1. **True/False Questions. Please indicate the correct answer by filling in the circle. *If you indicate the correct answer by any other way, you may receive 0 points for the question*.**

1.1. A study is conducted to compare the test scores of two groups of students who have had different test-taking training programs. Group A consists of 30 students who received Program X, while Group B consists of 30 students who received Program Y.

Ⓣ or Ⓕ To assess if the normality assumption is met, we would analyze the histogram of the difference in test scores Program X - Program Y.

1.2. In a matched pairs design, a study is conducted to assess the effectiveness of a new test-taking training program on students' performance. A group of 30 students is selected, and their test-taking performance is measured before and after the training intervention.

Ⓣ or Ⓕ To assess if the normality assumption is met, we would analyze separately the histograms of the scores before the training intervention and the scores after the training intervention.

1.3. Data is to be gathered from a population to test a statistical hypothesis regarding the mean of a population at a significance level of **α = 0.01**. It is known that the population is normal and has a standard deviation of **σ = 85**. The researchers believe that the population mean has increased from the value from previous studies of $\mu_0 = 150$. The researchers have enough resources to collect a sample of size **n = 100**. The critical value for **α = 0.01** is $z_{0.01}$ = **2.326348**.

Ⓣ or Ⓕ The power for detecting an alternative of $\mu_a = 180$ is given by $P\left(Z > \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_{0.01}\right) = P(Z > -1.2031)$, where $Z$ is a standard normal random variable.

1.4. A farmer wants to determine the most effective fertilizer for three different types of crops: corn, wheat, and soybeans. The farm has a large field where these crops are cultivated. The farmer suspects that the field's soil quality and drainage can vary across different sections of the field. To account for this variability, the farmer decides to section off the field and conduct a randomized block design experiment.

Ⓣ or Ⓕ In this scenario the blocks of the randomized block design would be the different types of fertilizer.

1.5. In situations where it is known or suspected that individual differences can affect the response and potentially mask the treatment effects then

Ⓣ or Ⓕ a completely randomized design is the most appropriate experimental design technique.

1.6. A **95% confidence interval** for the population mean is constructed from a sample from a population known to be normally distributed with $\sigma = 215.044$ and is found to be $(-52.3, 23.1)$. The critical value for a **95% confidence level** is **1.959964**.

Ⓣ or Ⓕ The sample size used to construct the interval is $n > 100$

**2. Multiple Choice Questions. Please indicate the correct answer by filling in the circle. *If you indicate the correct answer by any other way, you may receive 0 points for the question. Only one option should be selected in each multiple-choice problem.***

**2.1** When conducting a hypothesis test, the investigator obtained a $p$**-value** of **0.002**. It is essential to note that this conclusion is based on a **small sample size**, and the investigator did not verify the underlying assumptions. You may assume a significance level of **0.05**. This means that:

Ⓐ If the sample consists of only likely events from the population, then it would suggest that either the null hypothesis $\mathbf{H_0}$ or one of the model assumptions was incorrect.

Ⓑ If all model assumptions were correct and $\mathbf{H_0}$ is true, then the data was an unlikely event.

Ⓒ If all model assumptions were correct and $\mathbf{H_a}$ is true then this sample was an unlikely event.

Ⓓ The null hypothesis $\mathbf{H_0}$ is false.

Ⓔ Both options (a), and (b) are correct.

**2.2** In a preliminary study, a **99% confidence interval** for an **unknown population mean** was computed using a **sample of 32 observations**. The known **standard deviation (σ)** of the population was given as **16**, and the resulting confidence interval was $(\mathbf{127.7145, 142.2855})$.

Now, the researcher aims to conduct a follow-up study with improved precision. They want the **width** of the **confidence interval** to be **no more than 6 units long**. Using the same **99% critical value** of $z_{0.005} = \mathbf{2.5758}$, determine the **minimal sample size** required to achieve a **99% confidence interval** of **no more than 6 units in width**.

