Name: ______________________________ PUID ______________________________

STAT 350 Worksheet #13

In this worksheet, we transition into statistical inference, building upon the foundations established earlier in the course. We have already examined important concepts such as sampling techniques, experimental design, the Central Limit Theorem (CLT), and the properties of estimators, particularly unbiasedness. Up to now, our focus has primarily been on describing populations through parameters like the mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$), and calculating probabilities when the population distribution is fully known.

Statistical inference allows us to reverse this approach. Instead of starting with a known population and predicting what sample results we might observe, we now use information from a single sample to estimate unknown population parameters. This transition is fundamental in statistics because, in practice, we rarely know the true characteristics of an entire population. Instead, we typically have only a single random sample, from which we aim to infer these unknown characteristics.

However, any estimate derived from a single sample carries uncertainty due to natural sampling variability. To quantify this uncertainty, we introduce confidence intervals (CIs). Confidence intervals provide a range of plausible values for an unknown population parameter, constructed using the sample data we observe. The confidence interval is composed of two key parts: a **point estimate** derived from our **sample data**, and a **margin of error** that reflects the **precision** of this estimate. The general form of a confidence interval is expressed as:

$$\textbf{Point Estimate} \pm \textbf{Margin of Error}.$$

The **margin of error** depends on two main components: the **confidence level** (for example, 95%) and the **standard error** of the estimator. The **standard error** tells us how much the estimator, like the sample mean, would fluctuate across repeated samples of the same sample size. The **confidence level** expresses the proportion of such intervals that would include the true population parameter if we could sample repeatedly and construct a new interval each time. A higher confidence level typically expands the interval, showing that we are more certain it covers the true parameter.

In this worksheet, we will specifically explore how to construct confidence intervals for estimating the population mean when the population standard deviation is known. This will lay the foundation for future discussions, in which we transition to situations where the population standard deviation is not known.

1. Suppose you have a random sample (SRS) $\boldsymbol{X_1, X_2, \ldots, X_n}$ from a population with unknown mean ($\boldsymbol{\mu}$) and known standard deviation ($\boldsymbol{\sigma}$)**.** We want to estimate the unknown population mean $\boldsymbol{\mu}$ using the sample**.**
Recall from the Central Limit Theorem (CLT) that if the sample size is sufficiently large, the sampling distribution of the sample mean $\boldsymbol{\bar{X}}$ is approximately Normal with mean μ and standard deviation $\boldsymbol{\sigma/\sqrt{n}}$. This quantity $\boldsymbol{\sigma/\sqrt{n}}$ is also called the **standard error**.
A **pivotal quantity** is a statistic that:

- Involves the unknown parameter of interest (here $\boldsymbol{\mu}$).
- Has a known distribution that does not depend on any unknown parameters.

If $\boldsymbol{X_1, X_2, \ldots, X_n}$ come from a Normal distribution. Show clearly that the standardized sample mean $\boldsymbol{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a pivotal quantity.
Because we know the exact distribution of the pivotal quantity $\boldsymbol{Z}$ we can form a probability statement about it:

$$\boldsymbol{P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha}$$

i. Use algebra to rearrange this probability statement into an explicit confidence interval form for the population mean $\boldsymbol{\mu}$.
ii. Clearly identify the resulting $\boldsymbol{(1-\alpha)100}$% confidence interval and briefly explain why its distribution is helpful in constructing confidence intervals for the unknown but non-random $\boldsymbol{\mu}$.
iii. Carefully explain what it means to say that this is a $\boldsymbol{(1-\alpha)100}$% confidence interval. Use the concept of repeated sampling and interval construction to clearly illustrate your explanation.
iv. Also discuss what might change if the sample did not come from a Normal distribution, including the role of the Central Limit Theorem for large sample sizes.

Open Computer Assignment #5 Exploring Confidence Intervals, setup shiny, and then explore the app a little to build intuition by completing questions 1 and 2 in the computer assignment.

In real-world scenarios, we often need to determine how many observations (sample size) to collect to ensure our confidence interval achieves a specific level of precision. The precision of our estimate is quantified by the **margin of error**.

