3}{10/3} = 1.1$.

Thus $P(Y \leq 17) = P(Z \leq 1.1) \simeq 0.8643$.

You roll a die 80 times. What are the chances of rolling at most 17 sixes?

Option 2: See if we can use a Normal model to approximate this Binomial situation.

Note that $np = 80/6 \simeq 13.3 \geq 10$ and $nq = 80(5/6) \simeq 66.6 \geq 10$.

Let $Y = N(\mu = np, \sigma = \sqrt{npq}) = N(40/3, 10/3)$.

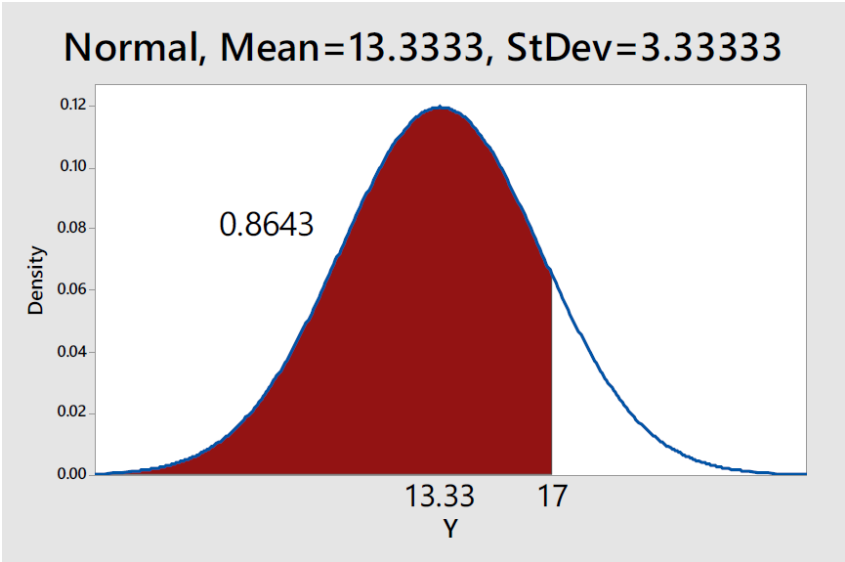
Since we expect more than 10 Successes/Failures,

$$\text{So } P(X \leq 17) \simeq P(Y \leq 17).$$

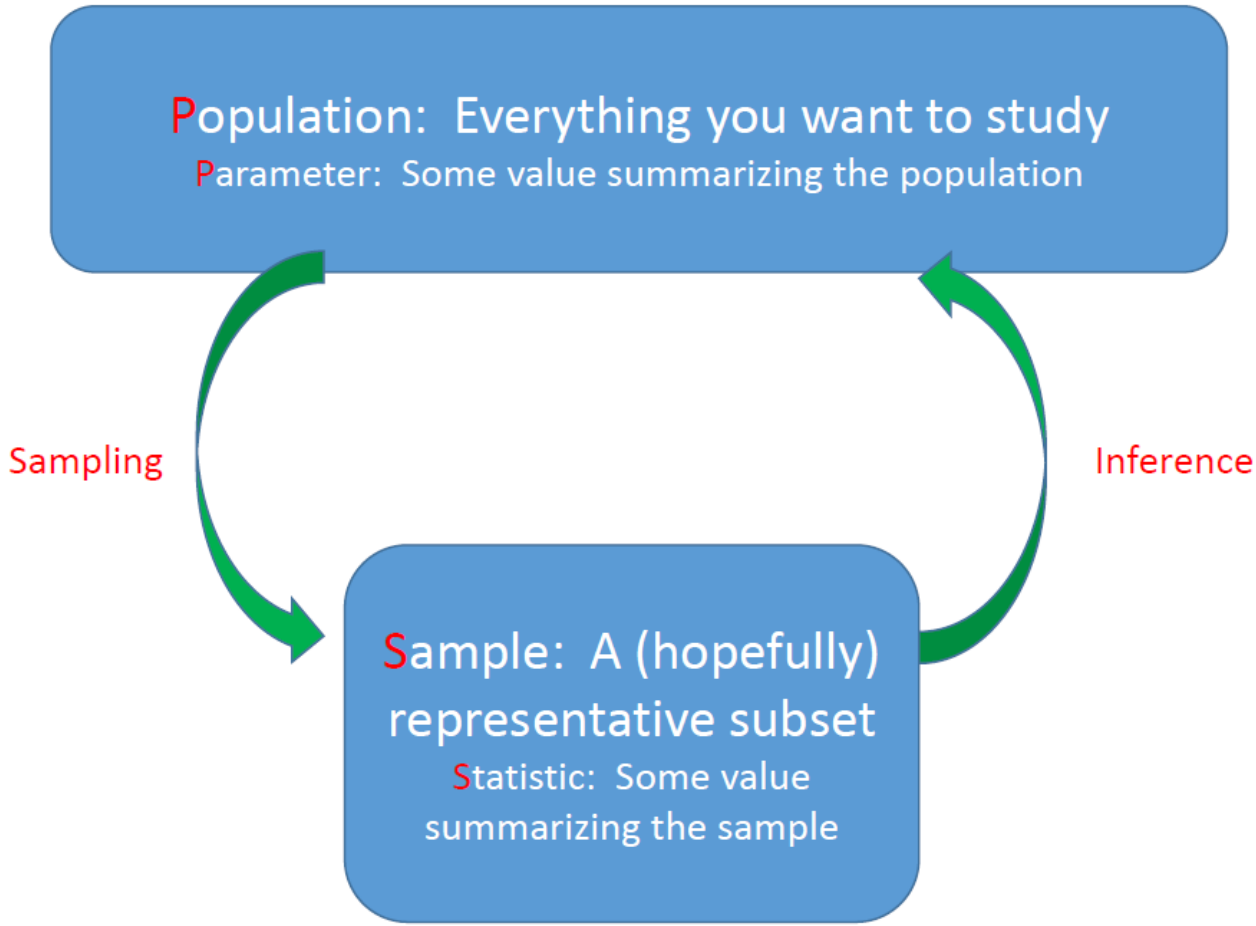
- If you want to use a table and z-scores, note that $\frac{17-40/3}{10/3} = 1.1$.

Thus $P(Y \leq 17) = P(Z \leq 1.1) \simeq 0.8643$.

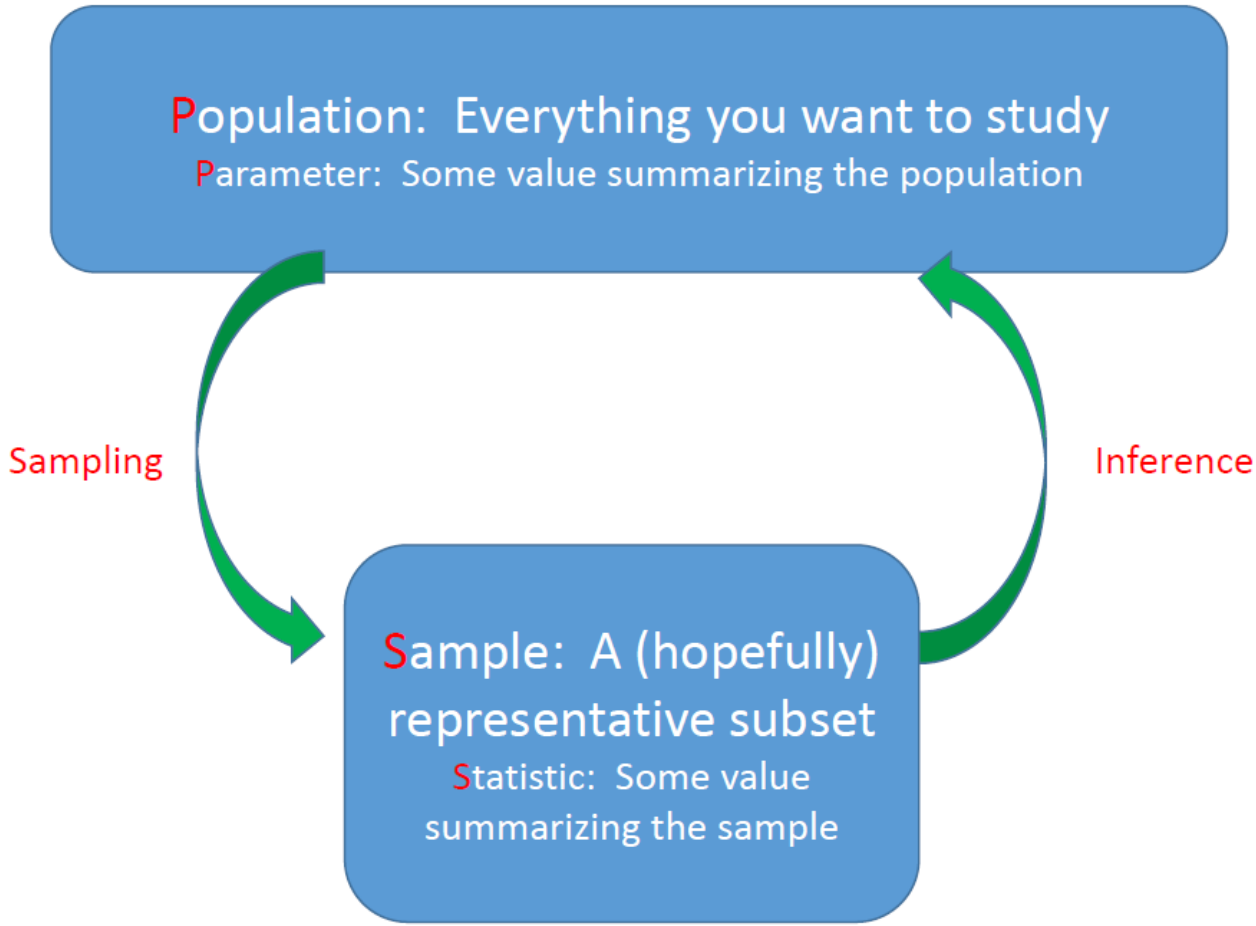
- If you want to use Minitab:



Bringing Probability and Statistics Together: Populations and Samples



Bringing Probability and Statistics Together: Populations and Samples



Statistics is really about learning to draw a representative sample, calculating a statistic, and understanding what inferences can be drawn about the population parameter.