Ⓐ 47

Ⓑ 47.18041

Ⓒ 48

Ⓓ 188

Ⓔ 188.7217

Ⓔ 189

**2.3** A one-sample analysis is performed for the mean $\mu$ where the normality assumption is valid for the sampling distribution of the estimator. An SRS is taken from the population and the test statistic for a **one-sided upper tail t-test** has a **p-value** of **0.04**. If you were to conduct a **two-tailed test**, what conclusion could you make at a significance level of **0.05**?

Ⓐ We have evidence at the 5% significance level that the population mean differs from the null value $\mu_0$.

Ⓑ We do not have evidence at the 5% significance level that the population mean differs from the null value $\mu_0$.

Ⓒ We have evidence at the 5% significance level that the population mean is less than the null value $\mu_0$.

Ⓓ We cannot conclude about a two-sided test from a **one-sided p-value**.

**2.4** A dietary supplement company plans to launch a product aimed at improving cholesterol levels. The supplement contains natural ingredients known for their cholesterol-lowering effects. Concerns arise about potential muscle cramps and spasms as adverse effects. To investigate, researchers conducted a six-month experiment with randomly assigned volunteers to the **placebo** or **supplement** group. Discomfort and pain levels are assessed through questions at the study's start (before taking supplement or placebo) and at the end (after six-months of using the supplement or placebo). A simple first step hypothesis is designed to test if the average difference in discomfort or pain scores $(\mathbf{D} = \mathbf{Pain}_{\mathbf{END}} - \mathbf{PAIN}_{\mathbf{START}})$ for the supplement group is greater than 0.

What is a **Type I error** in the scenario of the simple hypothesis and what are the negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error**?

Ⓐ A **Type I error** would mean concluding that the dietary supplement does increase discomfort or pain when there is no actual increase in discomfort or pain associated with the dietary supplement. The negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error** could include loss of a viable product, loss of potential profits, and money spent researching the product.

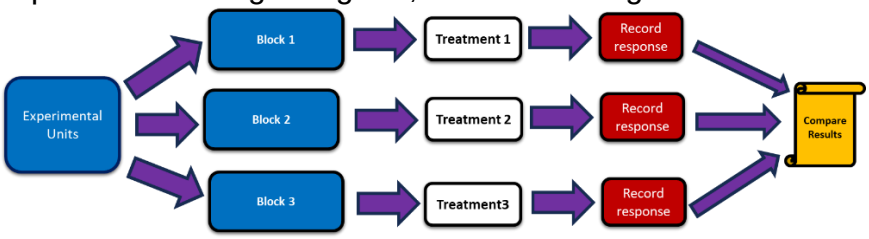
Ⓑ A **Type I error** would mean concluding that the dietary supplement has a significant effect on reducing discomfort or pain when it does not. The negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error** could include loss of consumer trust, and damaged reputation.

Ⓒ A **Type I error** would mean concluding that the dietary supplement has a significant effect on improving cholesterol levels when it does not. The negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error** could include loss of consumer trust, and damaged reputation.

Ⓓ A **Type I error** would mean concluding that the dietary supplement does not have a significant effect on improving cholesterol levels when it does. The negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error** could include loss of a viable product, loss of potential profits, and time spent researching the product.

Ⓔ A **Type I error** would mean concluding that the dietary supplement does not have a significant effect on reducing discomfort or pain when it does. The negative consequences for the dietary supplement company if the hypothesis test results in a **Type I error** could include loss of a viable product, loss of potential profits, and money spent researching the product.

**2.5** In the following experimental design diagram, what is missing?




Ⓐ There is no replication.

Ⓑ There is no randomization.

Ⓒ There is no control or comparison of levels.

Ⓓ There is no comparison of results.

Ⓔ The diagram has no issues.

**3.** A major source of profit for a new video streaming company, YouCube, is the advertisements that are played before their videos begin. The company charges their ad clients a rate per second watched. Therefore, it is of importance for YouCube to find out how long the users are willing to wait, in seconds, before the “Skip Ad” option appears on screen. To determine this information, the first 400 newly registered users were chosen. Then the company randomly assigned 4, 5, 6, or 7 seconds to wait for the “Skip Ad” option to appear on the screen for each user. Thus, each group has 100 participants. After one month, they will record the activity score of each user, ranging between 1 and 10, which reflects how active the user is on the platform at the moment of recording.