2. Recall that for estimating a population mean $\boldsymbol{\mu}$ when the standard deviation $\boldsymbol{\sigma}$ is known, the margin of error (ME) of a confidence interval is given by:

$$\textbf{Margin of Error} = z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}.$$

Suppose we want our confidence interval to have a margin of error no larger than a specified value $\boldsymbol{M}$, while maintain a confidence level of $(\mathbf{1}-\boldsymbol{\alpha})\mathbf{100\%}$.

i. Derive an explicit formula for the minimum required sample size $\boldsymbol{n}$ that ensures the margin of error does not exceed $\boldsymbol{M}$.

ii. The derived formula typically results in a non-integer value. Explain clearly why we must use the ceiling function ($\lceil\cdot\rceil$) to determine the sample size $\boldsymbol{n}$ and not the floor function ($\lfloor\cdot\rfloor$) or rounding to the closest integer.

So far, we've focused on constructing two-sided confidence intervals, which provide both a lower and an upper bound for an unknown population parameter. However, there are practical situations where our interest lies only in placing an upper bound or only a lower bound on a population mean $\boldsymbol{\mu}$. Such intervals are called **one-sided confidence bounds**. An **upper confidence bound** gives us a value that the population mean $\boldsymbol{\mu}$ is unlikely to exceed, while a **lower confidence bound** provides a value that $\boldsymbol{\mu}$ is unlikely to fall below. These bounds are provided below for your convenience.

Upper Confidence Bound

$$\boxed{\mu < \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{\left(-\infty, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)}$$

Lower Confidence Bound

$$\boxed{\mu > \bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{\left(\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right)}$$

3. An e-commerce company closely monitors the load time of its product webpage. Past experience and extensive historical monitoring have allowed the engineering team to reliably estimate the standard deviation of page load times to be $\boldsymbol{\sigma} = \mathbf{1}$ second. Recently, however, the development team implemented updates to the website that might have impacted the page's average load time $(\boldsymbol{\mu})$. The team now wants to verify if these recent changes have influenced the average load time.

To accomplish this, the team will use a confidence interval approach and perform the following analysis steps:

a) Suppose the team wants to estimate the average load time with a margin of error of at **most** $\mathbf{0.2}$ seconds at a $\mathbf{95\%}$ **level** of **confidence**. Derive the minimum required sample size $\boldsymbol{n}$ needed to satisfy this precision requirement.

b) After collecting the determined number of load time observations $\boldsymbol{n}$, the team calculates the sample mean load time as $\bar{x} = \mathbf{3.45}$ seconds. Construct a $\mathbf{95\%}$ confidence interval for the true mean load time. Interpret the meaning of this $\mathbf{95\%}$ confidence interval in practical terms using the concept of repeated sampling.

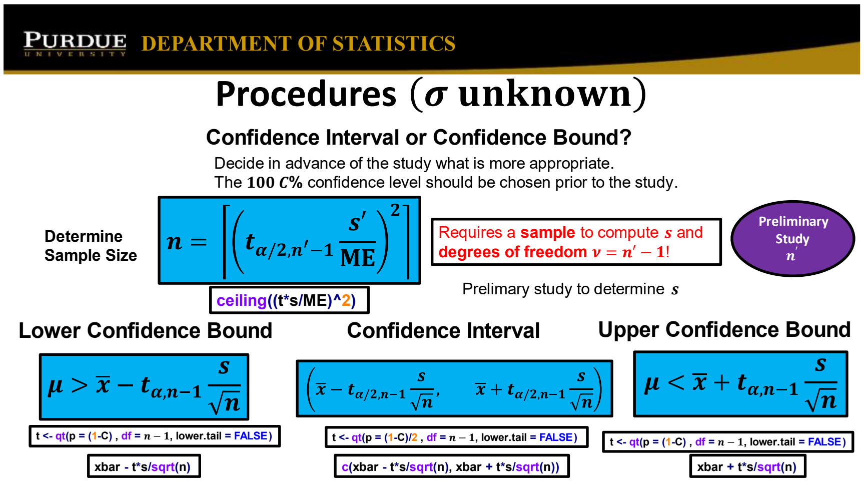
c) The management team has established a performance standard, requiring that average load times do not exceed a specific threshold. To verify compliance with this standard, the team decides to collect a sample of $\boldsymbol{n} = \mathbf{200}$ load times from the updated site. Use a one-sided upper confidence bound with a $\mathbf{99\%}$ confidence level to get an upper bound on the load times. The sample mean from the sample of **200 load times** was found to be $\bar{x} = \mathbf{3.6}$ seconds. Interpret your $\mathbf{99\%}$ upper confidence bound.

