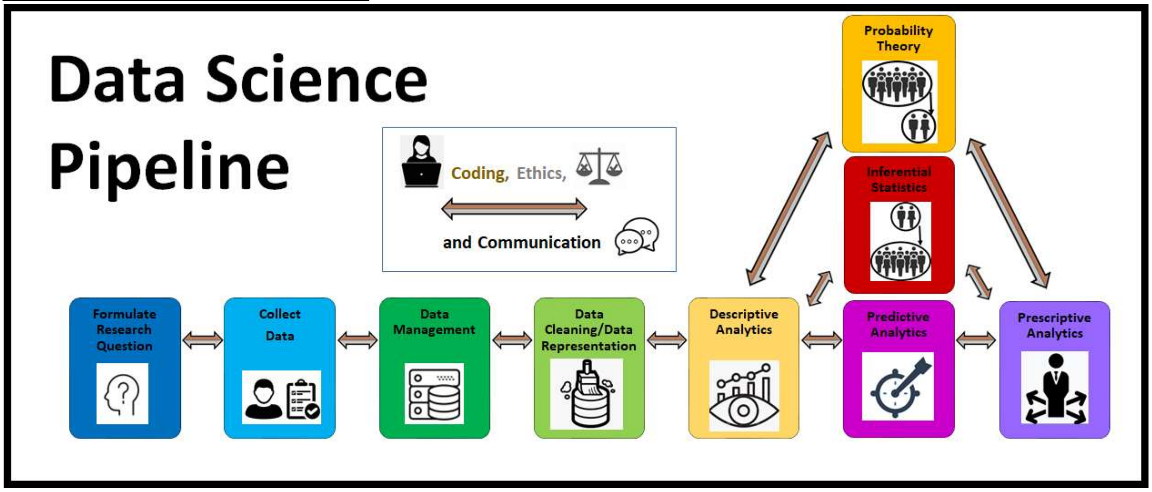
<u>Unit 9 Slides</u>: Introduction to Inference – The Central Limit Theorem and Confidence Intervals for μ and *p*

<u>Case Studies</u>:

- *Estimating a plausible range of values for the <u>average age of an adult living in the U.S (in 2017).</u>*

- *Estimating a plausible range of values for the <u>proportion of adults living in the U.S. that approve of the way things are going in the country (in 2017).</u>*

<u>Purpose of this Lecture:</u>



In this lecture we will cover the following topics.

1. <u>Two main Types of Inference for Unknown Population Parameters</u>
   1.1. Confidence Intervals
   1.2. Hypothesis Testing
2. <u>Proving Properties of Sampling Distribution of Sample Means</u>
   2.1. Proving that the <u>Mean</u> of **Sampling Distributions** ≈$E[\bar{X}] = \mu$
   2.2. Proving that the <u>Standard Deviation</u> of **Sampling Distributions** ≈SD$[\bar{X}] = \frac{\sigma}{\sqrt{n}}$
   2.3. <u>Central Limit Theorem</u>: Proving that the <u>Shape</u> of the **Sampling Distribution** of sample means is approximately normal *under certain conditions.*
3. <u>Confidence Intervals</u>
4. <u>Confidence Interval for a Population Mean μ</u>
   4.1. General framework
   4.2. What to do when you don't know σ?

4.3. What does “95% confident” mean?

5. Binomial Random Variables
6. Proving Properties of Sampling Distribution of Sample Means
   6.1. Proving $E[\hat{p}] = p$ and $SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$.
   6.2. Central Limit Theorem – for Sample Proportions: Proving that the Shape of the **Sampling Distribution of sample proportions** is approximately normal *under certain conditions.*
7. Confidence Interval for a Population Proportion p
   7.1. General framework
   7.2. What do you do when you need to plug in a ‘p’ in the conditions and confidence interval equation, but you don’t know p?
   7.3. What does “99% confident” mean?

Additional resources:

Sections 4.1-4.2 and 4.3-4.5 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

1. Two Main Types of Inference for Unknown Population Parameters

Suppose we were interested in a population parameter of a large, unknown population. But all we can collect is a random sample from this population.

What we wish we could know: What is the **average age** of ALL adults living in the U.S.?