**a.** What are the units, treatment factor including the levels, and response variable in this study?

units: the 400 newly registered users
minimum required: users
The following are not required: 400, newly registered
If they just state people, that is -0.5
treatment factor: time before the "Skip Ad" option appears
levels: 4, 5, 6, 7
response variable: activity score

Look carefully at the units and the levels – these are different between the two versions.

**b.** Is this an observational study or an experiment? Please explain your answer.

experiment.
Reason: The company is assigning the number of seconds until the “Skip Ad” option appears.

**c.** State one lurking variable in this scenario. Be sure to state whether it is confounding or common response.

Why the users are watching the videos
What type of videos are being watched
Are there are there other platforms that the users could watch the videos
The ads themselves: content/quality, etc.

**d.** You are asked to consult on the design for a new version of this study. Describe one specific change that can be made to ensure that the new study provides a more convincing result. Please explain why the design is better.

Improvement. (3 pts.) They do not necessarily need to correct what was stated in c). If it is reasonable, give full credit. Explanation of why it is better (3 pts.) Again, as long as the reason is correct, give full credit.
Possible changes

The obvious one is to randomly sample all users instead of newly registered users. This will reduce bias.

They can restrict the users to only certain types of users as mentioned in c). This will help with the principle of control in that all of the users are viewing the videos for the same reason.

**4.** A box of cereal produced by company A is labelled to contain 20 ounces (oz.) of cereal (567 grams). The manufacturing machinery used by company A is old and it is feared that the boxes were not being filled properly. To understand the situation better, the company took a random sample of 50 boxes. The sample mean and sample standard deviation was found to be 19.96 oz. and 0.0901 oz. respectively. **The R output of various tests is provided on the last page.** Note that more information is provided than is needed to answer the questions.

**a.** Does the data provide any evidence that the true mean weight of the cereal boxes produced by company A is not 20 oz.? Perform the 4-step hypothesis test with α = 0.08 using the provided information. Be sure to use at least 4 non-zero digits in the p-value. State which output on the data page you used.

**Step 1: Specify the parameter of interest**
Let $\mu$ (x) denote the population (or true) (x) mean (or average) (x) weight (x) of the box of cereal.

**Step 2: Specify the hypotheses**
$H_0: \mu = 20$, $H_a: \mu \neq 20$

**Step 3: Obtain test statistic, df, and p-value**
$t_{ts}$ = -2.7893 (x) see above for other choices
df = 50 – 1 = 49 (x)
p-value = 0.007501 (x) see above for the other choices.

no work is required for the df.

**Step 4: Conclusion**

Reject $H_0$ (x) because $0.007501 \leq 0.08$ (x, give full credit if they say $p < \alpha$) If they use the incorrect $\alpha$, that is -0.5. If the α = 0.06 (that is the value in V2, contact your instructor).

The data (x) does (x – no NOT) support (p = 0.007501) (x) the claim that the true (x) mean (x) weight (x) of the box of cereal is not (x) 20. (x)

**b.** Should you use a confidence interval, lower confidence bound, or upper confidence bound to be consistent with part a)? Please explain your answer. Also calculate the appropriate confidence level of the interval or bound.

This is an interval because the alternative hypothesis is ≠ so we want a range of values.

**Confidence level = C = 1 - α = 1 – 0.08 = 0.92**

**c.** If you stated an interval in part b), would you predict that the null value would be inside or outside of the calculated interval? If you stated a bound in part b), would you predict that the null value would be greater than or less than the calculated value? Please explain your answer.

The null value should be outside of the interval because the result of the hypothesis test is reject [the hypothesis test says that the value ≠ 20]

**d.** Write down the correct critical value for the confidence interval or bound in part b). No explanation is required. Please use at least 3 decimal places.