Up to now, we have assumed that the population standard deviation $\boldsymbol{\sigma}$ is **known**. In many real-world situations, however, $\boldsymbol{\sigma}$ is **not** known in advance and must be estimated from the sample data. Estimating $\boldsymbol{\sigma}$ introduces an additional layer of uncertainty, because our estimator s\,ss (the sample standard deviation) varies from sample to sample.
This increased uncertainty changes the distribution of our pivotal statistic. Previously, we relied on the standard Normal distribution (Z-distribution). Now, our pivotal quantity takes the form:

$$\boldsymbol{T = \frac{\overline{X} - \mu}{s/\sqrt{n}}}$$

which follows **Student's $\boldsymbol{t}$-distribution** rather than the standard Normal distribution. The t-distribution is similar in shape to the Normal distribution but has heavier tails, reflecting the extra uncertainty introduced by estimating $\boldsymbol{\sigma}$. The **$\boldsymbol{t}$-distribution** has an extra parameter, known as the *degrees of freedom* (df), which in this case scenario is equal to one less than the sample size $(\mathbf{df} = \boldsymbol{n} - \mathbf{1})$. Degrees of freedom represent how many independent pieces of information remain in the sample data after estimating both the mean $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$. Smaller degrees of freedom lead to thicker tails and wider confidence intervals, directly reflecting our reduced precision when estimating $\boldsymbol{\sigma}$ from limited data.

When $\boldsymbol{\sigma}$ is unknown, determining the sample size needed for a desired **margin of error** becomes more involved. Since the t-critical value itself depends on the sample size (through the degrees of freedom), we typically use a **pilot estimate** or **prior data** to approximate $\boldsymbol{\sigma}$. With this estimated value, we can then calculate a suitable sample size for our desired precision. The critical values are obtained using the **qt** function in R. You can use this similar to the **qnorm** function but need to specify the degrees of freedom. I am providing the slide with all of the important formulas below for your convenience.



PURDUE UNIVERSITY DEPARTMENT OF STATISTICS

Procedures (σ unknown)

Confidence Interval or Confidence Bound?

Decide in advance of the study what is more appropriate.
The $\mathbf{100}\ \boldsymbol{C}\%$ confidence level should be chosen prior to the study.

Determine Sample Size

$$n = \left\lceil \left(t_{\alpha/2, n'-1} \frac{s'}{\mathrm{ME}} \right)^2 \right\rceil$$

```
ceiling((t*s/ME)^2)
```

Requires a **sample** to compute s and **degrees of freedom $\nu = n' - 1$!**

Preliminary Study n'

Prelimary study to determine s

Lower Confidence Bound

$$\mu > \overline{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}}$$

```
t <- qt(p = (1-C) , df = n - 1, lower.tail = FALSE)
xbar - t*s/sqrt(n)
```

Confidence Interval

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

```
t <- qt(p = (1-C)/2 , df = n - 1, lower.tail = FALSE)
c(xbar - t*s/sqrt(n), xbar + t*s/sqrt(n))
```

Upper Confidence Bound

$$\mu < \overline{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}}$$

```
t <- qt(p = (1-C) , df = n - 1, lower.tail = FALSE)
xbar + t*s/sqrt(n)
```

4. A software company has recently upgraded its cloud servers to improve the performance of its analytics application. Specifically, the company hopes to reduce the average computation time required to run complex statistical analyses. The company does not yet know the new average computation time $\boldsymbol{\mu}$, nor the variability $\boldsymbol{\sigma^2}$ of computation times after the upgrade. To better understand the variability in performance, the engineering team first conducts a pilot study. From this pilot study of $\boldsymbol{n' = 20}$ independent analyses, they observe a sample standard deviation of $\boldsymbol{s' = 4.5}$ seconds.

a) The company wishes to collect enough data to estimate the new mean computation time with a margin of error of no more than 1 second at a $\mathbf{95\%}$ confidence level. Derive the approximate minimum sample size required to obtain this level of precision.

b) After collecting the required number of observations determined above, the company finds the average computation time from the full study to be $\boldsymbol{\overline{x} = 28.7}$ seconds with a sample standard deviation of $\boldsymbol{s = 4.8}$ seconds. Using the obtained statistics construct a **95% confidence interval** for the true mean computation time.