What we can figure out instead:

1. **Confidence Interval:** What is a plausible range of values for the **average** age of ALL adults living in the U.S.?
2. **Hypothesis Test:** Is there sufficient evidence to suggest some claim about the **average** age of ALL adults living in the U.S.?

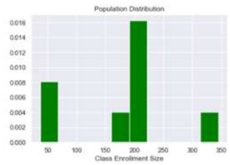
What we wish we could know: What is the **proportion** of ALL adults living in the U.S. that are **satisfied** with the way things are going in the country at the time of the survey (2017)?

What we can figure out instead:

1. **Confidence Interval:** What is a plausible range of values for **proportion** of ALL adults living in the U.S. that are **satisfied** with the way things are going in the country at the time of the survey (2017)?
2. **Hypothesis Test:** Is there sufficient evidence to suggesst some claim about the **proportion** of ALL adults living in the U.S. that are **satisfied** with the way things are going in the country at the time of the survey (2017)?

2. *Proving* Properties of Sampling Distributions of Sample Means

Population of Numerical Data



	course	section	enrolled
0	adv307	A	37
1	badm210	A	215
2	badm210	B	178
3	badm210	C	197
4	cs105	A	345
5	cs105	B	201
6	stat107	A	197
7	stat207	A	53

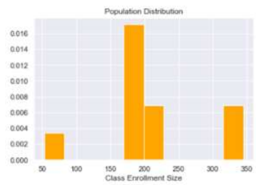
Mean of **Population** Distribution= __________ $= E[X]$

Standard Deviation of **Population** Distribution= __________ $= SD[X]$

The shape of the population distribution and the shape of the sample distribution _______________.

The standard deviation of the population and the standard deviation of the sample are approximately equal when ________________.

Sample of Numerical Data



Random Sample of n=10 Course Enrollments
197
53
345
345
215
178
197
178
215
197

Mean of **Sample** Distribution= __________

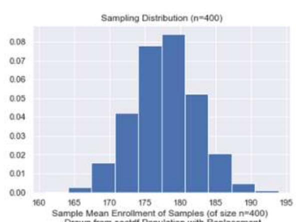
Standard Deviation of **Sample** Distribution= __________

Collect Many Random Samples (all of size n=10) drawn with replacement.

Random Sample of n=10 Course Enrollments *(drawn with replacement from population)*	Random Sample of n=10 Course Enrollments *(drawn with replacement from population)*	Random Sample of n=10 Course Enrollments *(drawn with replacement from population)*	...	Random Sample of n=10 Course Enrollments *(drawn with replacement from population)*
197	215	53	...	215
37	215	53	...	197
345	53	197	...	37
201	201	53	...	215
178	53	197	...	197
37	345	197	...	197
53	197	178	...	345
201	37	215	...	345
201	201	197	...	197
197	37	197	...	201

$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$

Sampling Distribution of **Sample Means**



Sample Means
164.7
155.4
153.7
...
214.6

Need to know the following to make an inference about *Unknown* Population Mean μ:

Mean of **Sampling** Distribution $\approx E[\bar{X}] =$ _______

Standard deviation of **Sampling** Distribution $\approx SD[\bar{X}] =$ _______

Shape of **Sampling** Distribution is ___________________ when either:

1. __________________________ OR
2. ____.

2.1. Proving that the Mean of Sampling Distributions $\approx E[\bar{X}] = \mu$

Suppose we were to collect a random sample of *n* numerical values $X_1, X_2, \dots, X_n$ (with replacement) from a population with mean μ and standard deviation σ.

What we know:

- Because $X_1, X_2, \dots, X_n$ are random variables, then $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ is a random variable.
- $E[X_i] = \mu$ for each i=1,2,…,n
- **Property of of E[ ]:**
    - If X and Y are random variables, and a and b are coefficients, then

$$E[aX + bY] = aE[X] + b[E[Y]$$

Proofs:

1. **By definition of E[ ]**, if we were to collect many, many random samples means (from samples of size n) then we would expect the **mean** of all these sample means to be $E[\bar{X}]$.

2. $E[\bar{X}] = \mu$ **(prove this in your lab assignment!)**

2.2. Proving that the Standard Deviation of Sampling Distributions (ie. the Standard Error) $\approx SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

Definition: We call the **standard error** the standard deviation of a sampling distribution.