1.787758

**e.** A statistician associated with company A argued that instead of using the sample standard deviation, they should use the population standard deviation known to be 0.10. Will the critical value change? Please explain your answer. If you state yes, also provide the new critical value.

The distribution changes from a t-distribution (do not know the standard deviation) to a z-distribution (do know sigma).
New critical value → 1.750686

**f.** What practical answer would you give to the company about whether the weight of the boxes have changed from 20 oz.? Box weights more than 0.001 oz. (0.03 gram) different from the stated weight will cause a problem for the company. The appropriate value from the confidence interval or bound is 19.99 oz. The sample mean is 19.96 oz. and the sample standard deviation is 0.0901 oz.

The difference is 20 – 19.99 = 0.01
The effect size is 0.01/0.0901 = 0.11

The difference is large based on company standards. The effect size is small. The results are not practically important. Even though statistical significance was obtained and the difference is large the effect size is small indicating that the measured difference is not large in comparison to the variability.

**g.** Company B also produces a box of cereal weighing 20 oz. If someone wants to determine if the mean weight of cereal boxes produced by company A and B are different, what type of inference would you need to perform? Choices are one-sample, two-sample independent, two-sample matched pair, or none of the above. Please explain your answer.

This is two-sample independent because there are two-populations for which there is no attribute to match sample units on.

**5.** A research team is conducting a study to evaluate the effectiveness of a general algorithmic modification for generative AI models that create natural images from text prompts. The team will use the same fixed set of text prompts to generate images both with and without the new algorithmic modification. A random sample of different AI models are selected for testing. The Fréchet Inception Distance (FID) is a metric that will be used as a measurement of the difference between the generated images and the real images. A lower FID score indicates better performance. You may assume that all assumptions are met in this study. **The R output of various tests is provided on last page.** Note more information is provided than is needed to answer the questions.

| Sour | n | $\bar{x}$ | s |
|---|---|---|---|
| **Algorithmic Modification (M)** | **38** | **20.401** | **10.910** |
| **Standard Algorithm (S)** | **38** | **21.872** | **10.081** |
| **M – S** | **38** | **-1.471** | **6.560** |

**a.** Should you use a two-sample independent procedure or a two-sample matched pairs procedure to analyze the data? Explain your answer. If you choose a matched pairs procedure, please state the common characteristic that makes these data paired.

Two-sample paired procedure. Each pair is matched by the fact that they have the same AI model or the same text prompts.

**b.** If you were to generate graphs to determine whether the assumptions are valid or not, which variable(s) would you be graphing?

The difference between the populations, M – S.

**c.** Is the average FID score lower when the algorithmic modification is implemented to generate the images at a 2.5% level of significance? Please write down the null and alternative hypotheses in symbols **ONLY**.

Let D= with modification (M) – standard (S) (REQUIRED)

$H_0: \mu_D \geq 0$ $H_a: \mu_D < 0$

**d.** For the hypothesis test in c), should you use a confidence interval, lower bound, or upper bound? Please explain your answer.

The alternative hypothesis is $<$ therefore it is appropriate to use an upper confidence bound.

**e.** Interpret the appropriate 97.5% confidence interval or bound mentioned in part d). Please include at least three decimal places for the appropriate number(s). Also state the number of the appropriate output used.

We are 97.5% confident that the difference between the population or true mean or average FID between the algorithmic modification and the standard algorithm (M - S) is less than the upper bound of 0.2712962.

**f.** Based on your answer for part e), what would the decision be for the appropriate hypothesis test for the hypotheses stated in part c)? Please explain your answer.