Biases: Some Examples

Vocabulary: The term **bias** is used when the sample is not representative of the population in some way. Good sampling is about reducing as much bias as possible

Biases: Some Examples

Vocabulary: The term **bias** is used when the sample is not representative of the population in some way. Good sampling is about reducing as much bias as possible

Population: A huge pot of soup
Sample: A small spoonful from the top

Biases: Some Examples

Vocabulary: The term **bias** is used when the sample is not representative of the population in some way. Good sampling is about reducing as much bias as possible

Population: A huge pot of soup
Sample: A small spoonful from the top
Reasons these samples may be biased:

- The soup has heavy ingredients that always sink to the bottom
- You recently added salt but didn't stir the pot before tasting

Biases: Some Examples

Vocabulary: The term **bias** is used when the sample is not representative of the population in some way. Good sampling is about reducing as much bias as possible

Population: A huge pot of soup
Sample: A small spoonful from the top
Reasons these samples may be biased:

- The soup has heavy ingredients that always sink to the bottom
- You recently added salt but didn't stir the pot before tasting

Population: All voters in the next election Sample: 1007 people contacted via home phone

Biases: Some Examples

Vocabulary: The term **bias** is used when the sample is not representative of the population in some way. Good sampling is about reducing as much bias as possible

Population: A huge pot of soup
Sample: A small spoonful from the top
Reasons these samples may be biased:

- The soup has heavy ingredients that always sink to the bottom
- You recently added salt but didn't stir the pot before tasting

Population: All voters in the next election Sample: 1007 people contacted via home phone
Reasons these samples may be biased:

- The person with a home phone may not be a typical voter
- We are trying to build a sample based on a future event!

Two Powerful Ideas about Samples

1. One of the best ways to avoid bias is by introducing random elements into the sampling process.

Soup: Stir the pot right before tasting
Voters: Reach out to random phone numbers (cell and/or landline)

Two Powerful Ideas about Samples

1. One of the best ways to avoid bias is by introducing random elements into the sampling process.

Soup: Stir the pot right before tasting
Voters: Reach out to random phone numbers (cell and/or landline)

2. The sample size does NOT need to be some percentage of the population size! Larger samples are better irrespective of the population size.

Tasting a small pot of soup gives you the same amount of info as tasting a big pot of soup.
Tasting 3 spoonfuls of a pot is better than tasting 1 spoonful.

Types of Sampling

Types of Sampling

Population: Everything you want to study

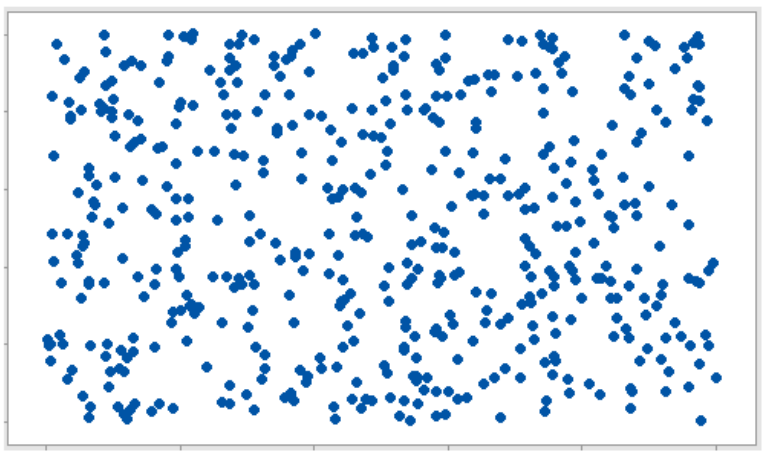
All people in the US

Sampling Frame
(The universe you will be picking from)

All people in the US with SSNs from SSA database

The Simple Random Sample (SRS)

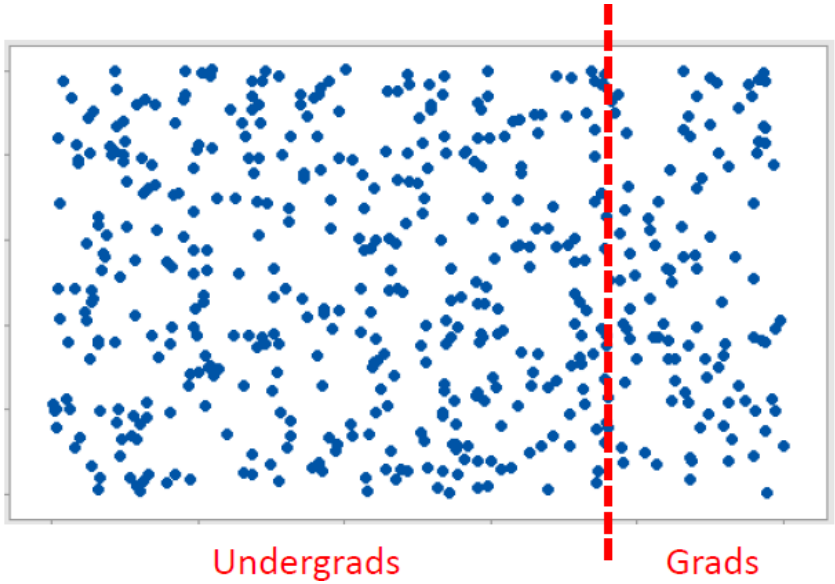
Imagine that each point in the sampling frame is a person. We just pick a certain number of random points.



Stratified Random Sampling

What is the average GPA of UCSD students?

Since grads and undergrads have much different average GPAs, you split the sample into two groups, do an SRS on each, and then combine the results.



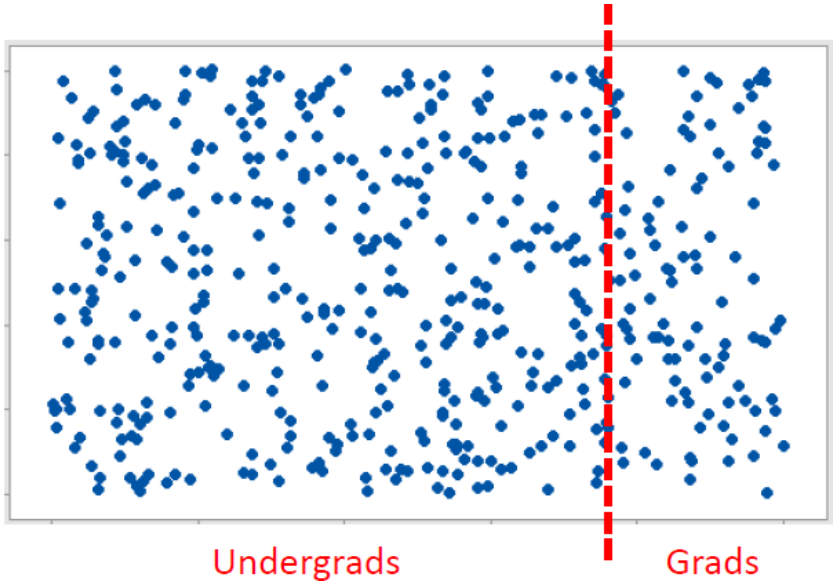
Sampling Frame:

All UCSD students

Stratified Random Sampling

What is the average GPA of UCSD students?

Since grads and undergrads have much different average GPAs, you split the sample into two groups, do an SRS on each, and then combine the results.



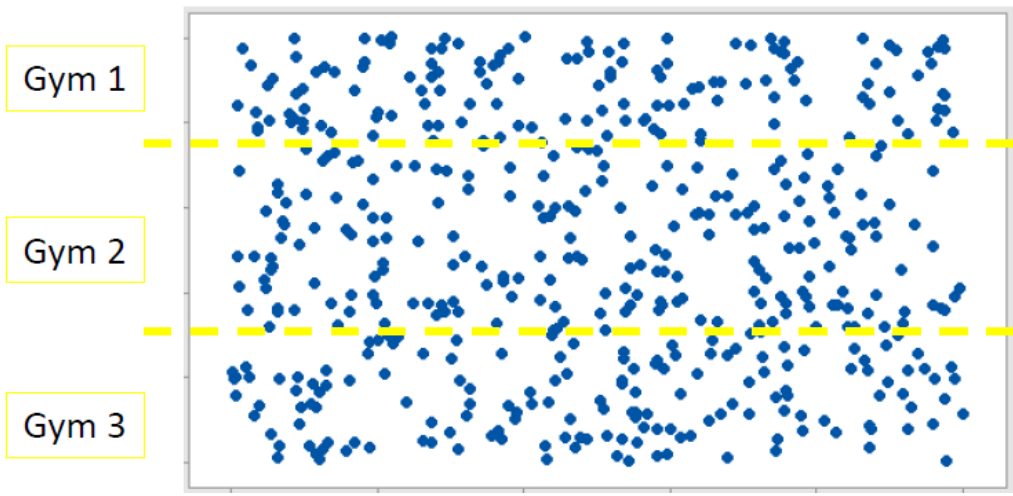
Sampling Frame:

All UCSD students

Cluster Sampling

What is the average GPA of UCSD students?

