The first part of this project is on the analysis of a population mean. You will obtain data on a specific measurement \( X \) by performing a random survey on a targeted population \( \Omega \) that is divided into two disjoint sub-populations \( \Omega_1 \) and \( \Omega_2 \). The project involves computing statistics, finding confidence intervals, and performing hypothesis tests on the overall mean \( \mu \) and on the difference \( \mu_1 - \mu_2 \) of the sub-population means.

**State the Measurement and Populations**

First, you will analyze the mean of some measurement \( X \) on a specific population and compare the differences within the two sub-populations.

1. Specify the measurement \( X \) that you wish to analyze and the population \( \Omega \) that you will target in your survey. State the two disjoint sub-populations \( \Omega_1 \) and \( \Omega_2 \) that you will consider. (After defining your sub-populations, refer to them throughout by name, not simply as \( \Omega_1 \) and \( \Omega_2 \).)

2. Give the approximate size \( N \) of your population \( \Omega \) and the sizes \( N_1 \) and \( N_2 \) of the sub-populations \( \Omega_1 \) and \( \Omega_2 \). Then specify the proportionate sizes of the sub-populations. (For example, \( \Omega_1 \) is 60% of the population and \( \Omega_2 \) is 40%.)

**Initial Estimates**

Before conducting the survey, you must make some initial estimates about the mean of your measurement. Later, you will use these estimates for your hypothesis tests.

1. Give estimates for the means \( \mu_1 \) and \( \mu_2 \) of the measurement \( X \) on your two sub-populations \( \Omega_1 \) and \( \Omega_2 \). Explain the reasoning you used for your estimates. (The difference of your estimates then becomes your estimate \( M \) for the difference in means \( \mu_1 - \mu_2 \) in Question 3 under Hypothesis Tests.)

2. Using your estimates in 1, calculate a weighted estimate of the overall population mean \( \mu \). (This estimate is to be used in Question 1 under Hypothesis Tests.)

3. How do you personally compare with your estimated average? That is, what is your measurement for \( X \) and you are above, below, or close to your estimated average?

**Determining Sample Size**

1. Give an estimate for the possible range \([c, d]\) of your measurements. Give a logical explanation for your chosen bounds.

2. For your measurement, choose a desired margin of error \( e \). This value should depend on the size of your measurements and should be about 5% of the size of your estimate for the overall population mean \( \mu \).

3. For both \( \Omega_1 \) and \( \Omega_2 \), find the sample sizes \( n_1 \) and \( n_2 \) required to obtain 95% confidence intervals that have no larger than your desired margin of error \( e \). (Over)
Recall: Using $N_1$ for the size of $\Omega_1$, the required sample size $n_1$ for $\Omega_1$ is found by

$$n_1 \geq \frac{N_1 \left( \frac{1.96 \times U}{e} \right)^2}{N_1 - 1 + \left( \frac{1.96 \times U}{e} \right)^2}, \text{ where } U = \frac{d - c}{2}$$

(Use $N_2$ for $\Omega_2$.)

Conducting the Survey

Next you must conduct a random survey on the targeted population to obtain sample measurements. For scientific purposes, there usually should be at least $n_1$ respondents from sub-population $\Omega_1$ and at least $n_2$ respondents from sub-population $\Omega_2$. However, for instructional purposes here, you may limit yourself to around 85 measurements from the entire population with at least 30 from each sub-population.

But you should use sample sizes that are in proportion to the sizes of the sub-populations. For example, if $\Omega_1$ is 60% of the population, then use a 60 : 40 breakdown such as $n_1 = 60$ and $n_2 = 40$, or perhaps $n_1 = 51$ and $n_2 = 34$.

1. Explain the sample sizes $n_1$ and $n_2$ that you will use. Then explain whether or not you will use the small population correction factors for each of $\Omega$, $\Omega_1$, and $\Omega_2$. (For $\Omega_1$, compute $\sqrt{(N_1 - n_1)/ (N_1 - 1)}$. If it is close to 1, then you do not need the correction factor for $\Omega_1$. Use the analogous formulas for $\Omega$ and $\Omega_2$.)

2. Take a random survey of people specifically within your target population $\Omega$ and record their measurements. Be sure to determine to which sub-population each respondent belongs.

3. State how the survey was conducted and how randomness was ensured.

4. Include the raw data of all responses collected above as an appendix.

Note: Because you will be using the same populations for both survey questions, you should collect the data for both at the same time. Data for your responses can be easily sorted in a chart like the one below. (Be sure to obtain an exact measurement from each respondent so that the statistics can be calculated accurately.)

<table>
<thead>
<tr>
<th>Response</th>
<th>In Pop $\Omega$</th>
<th>$\Omega_1$</th>
<th>$\Omega_2$</th>
<th>Meas. X</th>
<th>Yes</th>
<th>Not Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>√</td>
<td>√</td>
<td></td>
<td>20</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Person 2</td>
<td>√</td>
<td>√</td>
<td></td>
<td>14</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Person 3</td>
<td>√</td>
<td></td>
<td>√</td>
<td>12</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Display of Data

1. List the data in one frequency chart that reflects the possible measurements and the number of people having each measurement for each of $\Omega$, $\Omega_1$, and $\Omega_2$. (See below.)

2. Make one chart that shows the sample size, sample mean, sample deviation, mode, and five-number summary for each of $\Omega$, $\Omega_1$, and $\Omega_2$. (See below.)

3. For your overall population $\Omega$, determine the 1.5 IQR bounds and state the outliers.

4. For each of $\Omega$, $\Omega_1$, and $\Omega_2$, make a histogram, with appropriate bin lengths, and state the number of responses in each bin range.