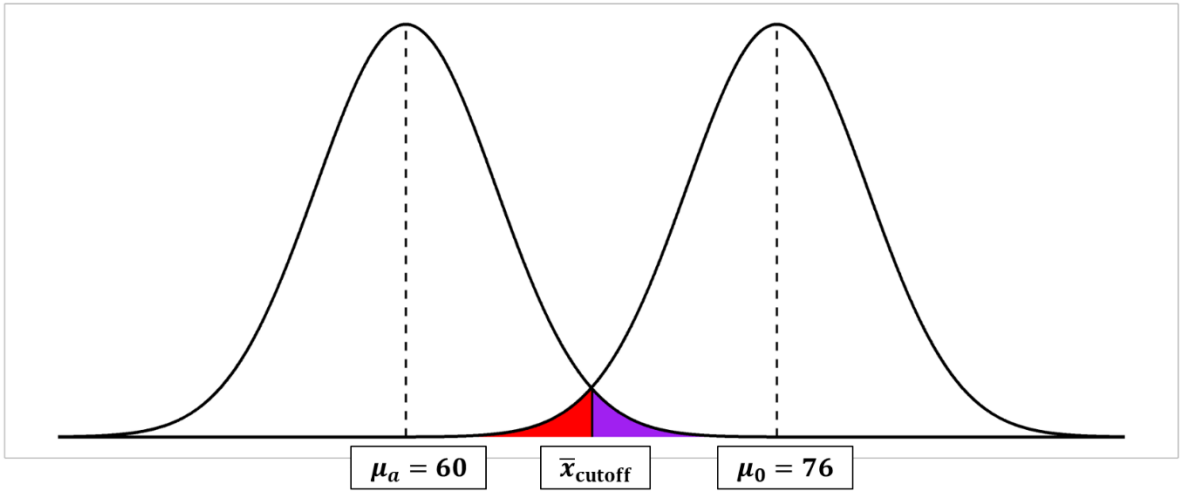
Confidence level of 97.5% and the significance of 2.5% add up to 100% and the null value of 0 is contained in the interval so we would fail to reject.

**6.** In Computer Assignment #5b you explored the relationship of **Type I** $(\alpha)$ and **Type II error** $(\beta)$, **Power** $(1-\beta)$, **sample size** $(n)$, the **standard deviation** $(\sigma)$, and the choice of an alternative mean value $(\mu_a)$ for a one-sample test of the population mean. You were asked to explore graphically what value of the sample size was required for the Type II error to be equivalent to the specified significance level **(Type I error)** for a given alternative $\mu_a$. Suppose we want to test the following hypothesis:

$\mathbf{H_0}: \mu \geq 76$

$\mathbf{H_a}: \mu < 76$

We will sample from a population that is known to be normally distributed with a standard deviation of $\sigma = 16$. We will require control of the **Type I error** at a level $\alpha = 0.02$ and require the **Type II error** to also be $\beta = 0.02$ for an alternative of $\mu_a = 60$. Determine the sample size required to ensure that the **Type I error** and **Type I error** probabilities are equal $(\alpha = \beta)$.




a. Select the correct code and output for determining the $z$ critical value (quantile) that forms the rejection region boundary of the standardized distribution. Is the correct value **(A)** or **(B)**?

**(A)**

```
> z <- qnorm(alpha, lower.tail = FALSE)
> z
[1] 2.05374
```

**(B)**

```
> z <- qnorm(0.02/2, lower.tail = FALSE)
> z
[1] 2.326348
```

**(A)** Is the correct answer.

b. Determine a formula for $\bar{x}_{\text{cutoff}}$ with respect to the null value $\mu_0$.

$P(\text{Type I Error}) = 0.02$

$P(\bar{X} < \bar{x}_{\text{cutoff}} | \mu = \mu_0) = 0.02$

$P(Z < -2.05374) = 0.02$ and $P\left(Z < \frac{\bar{x}_{\text{cutoff}} - \mu_0}{\sigma/\sqrt{n}}\right) = 0.02$

$$\frac{\bar{x}_{\text{cutoff}} - 76}{16/\sqrt{n}} = -2.05374$$

$$\bar{x}_{\text{cutoff}} = -2.05374 \times \frac{16}{\sqrt{n}} + 76$$

c. Write an expression for the probability of Type II error using $\bar{x}_{\text{cutoff}}$ in the probability statement.