You and two friends decide to ask people as they walk into various gyms on campus.



Sampling Frame:

All UCSD students that go to the gym

What's the Difference: Stratified vs. Cluster Sampling

In **stratified sampling**, you break the sample frame into pieces because you believe those pieces are homogeneous in relation to the parameter you are measuring.
(Undergrads have lower GPAs; grads have higher GPAs.)

What's the Difference: Stratified vs. Cluster Sampling

In **stratified sampling**, you break the sample frame into pieces because you believe those pieces are homogeneous in relation to the parameter you are measuring.
(Undergrads have lower GPAs; grads have higher GPAs.)

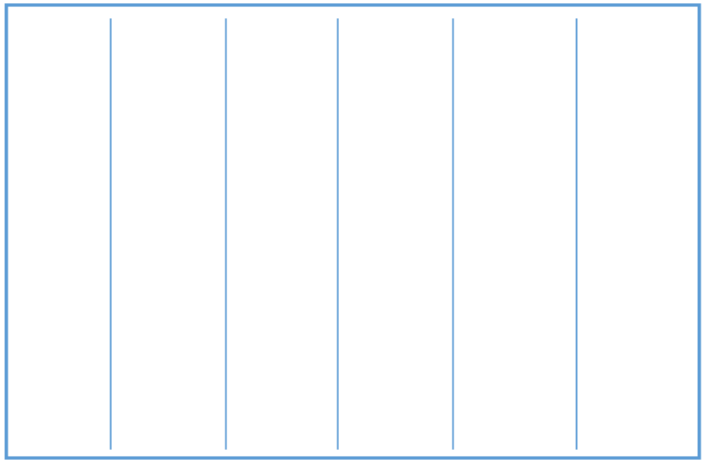
In **cluster sampling**, you break the sample frame into pieces because it makes life convenient. Your pieces will be heterogeneous in relation to the parameter you are measuring.
(Gym 1 has undergrads and grads; so do the other gyms.)

Types of Sampling

Systematic Sampling

What is the average GPA of UCSD students?

You ask every 10^{th} person that you see on campus.

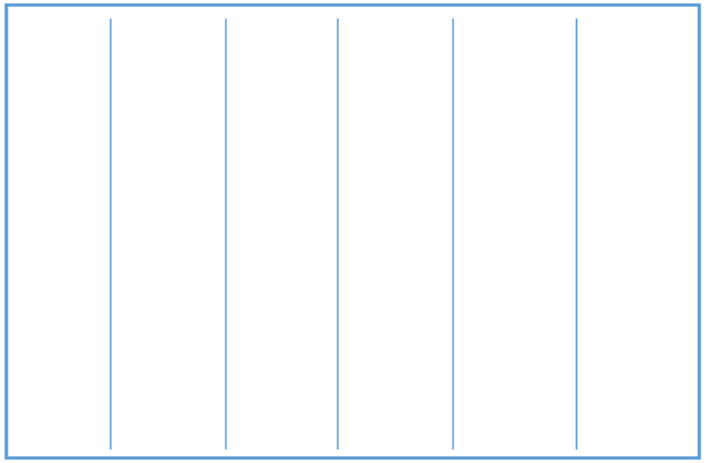


Types of Sampling

Systematic Sampling

What is the average GPA of UCSD students?

You ask every 10^{th} person that you see on campus.

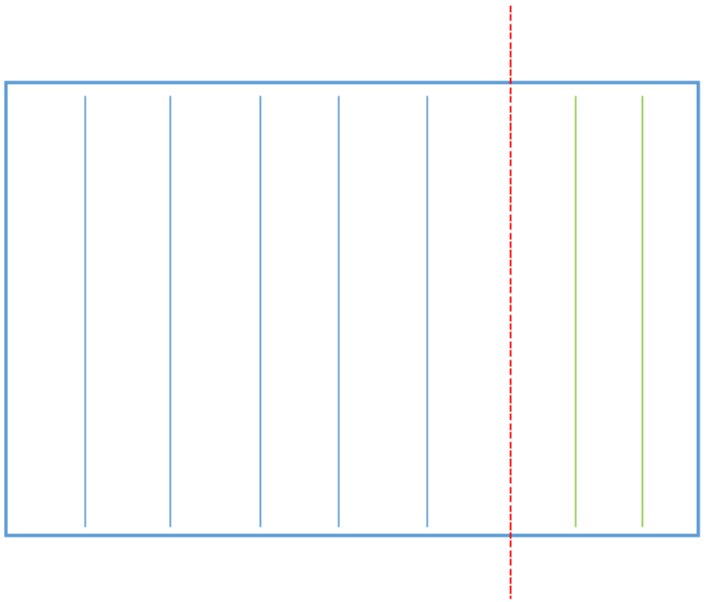


Multistage Sampling

What is the average GPA of UCSD students?

You focus on undergrads today, asking every 4^{th} person. You focus on grads tomorrow, asking every 4^{th} person.

Multistage sampling uses 2 or more of the previous methods (do NOT count a SRS as a method).

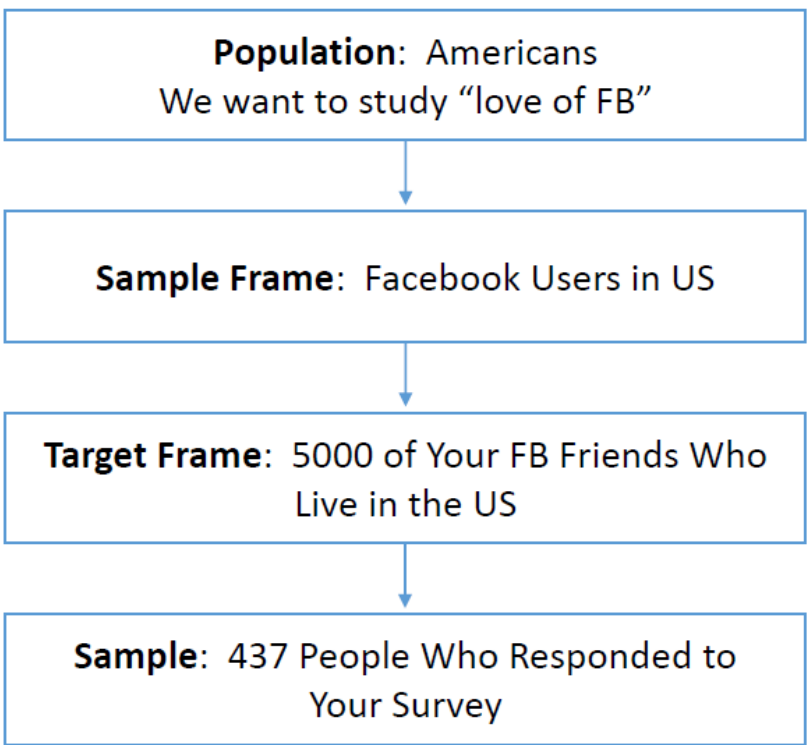


The Horrors of Sampling (AKA Common Sampling Biases)

Your boss at Facebook says

> "We want to know how much Americans love Facebook."

Here is what happens:



Bad Sample Frame Bias: We went from Americans to Americans on FB! We have completely underrepresented people not on FB (they probably hate it!).

Convenience Sample Bias: Maybe your friends also work at FB and are inclined to love (or hate) it.

Volunteer Bias: Those who complete something usually look different than those who don't.

Practice

You want to know the average number of siblings of UCSD students. This idea is a

1. Parameter
2. Statistic
3. Population
4. Sample

Practice

You want to know the average number of siblings of UCSD students. This idea is a

1. Parameter
2. Statistic
3. Population
4. Sample

Answer: 1. The parameter is the idea being studied in the population Population: All UCSD students

Practice

You ask some students and calculate the average. You found a:

1. Parameter
2. Statistic
3. Population
4. Sample

Practice

You ask some students and calculate the average. You found a:

1. Parameter
2. Statistic
3. Population
4. Sample

Answer: 2.
You drew a sample ("some students") and calculated the idea of interest in the sample. That is a statistic.

Practice

You do this study again. This time you ask every 6th student you se on campus. This is a

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Practice

You do this study again. This time you ask every 6th student you se on campus. This is a

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Answer: 4.

Practice

You do this study again. You randomly pick 4 dorms from the list of all dorms and then randomly ask people from those dorms.

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Practice

You do this study again. You randomly pick 4 dorms from the list of all dorms and then randomly ask people from those dorms.

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Answer: 3.
You've broken the population into pieces and the thing you want to study (sibling average) will likely look quite similar for each piece (dorms).

Practice

You do this study again. You read that LGBT students tend to have more siblings because your chance of being LGBT goes up as you get later in the birth order of your family.
You ask randomly chosen LGBT students and ask about their number of siblings. You also randomly choose non-LGBT students and ask.

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Practice

You do this study again. You read that LGBT students tend to have more siblings because your chance of being LGBT goes up as you get later in the birth order of your family.
You ask randomly chosen LGBT students and ask about their number of siblings. You also randomly choose non-LGBT students and ask.

