Name: ____________________________ PUID ____________________________

STAT 350 Worksheet #14

In this worksheet, we continue our exploration of statistical inference, now addressing situations where the population standard deviation ($\boldsymbol{\sigma}$) is unknown. When σ is unknown, we must estimate it from the sample data using the sample standard deviation ($\boldsymbol{s}$). This introduces additional uncertainty into our inference. As a result, our pivotal statistic changes from a standard Normal distribution (Z-distribution) to a Student's $\boldsymbol{t}$-distribution. The Student's t-distribution, developed by William Sealy Gosset, accounts for this additional uncertainty and has heavier tails than the Normal distribution, especially for smaller sample sizes. A crucial feature of the t-distribution is its degrees of freedom ($\mathbf{df}$), which equals the sample size minus one ($\boldsymbol{n}-\mathbf{1}$) Degrees of freedom reflect how many independent pieces of information are available to estimate the variability after accounting for the estimated mean.

Before applying confidence intervals based on the Student's $\boldsymbol{t}$-distribution, it's essential to verify that the necessary assumptions are adequately satisfied. Recall from previous explorations, including your investigation of the Central Limit Theorem (CLT) with different underlying population distributions, that inference procedures rely on specific conditions being met:

1. **Random Sampling (Independence)**:
   Data must originate from a simple random sample (SRS) drawn from the population. The formula for inference using Student's ttt-distribution explicitly assumes independence among observations. If your sampling scheme involves stratification, clustering, or systematic sampling, standard formulas require modification. Poor sampling methods that introduce systematic bias or lack randomness cannot be corrected simply by statistical formulas—there are no valid inference procedures for improperly collected data.
2. **Normality or Approximate Normality**:
   The procedures assume the underlying population distribution is Normal or approximately Normal. Based on your previous explorations, recall the CLT helps justify this assumption for large samples when the population is not severely skewed or heavily tailed:
   - When the population distribution is symmetric and unimodal, the $\boldsymbol{t}$-distribution methods are robust even for moderate or smaller samples.
   - With small samples from heavily skewed, discrete, or strongly non-normal populations, the confidence interval may fail to achieve the desired coverage. Specifically, discrete data can be particularly problematic if there are few unique values or a substantial lack of symmetry.
   - Your prior graphical analyses, such as histograms and QQ-plots to guide you in assessing whether normality assumptions are reasonable. If normality is questionable, interpret results with caution.
3. **Unknown Population Standard Deviation ($\boldsymbol{\sigma}$)**:
   When the population standard deviation is unknown and must be estimated from data ($\boldsymbol{s}$), additional uncertainty is introduced. The Student's $\boldsymbol{t}$-distribution explicitly accounts for this uncertainty by widening intervals, especially for small sample sizes (due to smaller degrees of freedom, $\mathbf{df}=\boldsymbol{n}-\mathbf{1}$).
4. **Resistance to Outliers**:
   The sample mean ($\bar{\boldsymbol{x}}$) and standard deviation ($\boldsymbol{s}$) are non-resistant statistics; thus, outliers can substantially distort both your point estimate and the resulting confidence interval. Resistant methods do exist (for example, median-based procedures or robust methods) if non-removable outliers or heavily tailed distributions are evident.
5. **Random Variability versus Systematic Errors**:
   The confidence interval margin of error specifically measures only random sampling variability. It does not capture or account for systematic errors, such as bias introduced through measurement errors, flawed sampling procedures, nonresponse, or improperly defined populations. Such systematic biases must be addressed separately through improved data collection and measurement practices.

In this worksheet, we will use the Student's t-distribution to construct confidence intervals for estimating unknown population means when the population standard deviation is unknown. After completing these exercises, we will transition into the statistical inference topic of hypothesis testing. We will explore important concepts including Type I and Type II errors, statistical power, and methods for determining the optimal sample size needed to achieve a certain desired power level.