$$P(\text{Type II Error}) = 0.02$$

$$P(\bar{X} > \bar{x}_{\text{cutoff}} | \mu = \mu_a) = 0.02$$

d. Standardize $\bar{x}_{\text{cutoff}}$ with respect to the alternative and equate the value to appropriate $z$ critical value (quantile). (Hint think about the standardized graphs.)

$$P(Z > 2.05374) = 0.02 \text{ and } P\left(Z > \frac{\bar{x}_{\text{cutoff}} - \mu_a}{\sigma/\sqrt{n}}\right) = 0.02$$

$$\frac{\bar{x}_{\text{cutoff}} - \mu_a}{\sigma/\sqrt{n}} = 2.05374$$

e. Solve your equation for the sample size and obtain the minimal sample size required to ensure that the **Type I error** and **Type I error** probabilities are equal ($\alpha = \beta$) for the given alternative of $\mu_a = 60$.

$$\frac{-2.05374 \times \frac{16}{\sqrt{n}} + 76 - 60}{16/\sqrt{n}} = 2.05374$$

$$\frac{16}{16/\sqrt{n}} - 2.05374 = 2.05374$$

$$\sqrt{n} = 2 \times 2.05374$$

$$n = \lceil (2 \times 2.05374)^2 \rceil$$

$$n = \lceil 16.87139 \rceil$$

$$n = 17$$

## Problem 4

In the following outputs on this page you can assume that **correct degrees of freedom (df)** and **confidence level (conf.level)** were utilized wherever appropriate.

**Output 1**

```
t.test(cereal_data, conf.level=??, alternative = "two.sided", mu = 0)
t = 1566.9, df = ??, p-value < 2.2e-16
```

**Output 2**

```
t.test(cereal_data, conf.level=??, alternative = "less", mu = 0)
t = 1566.9, df = ??, p-value = 1
```

**Output 3**

```
t.test(cereal_data, conf.level=??, alternative = "greater", mu = 0)
t = 1566.9, df = ??, p-value < 2.2e-16
```

**Output 4**

```
t.test(cereal_data, conf.level=??, alternative = "two.sided", mu = 20)
t = -2.7893, df = ??, p-value = 0.007501
```

**Output 5**

```
t.test(cereal_data, conf.level=??, alternative = "less", mu = 20)
t = -2.7893, df = ??, p-value = 0.003751
```

**Output 6**

```
t.test(cereal_data, conf.level=??, alternative = "greater", mu = 20)
t = -2.7893, df = ??, p-value = 0.9962
```

**Output 7**

| | |
|---|---|
| `> qnorm(0.08, lower.tail = TRUE)`<br>`[1] -1.405072` | `> qt(0.08, df, lower.tail = TRUE)`<br>`[1] -1.426726` |
| `> qnorm(0.08, lower.tail = FALSE)`<br>`[1] 1.405072` | `> qt(0.08, df, lower.tail = FALSE)`<br>`[1] 1.426726` |
| `> qnorm(0.08/2, lower.tail = TRUE)`<br>`[1] -1.750686` | `> qt(0.08/2, df, lower.tail = TRUE)`<br>`[1] -1.787758` |
| `> qnorm(0.08/2, lower.tail = FALSE)`<br>`[1] 1.750686` | `> qt(0.08/2, df, lower.tail = FALSE)`<br>`[1] 1.787758` |

## Problem 5

### Output 1

```
   Welch Two Sample t-test

97.5 percent confidence interval:
   -6.170136  3.227281
```

### Output 2

```
   Welch Two Sample t-test

97.5 percent confidence interval:
         -Inf 2.426865
```

### Output 3

```
      Welch Two Sample t-test

97.5 percent confidence interval:
    -5.36972      Inf
```

### Output 4

```
      Paired t-test

97.5 percent confidence interval:
 -3.5802517  0.6373965
```

### Output 5

```
   Paired t-test

97.5 percent confidence interval:
       -Inf  0.2712962
```

### Output 6

```
           Paired t-test

97.5 percent confidence interval:
   -3.214151       Inf
```