Suppose we were to collect a random sample of *n* numerical values $X_1, X_2, \dots, X_n$ (with replacement) from a population with mean μ and standard deviation σ.

What we know:

- Because $X_1, X_2, \dots, X_n$ are random variables, then $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ is a random variable.
- $V[X_i] = \sigma^2$ for each i=1,2,…,n
- $SD[X_i] = \sigma$ for each i=1,2,…,n
- **Property of of E[ ]:**
    - If X and Y are random variables, and a and b are coefficients, then

$$V[aX + bY] = a^2V[X] + b^2V[Y]$$

Proofs:

1. **By definition of SD[ ],** if we were to collect many, many random samples means (from samples of size n) then we would expect the **standard deviation** of all these sample means to be $SD[\bar{X}]$.
2. $V[\bar{X}] =$ ______________________________

$$= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$= \frac{\sigma^2}{n}$$

3. $SD[\bar{X}] = \sqrt{V[\bar{X}]}$

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

2.3. Proving that the sampling distribution of sample means is approximately normal if either:

1. _________________________ OR

2. _________________________.

Central Limit Theorem

- Suppose we were to collect a sample of *n* numerical values $X_1, X_2, \ldots, X_n$ that are ________________ from a population with mean μ and standard deviation σ.
- Then, $\overline{\boldsymbol{X}} = \frac{X_1+X_2+\cdots+X_n}{n}$ is an approximately normal random variable when either:
  - ____________________________ or
  - ____________________________.

What counts as large enough?

How can $X_1, X_2, \ldots, X_n$ be independent in a sample?

1. ____

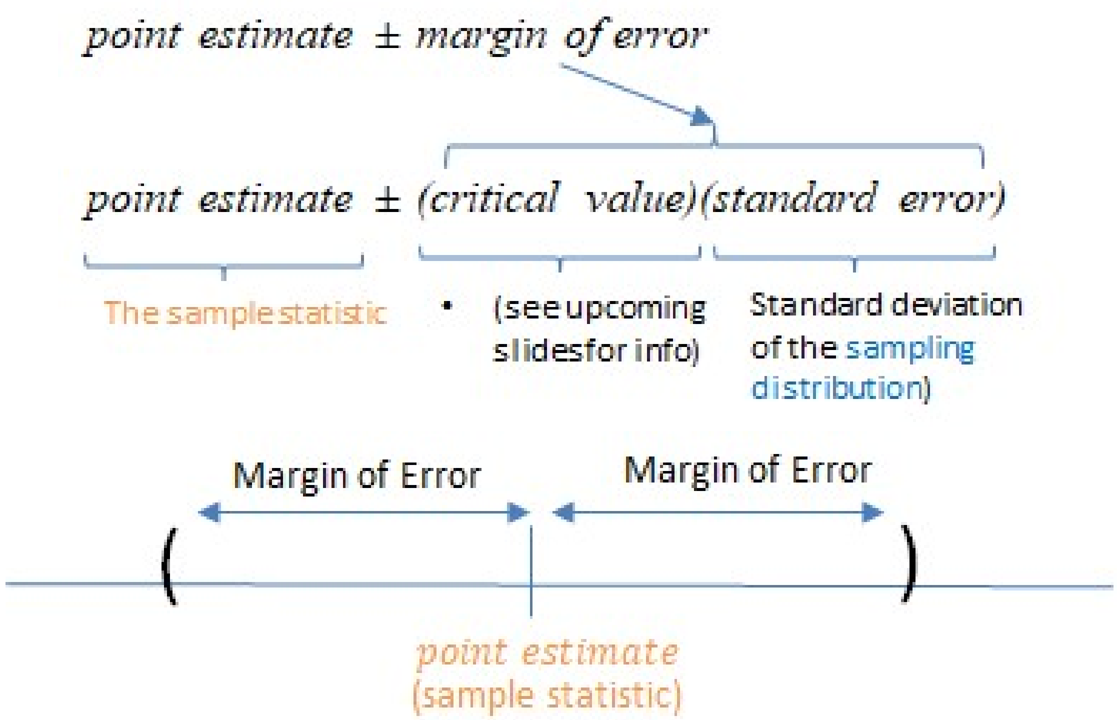
2. ____

3. Confidence Intervals

Definition: A **confidence interval** is one method of inference that gives a ________________________ for a _______________________.