1. Consider the built-in dataset ***ChickWeight*** in R, which records the weights of chicks at different time points as they grow. Suppose researchers are interested in studying the growth patterns of chicks and how diet influences their weight over time.

   a) First, compute descriptive statistics for the chick weights at each recorded time point. For each time point, compute the sample size ($\boldsymbol{n}$), mean ($\bar{\boldsymbol{x}}$), median ($\tilde{\boldsymbol{x}}$), sample standard deviation ($\boldsymbol{s}$), and interquartile range ($\mathbf{IQR}$). Report your results for time points 0, 10, and 21 below.

Time	$\boldsymbol{n}$	$\bar{\boldsymbol{x}}$	$\tilde{\boldsymbol{x}}$	$\boldsymbol{s}$	$\mathbf{IQR}$
0					
10					
21					

b) Next, visually assess the distribution of chick weights at each time point by plotting histograms with overlaid smooth kernel density curves and fitted Normal curves. Arrange these plots in a clear grid layout, labeling each time point clearly. To further investigate the normality assumption required for the validity of Student's t confidence intervals, create QQ plots for chick weights at each time point. Describe the steps needed to display these QQ plots in a clear, grid-like format.

c) Using the histograms and QQ plots you generated, evaluate the assumption of normality at each time point. Briefly comment on the appropriateness of using Student's ***t***-confidence intervals based on your visual assessments.

d) Construct 95% confidence intervals for the mean chick weights at each time point using the confidence interval formulas. Confirm your calculations are correct by using R's built in t.test function. Report your results for time points 0 and 21 below.

Time	**Manual Calculation (95% CI)**	**R t.test() Calculation (95% CI)**
0		
21		

e) Finally, graphically illustrate your results by plotting these confidence intervals.
To visually illustrate your confidence intervals, you can use the **ggplot2** package in R. Specifically, use:

- You will need to create a dataframe that stores your summary statistics and confidence intervals end points (lower_bound, and upper_bound) and pass it to ggplot for plotting.
- For the aesthetic mapping set: x-axis to be TimePoint, and y-axis to be the mean_weight.
- geom_point() to plot the mean chick weight at each time point.
- geom_errorbar() to add vertical lines representing the confidence intervals. Set ymin to your lower bound, and ymax to your upper bound.

Clearly interpret your findings in the context of chick growth and discuss whether there appear to be time points where the confidence intervals overlap substantially. What might this indicate regarding statistically significant differences between certain time points? Later we will learn statistical inference procedures to compare across multiple groups (our time points).

We now build on our understanding of statistical inference, transitioning from confidence intervals into hypothesis testing. In previous sections, we used confidence intervals to estimate unknown parameters. Hypothesis testing is another critical inferential method in statistics, which allows us to formally test claims or statements about these unknown population parameters.

A hypothesis test involves stating two competing claims: the **null hypothesis** ($\boldsymbol{H_0}$), representing the current belief or status quo, and the **alternative hypothesis** ($\boldsymbol{H_a}$), which represents the claim we suspect might be true instead. The goal of hypothesis testing is to determine, based on sample data, whether we have sufficient evidence to reject the null hypothesis in favor of the alternative.

However, due to the inherent randomness and uncertainty present in sample data, hypothesis tests always carry the risk of making incorrect conclusions. Specifically, we must carefully consider two possible types of errors:

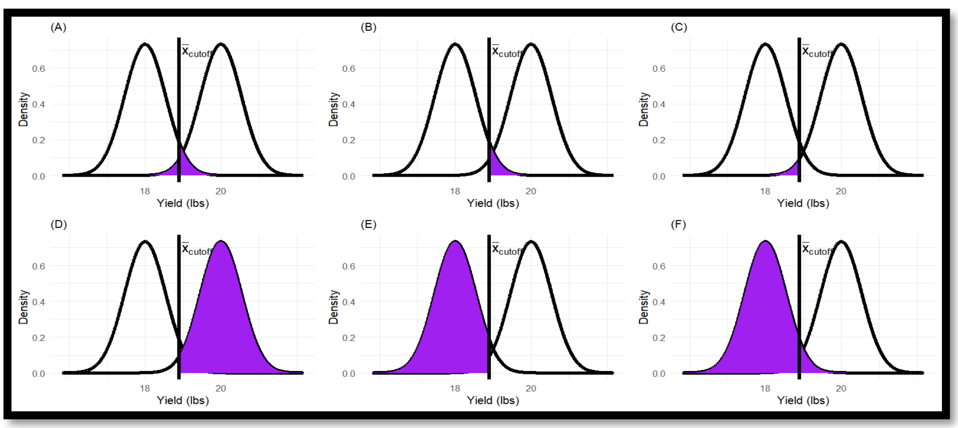
- **Type I error (False Positive):** Occurs if we reject the null hypothesis $\boldsymbol{H_0}$ when it is actually true. The probability of committing a **Type I error** is controlled by the significance level ($\boldsymbol{\alpha}$).
- **Type II error** occurs if we fail to reject the null hypothesis $\boldsymbol{H_0}$ when it is actually false. This represents a failure to detect an existing effect (false negative). The probability of committing a **Type II error** is indicated by ($\boldsymbol{\beta}$).

To quantify our ability to detect effects when they truly exist, we introduce the concept of **statistical power** $(1-\beta)$, defined as the probability that a hypothesis test correctly rejects a false null hypothesis. High power (close to 1) indicates the test is effective in detecting meaningful effects when they exist.
In practical applications, we often want to determine the optimal sample size needed to achieve a desired statistical power. Determining the sample size involves specifying the magnitude of the difference we wish to detect, the desired significance level (probability of Type I error), and the desired power (probability of correctly rejecting a false null hypothesis). We will explore these concepts, calculate power, and determine the necessary sample size through clear examples and calculations in the following exercises.

2. An agricultural researcher is studying a new fertilizer formulated to increase tomato yields. Historically, a certain variety of tomato plants grown under standard agricultural practices produces an average yield of **18 pounds per plant**. To assess the effectiveness of the new fertilizer, the researcher selects a simple random sample of tomato plants and applies the fertilizer throughout the growing season.

   The researcher wants to perform a one-sample hypothesis test to determine whether the fertilizer significantly increases the average yield per plant. An increase in average **yield of at least 2 pounds per plant** (to 20 lbs) would be considered practically meaningful. To ensure meaningful results, the researcher requires the hypothesis test to have at least **90% statistical power** to detect this yield increase, if it truly exists.
   The test is conducted at the significance level $\boldsymbol{\alpha = 0.05}$.
   $\boldsymbol{H_0: \mu \leq 18}$
   $\boldsymbol{H_a: \mu > 18}$

   Based on previous field studies, the population standard deviation for tomato yields with similar varieties is assumed to be known and equal to **4 pounds per plant**.
   The graphs shown below illustrate two probability distributions involved in hypothesis testing: one corresponding to the null hypothesis (left curve) and one corresponding to a specific alternative hypothesis (right curve). Each graph shades a different region between or under these curves.



   a) Carefully examine each plot **(A)-(F).** Your task is to clearly identify which plot represent Type I error $(\boldsymbol{\alpha})$**,** Type II error $(\boldsymbol{\beta})$, and Statistical Power $(\mathbf{1-\boldsymbol{\beta}})$. Additionally, briefly explain what the shaded regions in the remaining graphs might represent. If a graph's shaded region does not directly correspond to one of these main concepts, describe conceptually what that shaded area might illustrate in terms of hypothesis testing outcomes or probabilities.

   (A) **(C)** **(E)**

   (B) **(D)** **(F)**

b) Derive step-by-step the formula and calculate the **minimum required sample size** needed to achieve this desired statistical power level. Clearly present each step of your derivation and provide the numerical answer, rounded up to the nearest whole number.

c) Double check that your calculation for the minimum sample size is correct by computing power from scratch. Report the actual value for the power and provide the value for the cutoff $\bar{x}_{\mathbf{cutoff}}$