Name: _____ PUID _____

STAT 350 Worksheet #15

In previous lessons, we started our exploration of statistical inference by constructing confidence intervals to quantify uncertainty about the unknown population mean (μ). We then introduced the framework of hypothesis testing, an approach used to formally assess claims about unknown parameters. Specifically, we discussed types of errors that can occur in testing (Type I and Type II errors), the concept of statistical power as the ability of a test to detect an actual effect if it is present, and the tradeoffs involved.

- **Type I Error (False Positive)**: A **Type I error** is the **incorrect rejection** of a **true null hypothesis** ($\mathbf{H_0}$).
- **Type II Error (False Negative)**: A **Type II error** is the error that occurs when the **null hypothesis** is **not rejected**, even though **it is false**.
- **Statistical power : Statistical power** is the probability that a statistical test will **correctly reject** a **false null hypothesis**.

We defined the idea of a meaningful alternative hypothesis representing the specific **effect** or **difference** that a study needs to detect. We illustrated these concepts visually, exploring how changes in significance level (α), sample size (n), and variability (σ), affect power and errors. Furthermore, we calculated statistical power explicitly and learned to determine the sufficient sample size required to achieve a desired level of power for a specified alternative.

Building upon these foundational concepts, we now proceed to perform an actual **test of significance** for a single unknown population mean.
Recall that a **hypothesis** is a clearly stated mathematical claim about a population parameter. A **hypothesis test** formally assesses whether observed sample data provide sufficient evidence to support or contradict this claim. Specifically, a test evaluates two competing statements:

- **Null hypothesis** (H_0)**:** Represents the existing belief or status quo.
- **Alternative hypothesis** (H_a): Represents a claim we suspect to be true or wish to establish.

To objectively assess evidence against the null hypothesis, we rely on a carefully constructed numerical measure called a **test statistic**. This numeric measure assesses how closely our observed sample aligns with the claim made in the null hypothesis (H_0). Specifically, for a test concerning the unknown population mean (μ), when the population standard deviation (σ) is known, we use the following standardized test statistic, denoted Z_{TS}:

$$Z_{TS} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

This test statistic has several important elements:

- $\bar{X}$: This is the sample mean, serving the point estimator of the unknown population mean.
- μ_0: The **hypothesized value** under the null hypothesis, representing the status quo or baseline assumption we are testing against.
- $\sigma/\sqrt{n}$: This is the **standard deviation of the point estimate**, also known as the standard error, which quantifies the typical amount we expect our sample mean to vary from sample to sample due to randomness alone.

The computed value of the test statistic measures how far, in standard error units, our observed sample mean ($\bar{x}$) lies from the hypothesized mean (μ_0). The test statistic itself is a random variable just as it would change from sample to sample. To objectively assess the observed magnitude of the test statistic, we must understand its probability distribution under the assumption that the null hypothesis is true. If the following conditions are satisfied:

- The sample data are collected as a **simple random sample (SRS)**.
- Either the underlying population distribution is known to be approximately Normal, or the **Central Limit Theorem (CLT)** justifies normality due to sufficiently large sample size.

Then, under these conditions, the test statistic Z_{TS} follows a standard Normal distribution.

This known distribution enables us to calculate the probability (the **p-value**) of observing a test statistic at least as extreme as ours, thus providing a clear, consistent method to interpret the strength of evidence against the null hypothesis. The ***p*-value** is thus a direct measure of how strongly the data contradict or support the null hypothesis:

- If the ***p*-value** is small (less than or equal to our significance level α\alphaα), we conclude that the data provide strong evidence against the null hypothesis and thus reject H_0.
- If the ***p*-value** is larger (greater than α), we conclude that there is insufficient evidence to reject H_0

We now proceed to carefully explore each step of conducting a hypothesis test, demonstrating clearly how to interpret and communicate our statistical findings.