Parameters: A confidence interval always corresponds to a "$(1-\alpha)\cdot 100\%$ confidence level".

Calculation: It is calculated as follows with the following components:



When you can use it: When the sampling distribution is approximately **normal**, (ie. the Central Limit Theorem Conditions hold).

Interpretation: We are "$(1-\alpha)\cdot 100\%$ confident" that the population parameter is within this confidence interval range.

What does “XX% confident” mean in the interpretation of a confidence interval?

Suppose we have calculated a XX% confidence interval for a ______________________________ using a random sample of size __________.

Now suppose we do the following:

- Collect many, many random samples, each with a sample size of ____________.
- For each random sample we calculate _____________________.
- And then we construct a confidence interval centered around each ________________________.

When we say “XX% confident” in our original confidence interval interpretation, we are saying that we would expect ____________ of the confidence intervals that we would create, using the method described above would ______________________________.

4. Confidence Interval for a Population Mean μ

4.1. Confidence Interval for a Population Mean μ - General Framework

<u>Definition</u>: A **confidence interval for a population mean μ** is one method of inference that gives a ________________________ for a _______________________.

A **confidence interval for a population mean μ** can be calculated as follows:

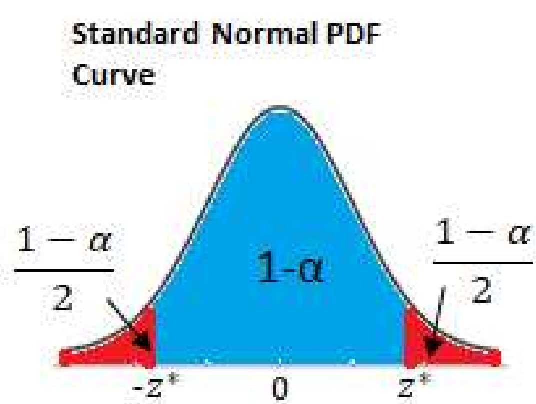
$$\textit{point estimate} \pm (\textit{critical value})(\textit{standard error})$$

$$\bar{x} \qquad \pm \qquad z^* \qquad \frac{\sigma}{\sqrt{n}}$$

A special kind of z-score

<u>Critical Value</u>:

<u>What is</u> the critical value z^* for $(1-\alpha)\cdot 100\%$ Confidence Interval?
The POSITIVE z-score in Standard Normal Distribution (z-tables) that creates this segmentation of area underneath of the standard normal pdf curve.



<u>When you can use it</u>: When the sampling distribution of **sample means** is approximately **normal**, (ie. the Central Limit Theorem Conditions hold):

- ____________________________________ AND
- ____________________________________ AND
- ____.

<u>Interpretation</u>: We are "$(1-\alpha)\cdot 100\%$ confident" that the **population mean μ** is within this confidence interval range.

__Ex__: Suppose we wanted to calculate a 95% confidence interval (ie. range of plausible values) for μ (the average age of ALL adults living in the U.S.). We have a random sample of size n=1489 that has a mean age of 50.49 years and a standard deviation of 17.84 years. **Suppose we also know that the standard deviation of ALL adults living in the U.S. is $\sigma = 18$.**

1. Learn more about the dataset, read it, and clean it

   See Unit 9 notebook section 4.1 for code involved in this problem.

2. Collect information from the problem.
   See Unit 9 notebook section 4.1 for code involved in this problem.

3. Are we allowed to calculate a confidence interval for μ using this sample that was collected and the equations we just learned? Why or why not?
   See Unit 9 notebook section 4.1 for code involved in this problem.

4. What if one of the conditions above was not met and we calculated our confidence interval anyway using the given equations?
   See Unit 9 notebook section 4.1 for code involved in this problem

5. What is the critical value for this 95% confidence interval?
   See Unit 9 notebook section 4.1 for code involved in this problem

6. Calculate the 95% confidence interval.
   See Unit 9 notebook section 4.1 for code involved in this problem

7. Interpret this 95% confidence interval.
   See Unit 9 notebook section 4.1 for code involved in this problem

Where it comes from:

If we assume that the Central Limit Theorem conditions are met, then we know that:

$$\bar{X} \sim N(mean = E[\bar{X}] = \mu, standard\ deviation = SD[\bar{X}] = \frac{\sigma}{\sqrt{n}})$$

Thus, $Z = \frac{\bar{X}-E[\bar{X}]}{SD[\bar{X}]} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is a standard normal random variable.

$$P\left(-Z_{0.975} \leq \frac{\bar{X} - E[\bar{X}]}{SD[\bar{X}]} \leq Z_{0.975}\right) = 0.95$$

$$P\left(-Z_{0.975} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{0.975}\right) = 0.95$$

$$P\left(\bar{X} - Z_{0.975}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

4.2. What to do if you don't know σ?

If we don't know μ, it is quite often the case that we don't know σ either.