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Answer: 2.
The groups here are internally homogeneous with respect to the thing we are trying to measure (# of siblings) since LGBTers will have higher totals and non-LGBTers will have lower totals. The group averages will be different from one another.

Practice

You do this study again. You choose 150 names at random from UCSD's complete list of current students

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Practice

You do this study again. You choose 150 names at random from UCSD's complete list of current students

1. Simple Random Sample (SRS)
2. Stratified Sample
3. Cluster Sample
4. Systematic Sample
5. Multistage Sample

Answer: 1.

Sampling Distributions

Population: Your universe of interest
Parameter: Some number summarizing the population

Sample: Some subset of population
Statistic: Some number summarizing the sample

Sampling Distributions

Population: Your universe of interest
Parameter: Some number summarizing the population

Sample: Some subset of population
Statistic: Some number summarizing the sample

Summary Number	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard Deviation	s	σ
Correlation	r	ρ
Regression Coefficient	b	β

Sampling Distributions

Population: Your universe of interest
Parameter: Some number summarizing the population

Sample: Some subset of population
Statistic: Some number summarizing the sample

Summary Number	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard Deviation	s	σ
Correlation	r	ρ
Regression Coefficient	b	β

Important Idea: Even if you "draw" (i.e. collect) a representative sample, your statistic may not be equal to the parameter.

Indeed, if we draw multiple samples, you will get different values for the statistic. How will these values relate to the parameter?

Sampling Distributions

The two most common parameters we gather on populations are:

- **Proportions**
  Examples:
  - % of people that go to college
  - % of people who are LGBT
- **Means**
  Examples:
  - Average weight of U.S. men
  - Mean SAT score of UCSD students

Sampling Distributions

The two most common parameters we gather on populations are:

- **Proportions**
  Examples:
  - % of people that go to college
  - % of people who are LGBT
- **Means**
  Examples:
  - Average weight of U.S. men
  - Mean SAT score of UCSD students

When the populations are big, we must draw a (random) sample and estimate these parameters using statistics

Sampling Distributions

The two most common parameters we gather on populations are:

- **Proportions**
  Examples:
  - % of people that go to college
  - % of people who are LGBT
- **Means**
  Examples:
  - Average weight of U.S. men
  - Mean SAT score of UCSD students

When the populations are big, we must draw a (random) sample and estimate these parameters using statistics

Because of randomness, there is variation in this statistic.
Example: 3 polls for a political candidate might show 53%, 49%, and 52.1% support.

Sampling Distributions

The two most common parameters we gather on populations are:

- **Proportions**
  Examples:
  - % of people that go to college
  - % of people who are LGBT
- **Means**
  Examples:
  - Average weight of U.S. men
  - Mean SAT score of UCSD students

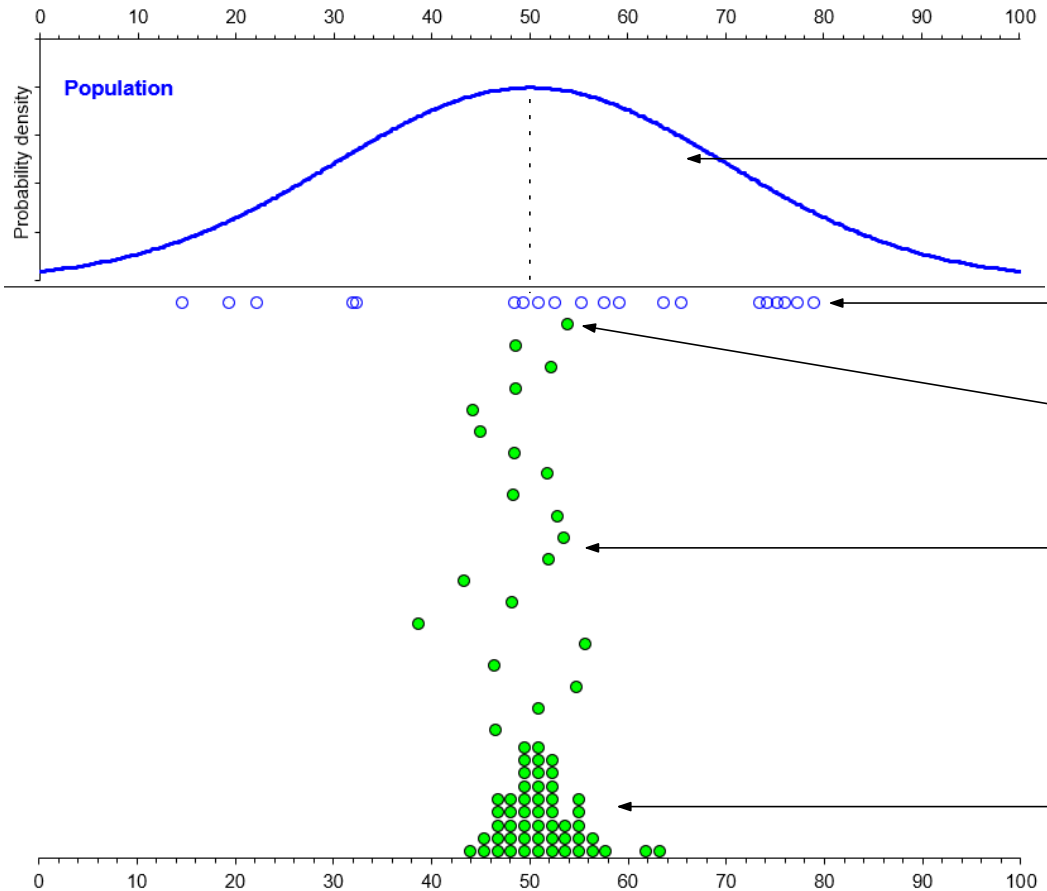
When the populations are big, we must draw a (random) sample and estimate these parameters using statistics

Because of randomness, there is variation in this statistic.
Example: 3 polls for a political candidate might show 53%, 49%, and 52.1% support.

We want to understand and visually display this variation.

Exploratory Software for Confidence Intervals

ESCI: Excel spreadsheets (by G. Cummings) to explore sampling and confidence intervals. (free download, enable Excel macros to use)



Distribution of the idea we're studying (e.g. weight of alien) This is invisible to us!

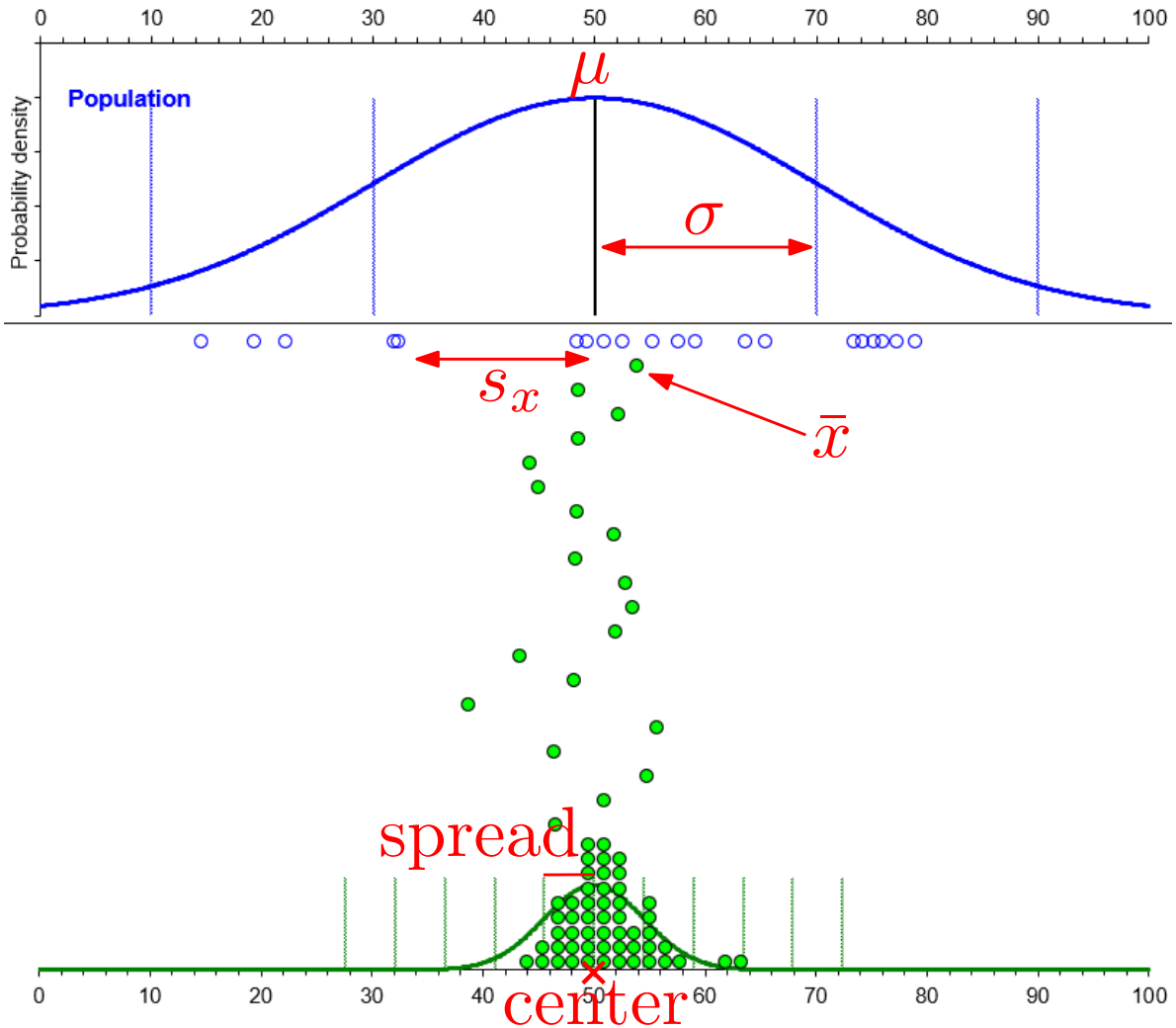
The individuals in the particular sample we drew