1. In this question, you will simulate and explore the behavior of the z**-test statistic** (Z_{TS}) and the p**-value** for testing a claim about the population mean.

a) Suppose you are conducting a hypothesis test for the population mean, where the hypotheses are stated as follows:
$H_0: \mu \leq 100$
$H_a: \mu > 100$
Assume the following conditions:

- The true population mean is exactly 100 (i.e., the null hypothesis H_0 is true).
- The population is Normal with standard deviation is known to be $\sigma = 15$.
- Your sample size is $n = 25$.
- The probability of **Type I error** is fixed at $\alpha = 0.05$.

To explore the distribution of your test statistic under these conditions, you will run a simulation in **R**. You will repeat your experiment a large number of times (**1500 repetitions**). Each repetition involves drawing a **simple random sample (SRS)** of size $n = 25$ from the specified population and computing the test statistic.

i. Obtain 1500 simple random samples from $N(\mu = 100, \sigma = 15)$ and compute 1500 sample means.

ii. **Calculate the z-Test Statistic**: For each of the 1500 simulations compute the test statistic $z_{TS} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

iii. Construct a histogram of the 1500 z_{TS}. Superimpose the smooth kernel and the theoretical standard Normal distribution density curve. Add a vertical dashed line marking the critical cutoff (rejection threshold) for a one-tailed test at significance level $\alpha = 0.05$. This is simply the critical value $z_{0.05}$ as we are plotting in the distribution of the test statistic not the distribution of $\bar{X}$.

iv. Calculate the proportion of times that the test statistic exceeds the critical value $z_{0.05}$.

Proportion (Null True scenario):

v. Compute the p**-values** for each of the 1500 test statistics and obtain a histogram of the p**-values**. Calculate and clearly report the proportion of simulated experiments where the p**-value** is less than 0.05 (i.e., the proportion of experiments that correctly reject the null hypothesis).

Proportion of p-values < 0.05 (Null True scenario):

b) Now, assume that the true population mean is $\mu = 105$, but you are still testing the same null hypothesis as above (assuming a null value $\mu_0 = 100$). Thus, the null hypothesis H_0 is false in this scenario.

i. Repeat the above steps (using again 1500 repetitions and samples of size $n = 25$) with this new true population mean of **105**.

Proportion (Null False scenario):

ii. Also, compute the p**-values** for each of the 1500 test statistics and obtain a histogram of the p**-values**. Calculate and clearly report the proportion of simulated experiments where the p**-value** is less than 0.05 (i.e., the proportion of experiments that correctly reject the null hypothesis).

Proportion of p-values < 0.05 (Null False scenario):

c) What did you observe about the distribution of test statistics under each scenario? What differences did you observe in the distribution of the test statistics under the scenario when the null hypothesis was true $(\mu = 100)$ versus when it was false $(\mu = 105)$**.** How did changing the true mean affect the proportion of test statistics exceeding the cutoff value?

d) Describe and clearly explain how the shape and pattern of simulated p**-values** differ when the null hypothesis is true (mean = 100) versus when the null hypothesis is false (mean = 105). Why do you observe these differences, and what does this illustrate about the concepts of Type I error and statistical power in hypothesis testing? Hint: calculate the power if the alternative was $\mu_a = 105$.

In real-world scenarios, the population standard deviation (σ\sigmaσ) is typically not known. As we discussed previously when exploring confidence intervals, this means we must estimate the population standard deviation from our sample data, using the **sample standard deviation** (s).

However, estimating the standard deviation introduces an additional layer of uncertainty. Even if our original data come from a population that is exactly normal, our test statistic no longer follows a standard Normal distribution. Instead, because we now rely on an estimated variability, the test statistic has greater variability and consequently follows a distribution with heavier tails, i.e., a **Student's t-distribution**.