Useful property:

As n increases, $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}$ gets closer to σ.

<u>"Rule of Thumb":</u> How large does n have to be to approximate $s \approx \sigma$?

<u>Ex</u>: Suppose we wanted to calculate a 90% confidence interval (ie. range of plausible values) for μ (the average age of ALL adults living in the U.S.). We have a random sample of size n=1489 that has a mean age of 50.49 years and a standard deviation of 17.84 years. **Suppose we didn't know what the population standard deviation was.**

<u>See Unit 9 notebook section 4.2 for code involved in this problem</u>

4.3. What does “95% confident” mean?

Ex: We are 95% confident that μ the average age of adults living in the U.S. is between 49.303 and 51.674.

Suppose we do the following:

- Collect many, many random samples, each with a sample size of ____________.
- For each random sample we calculate _____________________.
- And then we construct a confidence interval centered around each

  ________________________.

When we say “_____% confident” in our original confidence interval interpretation, we are saying that we would expect ____________ of the confidence intervals that we would create, using the method described above would ____.

$$P\left(-Z_{0.975} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{0.975}\right) = 0.95$$

$$P\left(\mu - Z_{0.975}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + Z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

5. Binomial Random Variables

***Binomial Random Variable*:**

Definition: A **binomial random variable X** = number of independent trials (out of **n**) that are a success. We assume that the probability of success of any given trial is **p.**

Short-Hand: ______________________

Probability Mass Function: Y is a **Bernoulli random variable** if and only if

$$\boldsymbol{p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}}$$

Mean, Variance, and Standard Deviation

$\boldsymbol{E[\overline{X}] = np}$

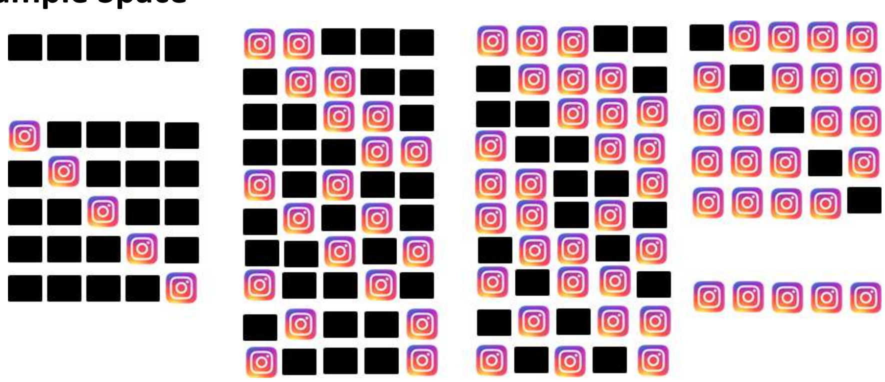
$\boldsymbol{V[\overline{X}] = np(1 - p)}$

$\boldsymbol{SD[\overline{X}] = \sqrt{np(1 - p)}}$

Example:

About 35% of American adults use Instagram. We decide to collect a random sample of 5 American adults and ask if they use Instagram or not.

Sample Space



Let X = # of randomly selected American adults (out of 5) that are Instagram users.

1. Is X a binomial random variable? If so, why?

2. What is the probability that we select a random sample with 2 people that use Instagram (using the binomial random variable probability mass function)?

3. What is the probability that we randomly select two people (the second and third) that use Instagram and the others (the first, fourth, and fifth) that do not?

4. What is the probability that we randomly select two people (the first and fifth) that use Instagram and the others (the second, third, and fourth) that do not?

5. How many possible ways is there to choose <u>two</u> **positions** out of **<u>five</u> positions in the sample** to be Instagram users?

