CLT - Conditions

Certain conditions must be met for the CLT to apply:

1. **Independence:** Sampled observations must be independent.
   
   This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - $n < 10\%$ of the population.

2. **Sample size/skew:** the population distribution must be nearly normal
   or $n > 30$ and the population distribution is not extremely skewed.
   
   This is also difficult to verify for the population, but we can check it
   using the sample data, and assume that the sample mirrors the population.
From last time

Example - Review

To the right is a plot of a population distribution. Match each of the following descriptions to one of the three plots below.

(1) a single random sample of 100 observations from this population
(2) a distribution of 100 sample means from random samples with size 7
(3) a distribution of 100 sample means from random samples with size 49
A plausible range of values for the population parameter is called a confidence interval.

Using only a point estimate to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we are more likely to miss. If we toss a net in that area, we have a better chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.
Example - Relationships

A sample of 50 Duke students were asked how many long term exclusive relationships they have had. The sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

The approximate 95% confidence interval is defined as

\[
\text{point estimate} \pm 2 \times SE
\]

\[
\bar{x} = 3.2 \quad s = 1.74 \quad SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25
\]

\[
\bar{x} \pm 2 \times SE = 3.2 \pm 2 \times 0.25
\]

\[
= (3.2 - 0.5, 3.2 + 0.5)
\]

\[
= (3.15, 3.25)
\]

We are 95% confident that Duke students on average have been in between 3.15 and 3.25 exclusive relationships.
What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation \( \text{point estimate} \pm 2 \times SE \).
- Then about 95% of those intervals would contain the true population mean (\( \mu \)).

- The figure on the left shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.

- It **does not** mean there is a 95% probability the CI contains the true value.
A more accurate interval

Confidence interval, a general formula

\[ \text{point estimate} \pm Z^* \times SE \]

Conditions when the point estimate $= \bar{x}$:

1. **Independence**: Observations in the sample must be independent
   - random sample/assignment
   - $n < 10\%$ of population

2. **Sample size / skew**: $n \geq 30$ and distribution not extremely skewed

*Note*: We’ll talk about what happens when $n < 30$ next week.
Changing the confidence level

\[ \text{point estimate} \pm Z^* \times SE \]

- In order to change the confidence level all we need to do is adjust \( Z^* \) in the above formula.
- Commonly used confidence levels in practice are 90\%, 95\%, 98\%, and 99\%.
- For a 95\% confidence interval, \( Z^* = 1.96 \).
- Using the Z table it is possible to find the appropriate \( Z^* \) for any confidence level.
Example - Calculating $Z^*$

What is the appropriate value for $Z^*$ when calculating a 98% confidence interval?
Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?
Example - Sample Size

Coca-Cola wants to estimate the per capita number of Coke products consumed each year in the United States, in order to properly forecast market demands they need their margin of error to be 5 items at the 95% confidence level. From previous years they know that $\sigma \approx 30$. How many people should they survey to achieve the desired accuracy? What if the requirement was at the 99% confidence level?
Common Misconceptions

1. The confidence level of a confidence interval is the probability that the interval contains the true population parameter.

   *This is incorrect, CIs are part of the frequentist paradigm and as such the population parameter is fixed but unknown. Consequently, the probability any given CI contains the true value must be 0 or 1 (it does or does not).*

2. A narrower confidence interval is always better.

   *This is incorrect since the width is a function of both the confidence level and the standard error.*

3. A wider interval means less confidence.

   *This is incorrect since it is possible to make very precise statements with very little confidence.*
We start with a *null hypothesis* ($H_0$) that represents the status quo.

We develop an *alternative hypothesis* ($H_A$) that represents our research question (what we’re testing for).

We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.

If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We’ll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.
In 2001 the average GPA of students at Duke University was 3.37. Last semester Duke students in a Stats class were surveyed and asked for their current GPA. This survey had 147 respondents and yielded an average GPA of 3.56 with a standard deviation of 0.31.

Assuming that this sample is random and representative of all Duke students, do these data provide convincing evidence that the average GPA of Duke students has \textit{changed} over the last decade?
Setting the hypotheses

- The *parameter of interest* is the average GPA of current Duke students.

- There may be two explanations why our sample mean is higher than the average GPA from 2001.
  - The true population mean has changed.
  - The true population mean remained at 3.37, the difference between the true population mean and the sample mean is simply due to natural sampling variability.

- We start with the assumption that nothing has changed.

\[ H_0 : \mu = 3.37 \]

- We test the claim that average GPA has changed.

\[ H_A : \mu \neq 3.37 \]
Making a decision - p-values

We would know like to make a decision about whether we think $H_0$ or $H_A$ is correct, to do this in a principled / quantitative way we calculate what is known as a *p-value*.

- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is *low* (lower than the significance level, $\alpha$, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* $H_0$.
- If the p-value is *high* (higher than $\alpha$) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* $H_0$.
- We never accept $H_0$ since we’re not in the business of trying to prove it. We simply want to know if the data provide convincing evidence against $H_0$. 
In order to perform inference using this data set, we need to use the CLT and therefore we must make sure that the necessary conditions are satisfied:

1. Independence:
   - We have already assume this sample is random.
   - $147 < 10\%$ of all current Duke students.
   \[ \Rightarrow \] we are safe assuming that GPA of one student in this sample is independent of another.

2. Sample size / skew: The distribution appears to be slightly skewed (but not extremely) and $n > 30$ so we can assume that the distribution of the sample means is nearly normal.
Calculating the p-value

**p-value:** probability of observing data at least as favorable to $H_A$ as our current data set (a sample mean greater than 3.56 or less than 3.18), if in fact $H_0$ was true (the true population mean was 3.37).

\[
P(\bar{x} > 3.56 \text{ or } \bar{x} < 3.18 \mid \mu = 3.37) \\
= P(\bar{x} > 3.56 \mid \mu = 3.37) + P(\bar{x} < 3.18 \mid \mu = 3.37) \\
= P\left(Z > \frac{3.56 - 3.37}{0.31/\sqrt{147}}\right) + P\left(Z < \frac{3.18 - 3.37}{0.31/\sqrt{147}}\right) \\
= P(Z > 7.43) + P(Z < -7.43) \\
= 10^{-13} \approx 0
\]
Drawing a Conclusion / Inference

\[ p - value \approx 10^{-13} \]

- If the true average GPA Duke students applied to is 3.37, there is approximately a 10^{-11} % chance of observing a random sample of 147 Duke students with an average GPA of 3.56.
- This is a very low probability for us to think that a sample mean of 3.56 GPA is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject \( H_0 \).
- The data provide convincing evidence that Duke students average GPA has changed since 2001.
- The difference between the null value of a 3.37 GPA and observed sample mean of 3.56 GPA is not due to chance or sampling variability.
Example - College applications

A similar survey asked how many colleges each student had applied to. 206 students responded to this question and the sample yielded an average of 9.7 college applications with a standard deviation of 7. The College Board website states that counselors recommend students apply to roughly 8 colleges. What would be the correct set of hypotheses to test if these data provide convincing evidence that the average number of colleges Duke students apply to is higher than the number recommended by the College Board. Are the conditions for inference met?

http://www.collegeboard.com/student/apply/the-application/151680.html
**College Applications - p-value**

**p-value:** probability of observing data at least as favorable to \( H_A \) as our current data set (a sample mean greater than 9.7), if in fact \( H_0 \) was true (the true population mean was 8).

\[
P(\bar{x} > 9.7 \mid \mu = 8) = P \left( Z > \frac{9.7 - 8}{7/\sqrt{206}} \right) = P(\text{ } Z > 3.4) = 0.0003
\]
Hypothesis testing

College Applications - Making a decision

\[ p - value \approx 0.0003 \]

- If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
- This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject \( H_0 \).
- The data provide convincing evidence that Duke students average apply to more than 8 schools.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is not due to chance or sampling variability.
A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 Duke students (you!) yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all Duke students, a hypothesis test was conducted to evaluate if Duke students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

What are the hypotheses being tested?
What is the correct inference for this situation?
Two-sided hypothesis test

If the research question had been “Do the data provide convincing evidence that the average amount of sleep Duke students get per night is different than the national average?”, the alternative hypothesis would be different.

\[ H_0 : \mu = 7 \]
\[ H_A : \mu \neq 7 \]

- Hence the p-value would change, as well as our decision to reject:

\[
p-value = 0.0485 \times 2 = 0.097
\]

Fail to reject \( H_0 \)!
Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a test statistic and a p-value.
4. Make a decision, and interpret it in context of the research question.
Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - \( H_0 : \mu = \text{null value} \)
   - \( H_A : \mu < \text{or} > \text{or} \neq \text{null value} \)

2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or \( n \geq 30 \), no extreme skew

3. Calculate a test statistic and a p-value (draw a picture!)
   \[
   Z = \frac{\bar{X} - \mu}{SE}, \quad \text{where} \quad SE = \frac{s}{\sqrt{n}}
   \]

4. Make a decision, and interpret it in context of the research question
   - If p-value < \( \alpha \), reject \( H_0 \), data provide strong evidence for \( H_A \)
   - If p-value > \( \alpha \), do not reject \( H_0 \)