The mean weight from the pictured sample

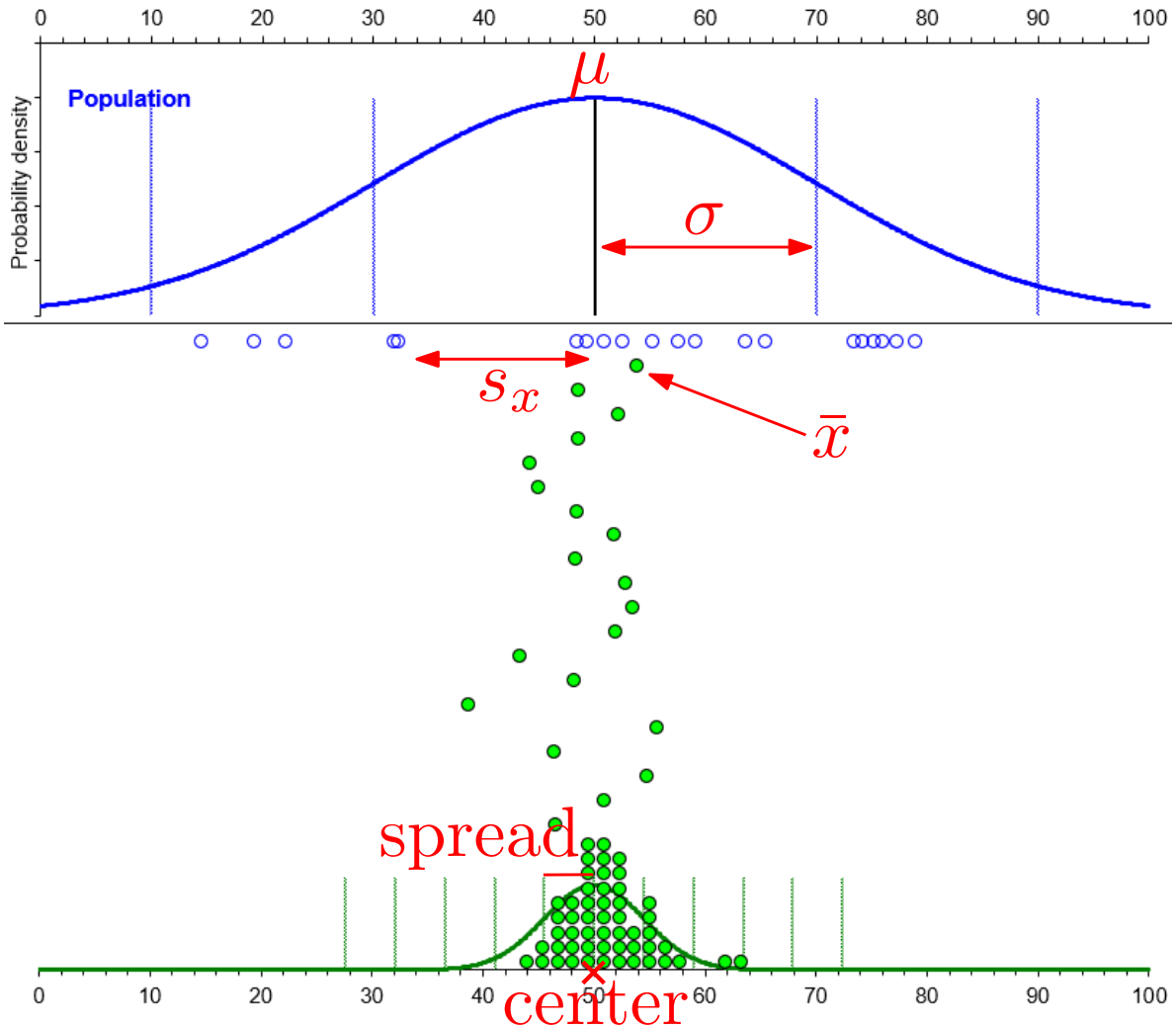
Means from other samples

Making a histogram of all the means from all the samples

Some Notation

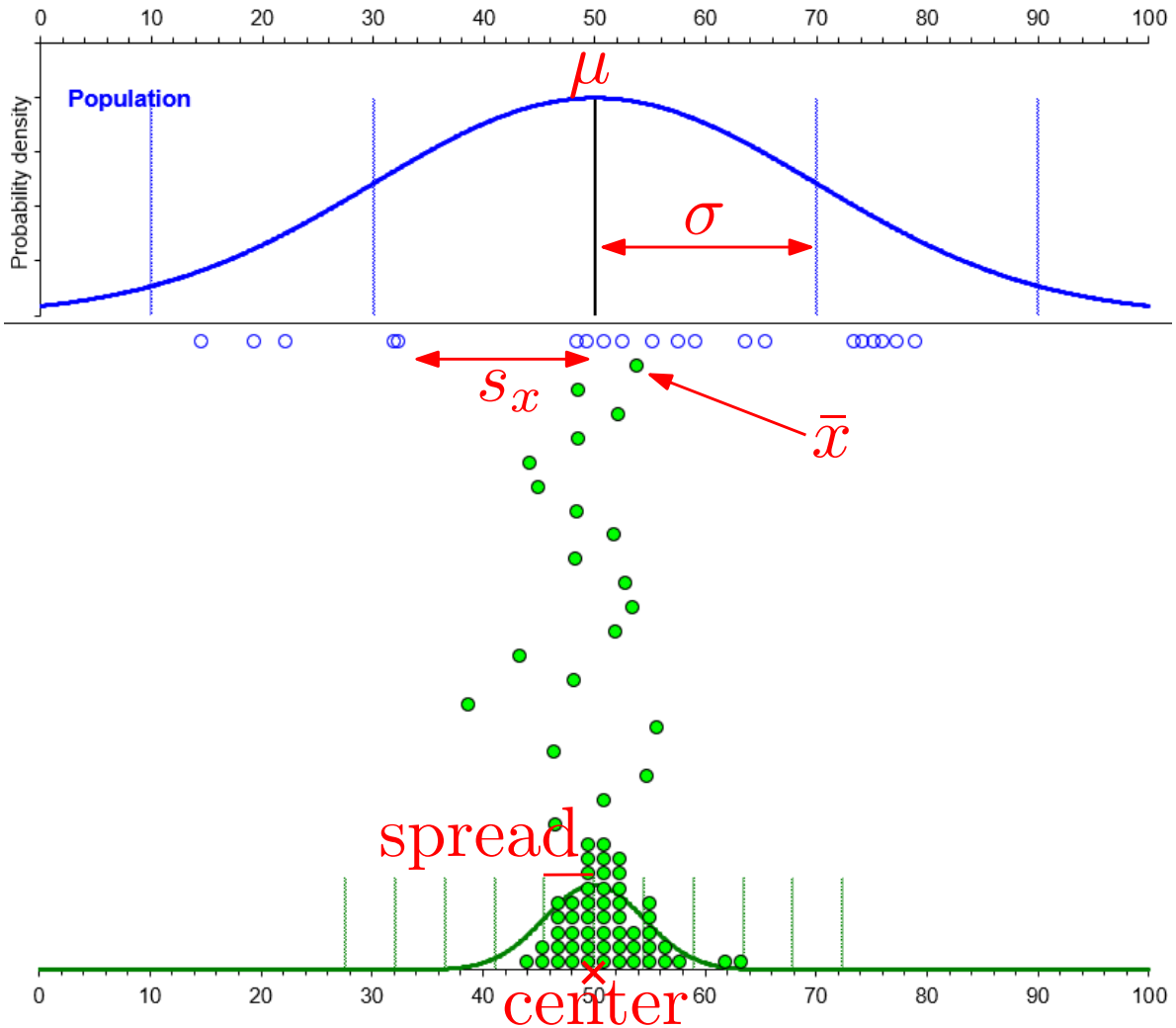


Some Notation



Population:
Center: μ
Spread: σ

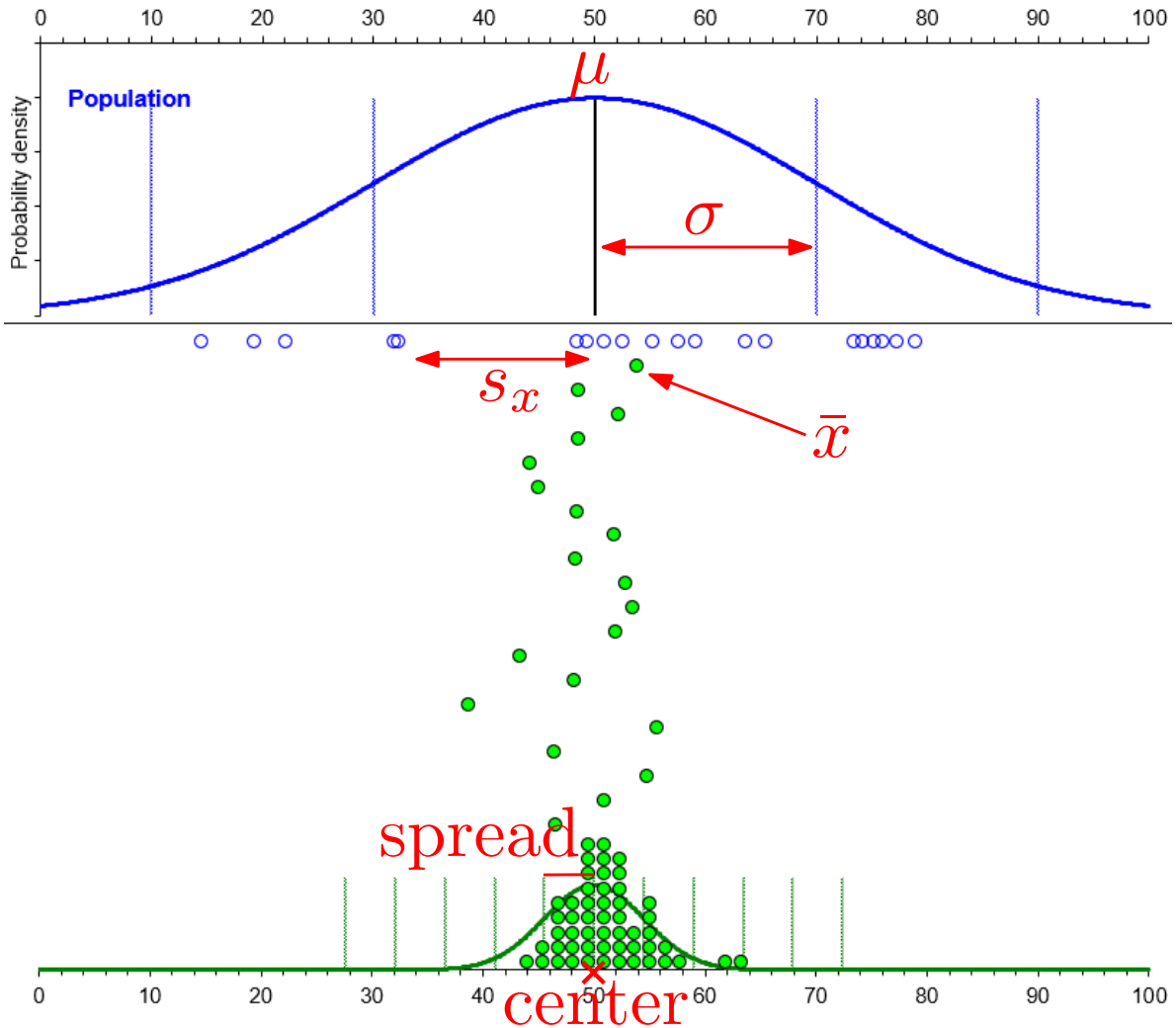
Some Notation



Population:
Center: μ
Spread: σ

One sample:
Center: $\bar{x}$
Spread: s_x

Some Notation



Population:
Center: μ
Spread: σ

One sample:
Center: $\bar{x}$
Spread: s_x

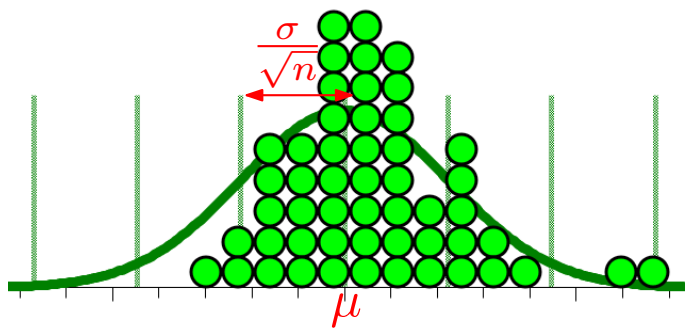
Sampling distribution:
Center: ?
Spread: ?

A Major Discovery!

As the picture suggests, the center of the sampling distribution is also at μ. (no new letter needed!)

An important result from statistics proves that the spread of the sampling distribution, called the **Standard Error** (SE), is just $\frac{\sigma}{\sqrt{n}}$.