6. What is the probability that we select a random sample with 2 people that use Instagram (using basic probability rules)?

7. What is the probability that we select a random sample with 2 people that use Instagram (using Python)?

   See Unit 9 notebook section 5 for code involved in this problem

8. What is the probability that we select a random sample with more than 2 people that use Instagram (using Python)?

   See Unit 9 notebook section 5 for code involved in this problem

6. *Proving* Properties of Sampling Distributions of Sample Proportions

Population of Categorical Data

	toss	value
0	heads	1
1	tails	0

Population Proportion: **p**

Sample of Categorical Data

```
Random Sample
0    1
1    0
1    0
0    1
0    1
1    0
1    0
1    0
0    1
0    1
```

Sample Proportion: $\hat{p}$

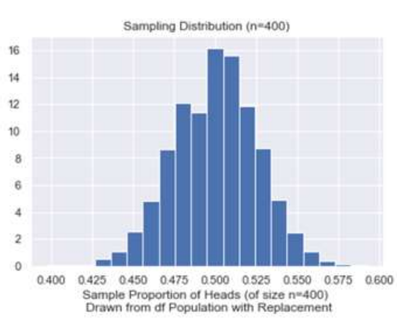
Collect Many Random Samples (all of size n=10) drawn with replacement.

Random Sample of n=10 Tosses *(drawn with replacement from population)*	Random Sample of n=10 Tosses *(drawn with replacement from population)*	Random Sample of n=10 Tosses *(drawn with replacement from population)*	...	Random Sample of n=10 Tosses *(drawn with replacement from population)*
1	1	1	...	0
0	1	0	...	0
1	1	1	...	0
0	1	0	...	0
1	1	0	...	1
0	1	0	...	0
0	1	1	...	0
1	1	1	...	0
0	1	0	...	1
1	0	0	...	1

$\hat{p}$ $\quad$ $\hat{p}$ $\quad$ $\hat{p}$ $\quad$ $\hat{p}$

Sampling Distribution of Sample Proportions

Sample Proportions
0.5
0.9
0.4
...
0.3



Need to know the following to make an inference about *Unknown* Population Proportion *p*:

Mean of **Sampling** Distribution $\approx E[\hat{p}] =$ ________

Standard deviation of **Sampling** Distribution $\approx SD[\hat{p}] =$ ________

Shape of **Sampling** Distribution is ____________________ when:

____.

6.1. Proving $E[\hat{p}] = p$ and $SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$

***Binomial Proportion Random Variable*:**

Definition: A **binomial random variable** $\hat{p}$ = **proportion** of independent trials (out of **n**) that are a success. We assume that the probability of success of any given trial is **p.**

Relationships:

$\hat{p} = \frac{X}{n}$, where $X \sim Bin(n, p)$ (ie. X is a binomial random variable with parameters n and p)

Mean, Variance, and Standard Deviation

$$E[\hat{p}] = p$$

$$V[\hat{p}] = \frac{p(1-p)}{n}$$

$$SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$$

6.2. Proving that the distribution of sample proportions is approximately normal if ___________________________________.

Central Limit Theorem – for Sample Proportions

- Suppose we were to collect a sample of *n* **Bernoulli random variable values** $X_1, X_2, \dots, X_n$ that are _________________ where $X_i \sim Bernoulli(p)$.
- Then, $\hat{p} = \frac{X_1+X_2+\cdots+X_n}{n}$ is an approximately normal random variable when:

____

“Rule of Thumb”: What counts as large enough?

How can $X_1, X_2, \dots, X_n$ **be independent in a sample?**

1. ____

2. ____

7. Confidence Interval for a Population Proportion *p*

7.1. Confidence Interval for a Population Proportion *p* - General Framework

Definition: A **confidence interval for a population proportion p** is one method of inference that gives a plausible range of values for the population proportion p.