Specifically, when performing hypothesis tests for an unknown population mean (μ), and when estimating σ by s, the correct form of the test statistic is:

$$t_{\text{TS}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

This statistic follows a Student's t-distribution with degrees of freedom $(\mathbf{df})$ equal to $n - 1$:

$$t_{\text{TS}} \sim t_{df=n-1}$$

The Student's t **-distribution** differs from the Normal distribution primarily in having heavier tails. This reflects the increased uncertainty from estimating σ:

- **Heavier tails:** The probability of observing more extreme values is higher compared to the Normal distribution, appropriately accounting for the uncertainty introduced by estimating variability.
- **Degrees of freedom (df):** The shape of the t **-distribution** depends on the degrees of freedom, defined $df = n - 1$**.** As sample size n increases, the t **-distribution** approaches a standard Normal distribution because the estimate of the standard deviation becomes more precise:

Relationship Between Confidence Intervals/Bounds and Hypothesis Tests

Hypothesis testing and confidence intervals (or confidence bounds) are closely connected concepts. Both approaches use the same underlying principles and assumptions, and they complement each other in interpreting data. Specifically:

- **A two-sided hypothesis test (testing $H_0: \mu = \mu_0$)** directly corresponds to constructing a two-sided confidence interval:

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

- For an **upper-tailed alternative** $(H_a: \mu > \mu_0)$, we construct a **lower confidence bound**:

$$\left(\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \infty\right)$$

- For a **lower-tailed alternative** $(H_a: \mu < \mu_0)$, we construct an **upper confidence bound**:

$$\left(-\infty,\ \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right)$$

Thus, confidence intervals or bounds can intuitively represent the plausible values for μ. If the null value μ_0 is not plausible (not included in the interval or lies beyond the bound), the hypothesis test naturally rejects the null hypothesis.

In practice, confidence intervals (or bounds) provide valuable context. They give a range of reasonable values for the parameter based on the data, enhancing the interpretability of hypothesis testing conclusions. The connection between these two inferential approaches is central to practical statistical reasoning and interpretation.

Next, we will illustrate these ideas clearly through a structured exercise.

2. The Environmental Protection Agency (EPA) regulates ozone concentrations due to potential harmful health effects. Historically, an 8-hour ozone concentration above 70 parts per billion (ppb) is considered unhealthy. The dataset in the base R package includes a two-hour daily average (from 1:00 PM to 3:00 PM) rather than an 8-hour average, we'll use 70 ppb as a benchmark to evaluate ozone concentrations recorded in New York during the summer of 1973.

   Conduct a hypothesis test to determine if the mean two-hour ozone concentration during the summer months exceeds the EPA threshold.

   a) Assumptions and Exploratory Analysis (in R):
      i. Clean the dataset **(airquality**) and isolate summer months (June, July, August).
      ii. Compute and clearly report the mean, standard deviation, and sample size for summer ozone concentrations.
      iii. Construct a histogram with a normal density curve and a QQ-plot to visually check the normality assumption.
      iv. Clearly interpret the plots and comment on the appropriateness of using the t-distribution for your hypothesis test.

   b) Regardless of your conclusions regarding the assumptions perform the full four step hypothesis testing procedure utilizing the $\boldsymbol{t.test}$ function in R. Use $\boldsymbol{\alpha = 0.05}$.
      i. **Step 1:** Clearly Identify the Parameter of Interest.
      ii. **Step 2:** State the Null and Alternative Hypotheses symbolically.
      iii. **Step 3:** Calculate the test statistic and $\boldsymbol{p}$**-value** and provide them below**.** Be sure to also state the degrees of freedom**.**
      iv. **Step 4**: Provide a formal conclusion using the template in the slides.

   c) Verify that the $\boldsymbol{t.test}$ agrees with the formulas by computing the test statistic manually. Compute the $\boldsymbol{p}$**-value** by utilizing the $\boldsymbol{pt}$ function in R. Show your work for computing the test statistic below and write the $\boldsymbol{p}$**-value** as a probability statement.

   d) Compute the appropriate **95% one-sided confidence bound** related to your stated hypothesis. Determine clearly if the null hypothesis value (70 ppb) is above or below this bound and explain explicitly how this relates to your hypothesis test conclusion.

   e) Calculate the power associated with an alternative of $\boldsymbol{\mu_a = 120}$ **ppm** which is the EPA standard for being considered very unhealthy.