Finally, the sampling distribution is a Normal Curve.

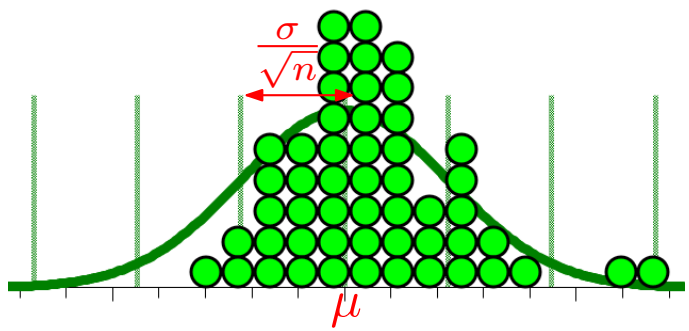


A Major Discovery!

As the picture suggests, the center of the sampling distribution is also at μ. (no new letter needed!)

An important result from statistics proves that the spread of the sampling distribution, called the **Standard Error** (SE), is just $\frac{\sigma}{\sqrt{n}}$.

Finally, the sampling distribution is a Normal Curve.



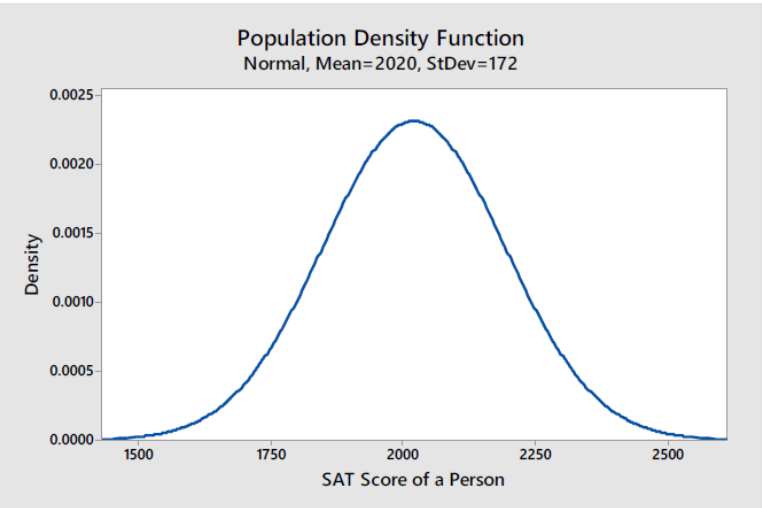
The green curve is simply $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

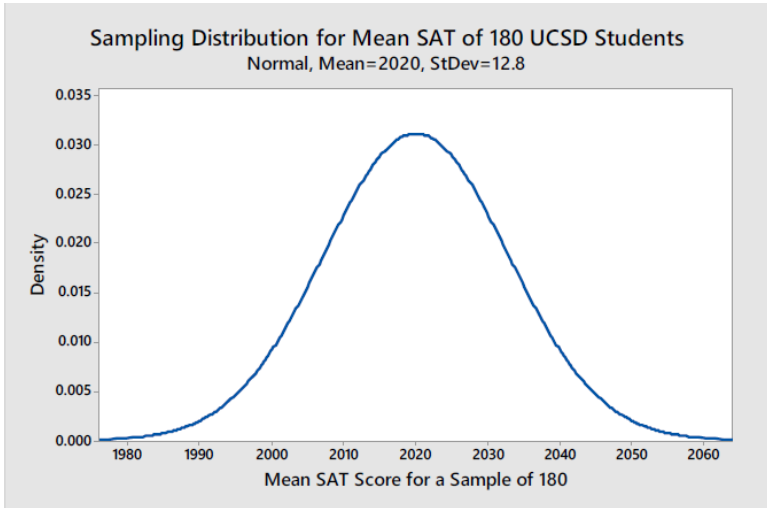
Wait A Second!

Does it matter what the population distribution looks like?

Wait A Second!

Does it matter what the population distribution looks like?