A **confidence interval for a population proportion *p*** can be calculated as follows:

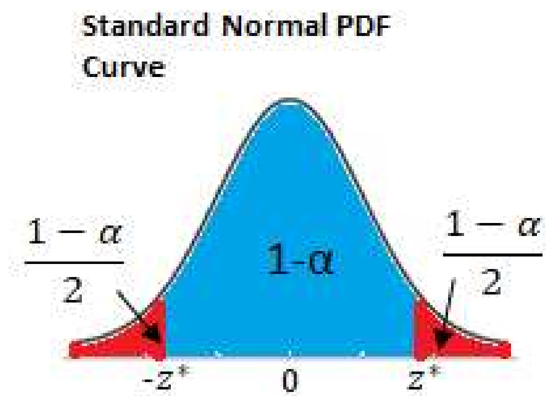
point estimate ± *(critical value)(standard error)*

$$\hat{p} \quad \pm \quad z^* \quad \sqrt{\frac{p(1-p)}{n}}$$

A special kind of z-score

Critical Value:

What is the critical value z^* for $(1-\alpha)\cdot 100\%$ Confidence Interval?
The POSITIVE z-score in Standard Normal Distribution (z-tables) that creates this segmentation of area underneath of the standard normal pdf curve.



When you can use it: When the sampling distribution of **sample proportions** is approximately **normal**, (ie. the Central Limit Theorem Conditions hold):

- ______________________________________ AND
- ______________________________________ AND
- ____.

Interpretation: We are "$(1-\alpha)\cdot 100\%$ confident" that the **population proportion p** is within this confidence interval range.

7.2. What do you do when you need to plug in a ‘p’ in the conditions and confidence interval equation, but you don’t know p?

<u>Problem</u>: We’re making an inference about p, so we can’t assume to know p to plug into the following places.

<u>When you can use it</u>: When the sampling distribution of **sample proportions** is approximately **normal**, (ie. the Central Limit Theorem Conditions hold):

- ____________________________________ AND
- ____________________________________ AND
- ____.

A **confidence interval for a population proportion *p*** can be calculated as follows:

$$\text{point estimate} \pm (\text{critical value})(\text{standard error})$$

$$\hat{p} \quad \pm \quad z^* \quad \sqrt{\frac{p(1-p)}{n}}$$

A special kind of z-score

<u>“Work-around”</u>: When creating a confidence interval, substitute $\hat{p}$ for p in these places.

<u>Ex</u>: Suppose we wanted to calculate a 99% confidence interval (ie. range of plausible values) for p, the proportion of all adults living in the U.S. that are satisfied with the way things are going in the country at the time of the survey (2017). We collect a sample of size n=1435 that has a sample proportion of 0.301.

1. Learn more about the dataset, read it, and clean it

   <u>See Unit 9 notebook section 7.2 for code involved in this problem.</u>

2. Collect information from the problem.
   <u>See Unit 9 notebook section 7.2 for code involved in this problem.</u>

3. Are we allowed to calculate a confidence interval for p using this sample that was collected and the equations we just learned? Why or why not?
   <u>See Unit 9 notebook section 7.2 for code involved in this problem.</u>

4. What is the critical value for this 99% confidence interval?
   <u>See Unit 9 notebook section 7.2 for code involved in this problem.</u>

5. Calculate the 99% confidence interval.
   See Unit 9 notebook section 7.2 for code involved in this problem.

6. Interpret this 99% confidence interval.
   See Unit 9 notebook section 7.2 for code involved in this problem.

Where it comes from:

If we assume that the Central Limit Theorem conditions are met, then we know that:

$$\hat{P} \sim N(mean = E[\hat{p}] = p, standard\ deviation = SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}})$$

Thus, $Z = \frac{\hat{p}-E[\hat{p}]}{SD[\hat{p}]} = \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}}$ is a standard normal random variable.

$$P\left(-Z_{0.975} \leq \frac{\hat{p}-E[\hat{p}]}{SD[\hat{p}]} \leq Z_{0.975}\right) = 0.99$$

$$P\left(-Z_{0.975} \leq \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq Z_{0.975}\right) = 0.99$$

$$P\left(\hat{P} - Z_{0.975}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + Z_{0.975}\sqrt{\frac{p(1-p)}{n}}\right) = 0.99$$

7.3. What does “99% confident” mean?

Ex: We are 95% confident that ____

Suppose we do the following:

- Collect many, many random samples, each with a sample size of ____________.
- For each random sample we calculate _____________________.
- And then we construct a confidence interval centered around each

  ________________________.

When we say “_____% confident” in our original confidence interval interpretation, we are saying that we would expect ____________ of the confidence intervals that we would create, using the method described above would ____.