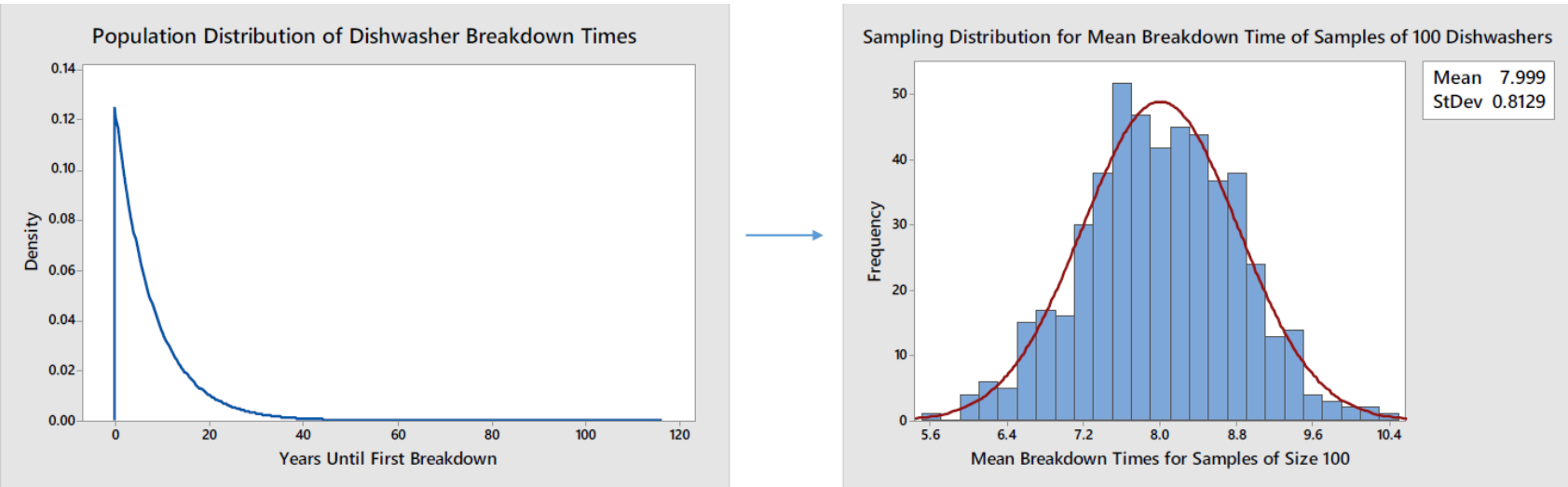
Perhaps its not surprising that if your population is Normal, you end up with a Normal sampling distribution

What if we start with a very skewed idea?

Assume the amount of time before a dishwasher breaks down is modelled by $Exp(\lambda = 1/8)$.
If we take samples of 100 dishwashers and calculate the mean time for the first breakdown, what does the sampling distribution look like?

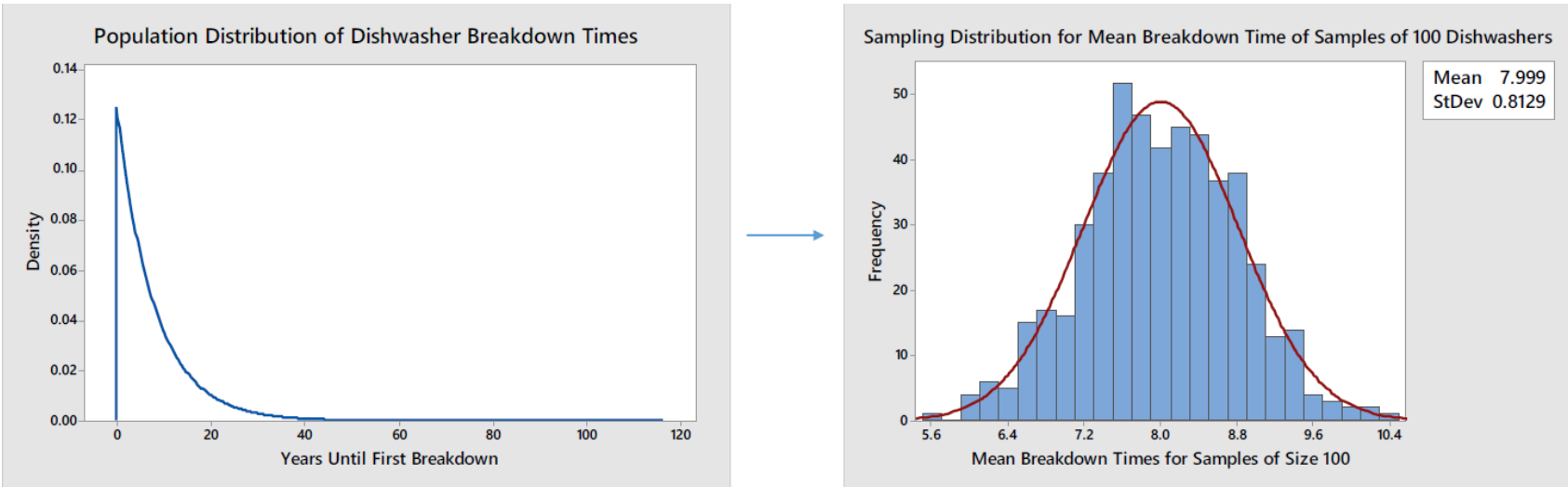
What if we start with a very skewed idea?

Assume the amount of time before a dishwasher breaks down is modelled by $Exp(\lambda = 1/8)$.
If we take samples of 100 dishwashers and calculate the mean time for the first breakdown, what does the sampling distribution look like?



What if we start with a very skewed idea?

Assume the amount of time before a dishwasher breaks down is modelled by $Exp(\lambda = 1/8)$.
If we take samples of 100 dishwashers and calculate the mean time for the first breakdown, what does the sampling distribution look like?



Note: $\mu = 1/\lambda = 8$ years, and $\sigma = 1/\lambda = 8$ years.

Hence, theory predicts:

Center: $\mu = 8$ Spread: $SE = \sigma/\sqrt{100} = 0.8$.

Do We Always Get a Normal Model for the Sampling Distribution of a Mean?

Do We Always Get a Normal Model for the Sampling Distribution of a Mean?

We do if two conditions are met:

1. **Independence Assumption**: The items in each sample (people, SAT scores, etc.) must be independent of one another.

   Typically we cannot easily determine this, so it better to check these two conditions (which effectively create independence):
   - The **Randomization Condition**: Are the items in your sample randomly chosen?
   - The **<10% Condition**: Is your sample size less than 10% of the population size?

Do We Always Get a Normal Model for the Sampling Distribution of a Mean?

We do if two conditions are met:

1. **Independence Assumption**: The items in each sample (people, SAT scores, etc.) must be independent of one another.

   Typically we cannot easily determine this, so it better to check these two conditions (which effectively create independence):
   - The **Randomization Condition**: Are the items in your sample randomly chosen?
   - The **<10% Condition**: Is your sample size less than 10% of the population size?

2. **Nearly Normal Condition (Sample Size Condition)**: The population histogram should be nearly normal (usually checked by looking at the histogram for your sample). If this histogram shows skew, you are still ok if the sample size is large (say $n > 30$ for moderate skew, and $n > 60$ for large skew)

Do We Always Get a Normal Model for the Sampling Distribution of a Mean?

We do if two conditions are met:

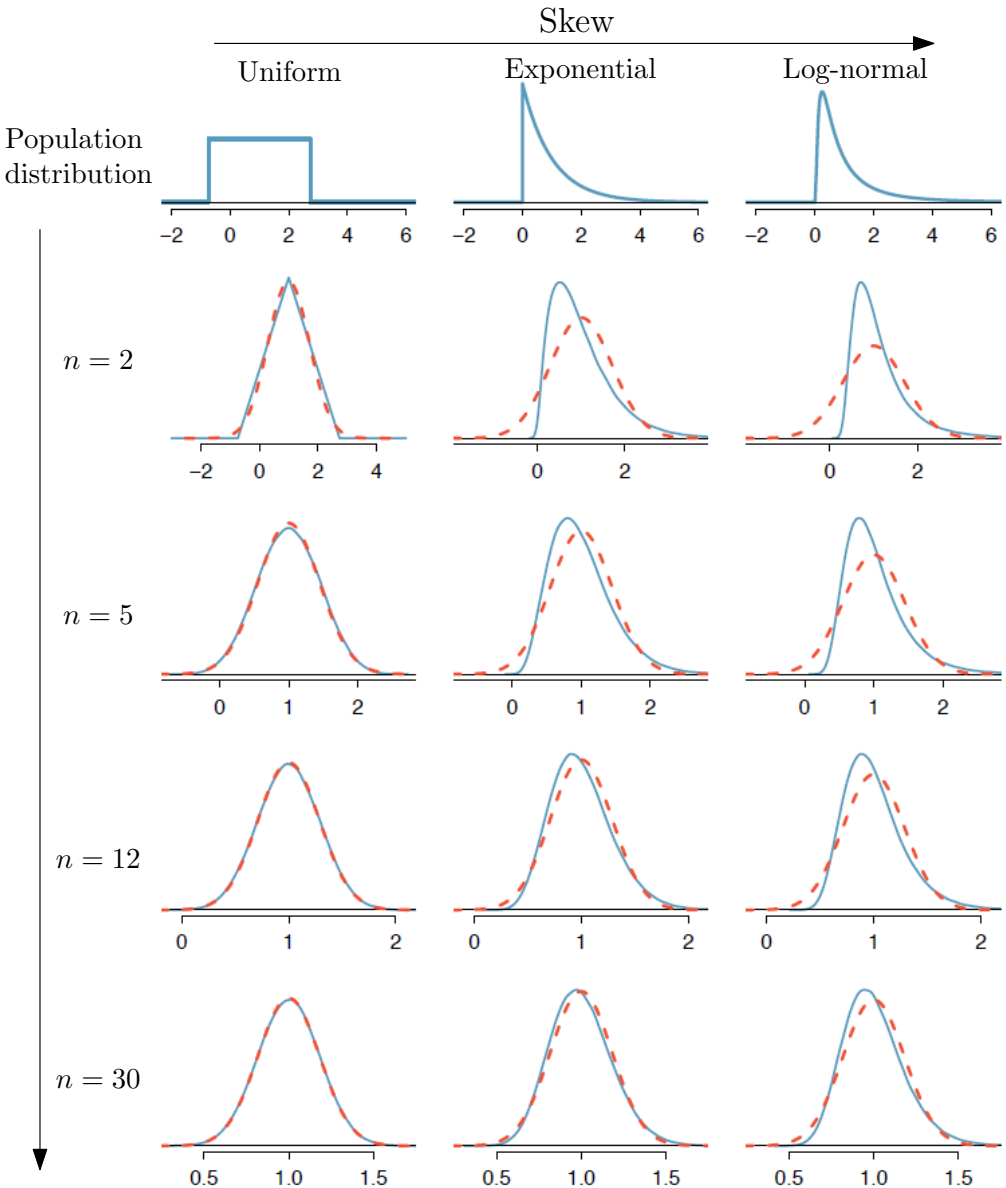
1. **Independence Assumption**: The items in each sample (people, SAT scores, etc.) must be independent of one another.

   Typically we cannot easily determine this, so it better to check these two conditions (which effectively create independence):
   - The **Randomization Condition**: Are the items in your sample randomly chosen?
   - The **<10% Condition**: Is your sample size less than 10% of the population size?

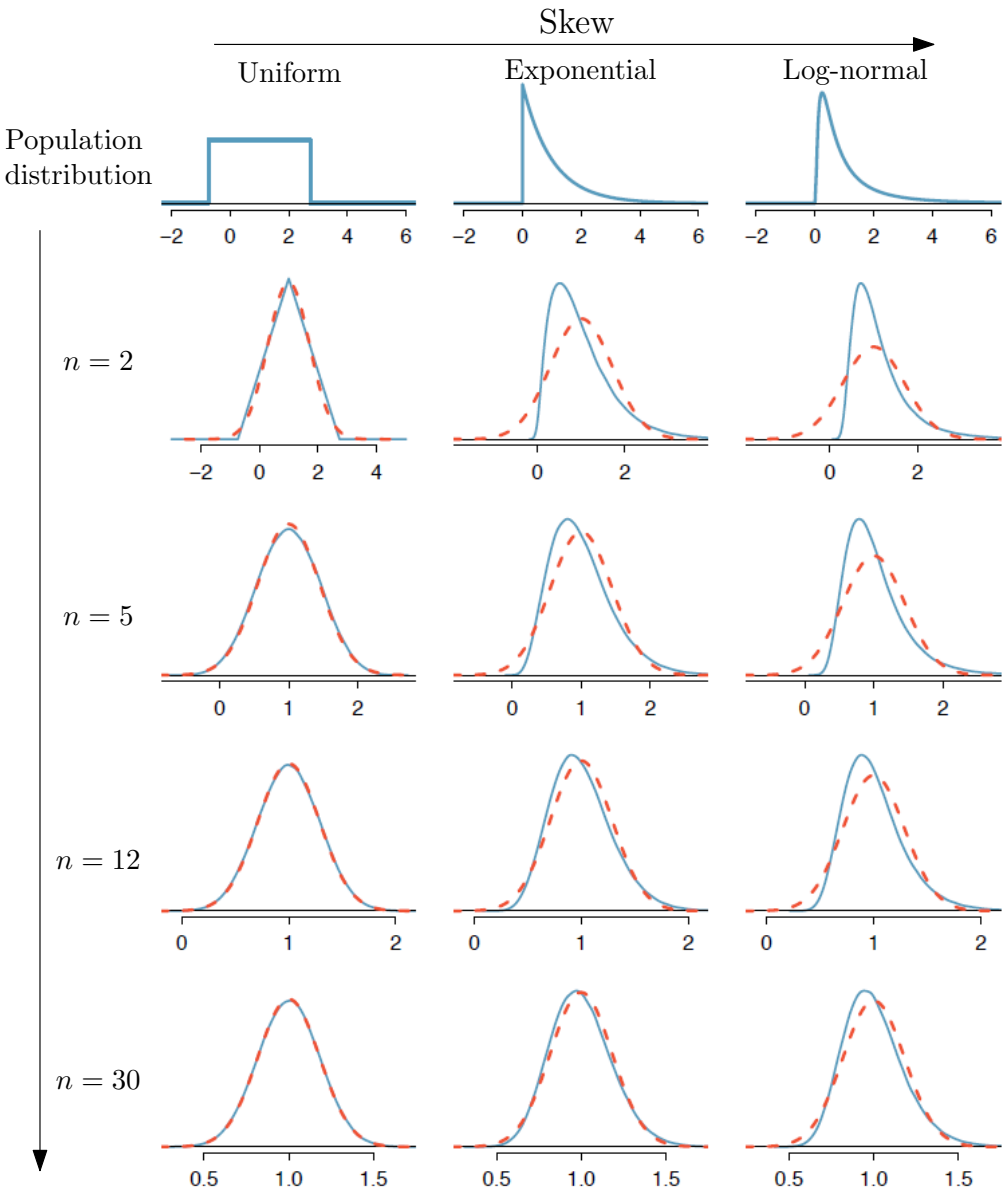
2. **Nearly Normal Condition (Sample Size Condition)**: The population histogram should be nearly normal (usually checked by looking at the histogram for your sample). If this histogram shows skew, you are still ok if the sample size is large (say $n > 30$ for moderate skew, and $n > 60$ for large skew)

Check these conditions any time you are using a Normal model to answer questions about a sampling distribution.

Central Limit Theorem



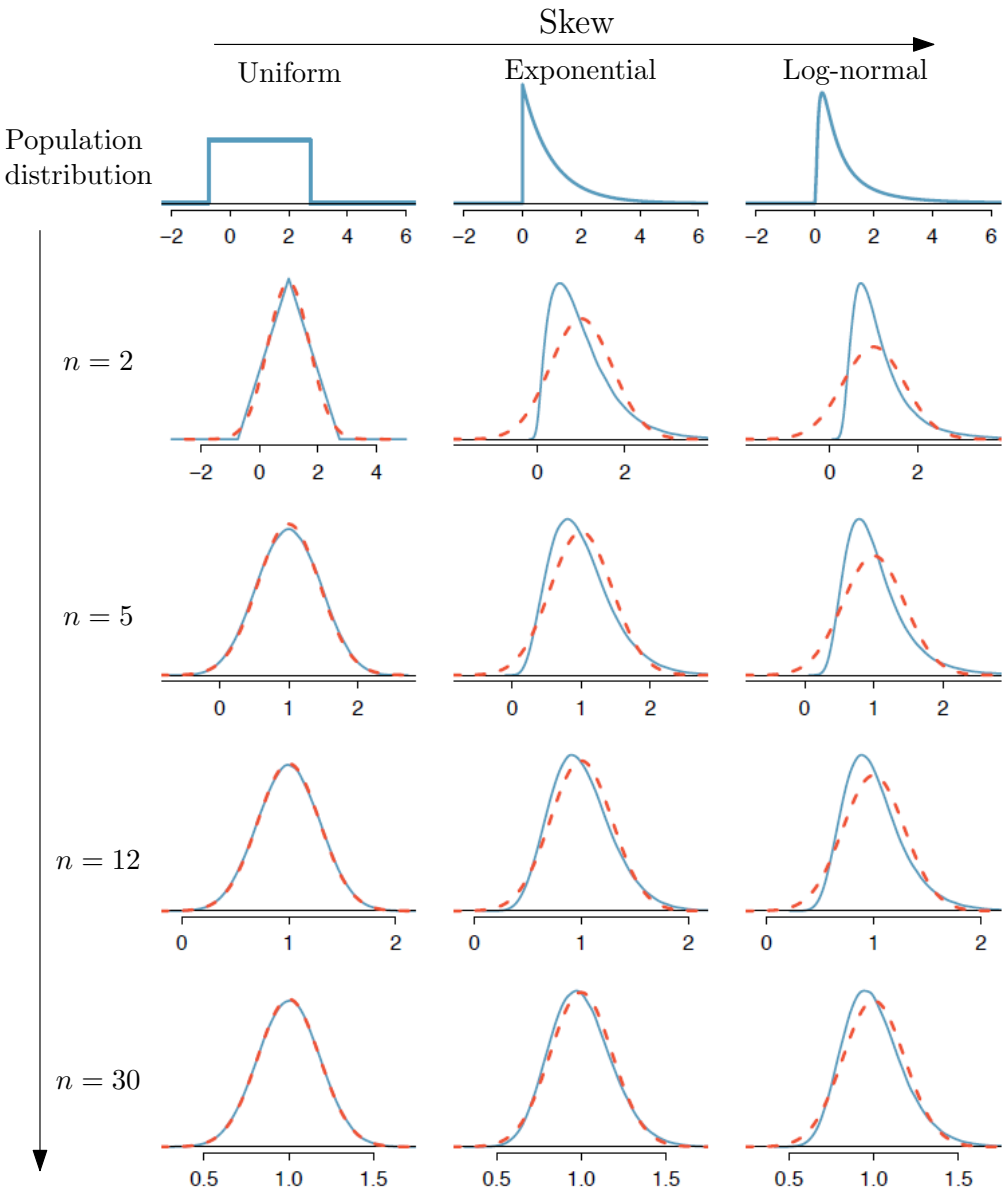
Central Limit Theorem



As the sample size grows, the sampling distribution tends to look more and more normal

The greater the skew in the population, the higher n must be to get a normal sampling distribution

Central Limit Theorem



As the sample size grows, the sampling distribution tends to look more and more normal

The greater the skew in the population, the higher n must be to get a normal sampling distribution

The **Central Limit Theorem** (CLT) proves that the sampling distribution of a proportion statistic or mean statistic will roughly be a Normal distribution regardless of the population distribution.
(assuming we have the conditions outlined on the previous slide)