<table>
<thead>
<tr>
<th>Frequency Chart</th>
<th>Sample Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$\Omega$</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Initial Analysis

1. How do $\bar{x}_1$ and $\bar{x}_2$ compare? That is, based on this data, does one sub-population seem to have a higher average than the other? Based on these values of $\bar{x}_1$ and $\bar{x}_2$, do you think it seems possible for the true sub-population means $\mu_1$ and $\mu_2$ to be equal?

2. How does the population median compare with the mean? Specifically, state whether more than half the respondents are above average or below average.

3. For the overall population $\Omega$, compute the percentage of measurements that are within one sample deviation of the sample average, within two sample deviations of the sample average, and within three sample deviations of the sample average.

Check for Normality

1. From looking at your histogram, statistics, and the statistical intervals in Question 3 above, explain whether the overall population measurement $X$ appears to be normally distributed. If not, explain what properties of normality fail.

Complete the remainder of this survey analysis based on your conclusion from Question 1 above. That is, if you assume a normally distributed measurement, then use $t$-intervals and $t$-scores. Otherwise, assume you have “large enough samples” and use the $z$-intervals and $z$-scores with each $S$ being a reasonable estimate of each $\sigma$. 

Confidence Intervals

1. Compute 95% confidence intervals for $\mu$, $\mu_1$, and $\mu_2$.

2. Do you believe that the margins of error are small enough to reasonably pinpoint the true averages? Explain.

3. Based on these confidence intervals, is it now possible for $\mu_1$ and $\mu_2$ to be equal; that is, is there any overlap in their confidence intervals?

4. Find a 95% confidence interval for the difference $\mu_1 - \mu_2$ and explain the interval in words. Based on this confidence interval, is it now possible for $\mu_1$ and $\mu_2$ to be equal?

Hypothesis Tests

Next, you will use the sample data to perform hypothesis tests on the initial personal estimates that you made.

1. Let $M$ denote your personal estimate for $\mu$. (See Initial Estimates Question 2.) Use your data to test the null hypothesis $H_0: \mu = M$ with a one-sided alternative. Explain your conclusion in terms of the $p$-value.

2. Test the null hypothesis $H_0: \mu_1 = \mu_2$ with a one-sided alternative. Explain your conclusion in terms of the $p$-value. What do you infer now about the means being the same for each sub sub-population?

3. Now let $M$ denote your personal estimate of the difference $\mu_1 - \mu_2$ (refer to Question 1 under Initial Estimates and use your difference of the individual guesses). Use your data to test the null hypotheses $H_0: \mu_1 - \mu_2 = M$ with a one-sided alternative. Explain your conclusion in terms of the $p$-value.

State any final conclusions and add any other comments that you like.
Survey of Population Proportions

This second part of the project is on the analysis of a population proportion. Now you will analyze the “Yes/No” data from your survey on your targeted population and its two sub-populations. This part of the project involves constructing confidence intervals and performing hypothesis tests for the overall population proportion \( p \) and for the difference \( p_1 - p_2 \) of the sub-population proportions.

State the Question and Populations

1. State your (Yes/No) question and which response you want to measure. This will be called a favorable response.

2. State again the population \( \Omega \) and the two sub-populations \( \Omega_1 \) and \( \Omega_2 \) that you are considering. Henceforth, refer to them by name and not just as \( \Omega_1 \) and \( \Omega_2 \). (These must be the same populations used in the first part of the project.)

Initial Estimates

Before conducting the survey, make some initial estimates of the proportions. You will use these estimates later for your hypothesis tests.

1. Give estimates of the true proportions \( p_1 \) and \( p_2 \) within each sub-population \( \Omega_1 \) and \( \Omega_2 \) that you think will respond favorably. Give reasons for your estimates. (The difference of your estimates then becomes your estimate for the difference in proportions \( p_1 - p_2 \) for Question 3 in Hypothesis Tests.)

2. Using your estimates in Question 1, calculate a weighted estimate \( P \) for the overall population proportion \( p \). (This estimate is to be used in Question 1 under Hypothesis Tests.)

3. How would you personally respond to your question? Do you think that you would be with the majority or minority with your response?

Determining Sample Size

1. For your proportion, choose a desired margin of error \( e \) from 0.02 to 0.04.

2. For both \( \Omega_1 \) and \( \Omega_2 \), find the sample sizes \( n_1 \) and \( n_2 \) required to obtain 95% confidence intervals that have no larger than your desired margin of error \( e \).

Recall: Using \( N_1 \) for the size of \( \Omega_1 \), the required sample size \( n_1 \) for \( \Omega_1 \) is found by

\[
n_1 \geq \frac{N_1 \times \left( \frac{1.96 \times 0.5}{e} \right)^2}{(N_1 - 1) + \left( \frac{1.96 \times 0.5}{e} \right)^2}
\]

(Use \( N_2 \) for \( \Omega_2 \).)
Conducting the Survey

Now you must analyze the Yes/No data obtained while doing Survey 1.

1. Give the overall sample proportion \( \bar{p} = \frac{m}{n} \) as a fraction (\# yes / \# surveyed) and as a percentage. How does \( \bar{p} \) compare with your initial estimate of \( p \)?

2. For both \( \Omega_1 \) and \( \Omega_2 \), give the respective sample proportions \( \bar{p}_1 = \frac{m_1}{n_1} \) and \( \bar{p}_2 = \frac{m_2}{n_2} \) in fraction form and as percentages. Do these values compare favorably with your estimates?

3. Does it seem that one sub-population is much more likely to respond favorably?

4. Based on the values of \( \bar{p}_1 \) and \( \bar{p}_2 \) and possible margin of error, do you think it is possible for the true sub-population proportions \( p_1 \) and \( p_2 \) to be equal? Explain.

Two-Way Table

We next wish to display the data graphically with two-way tables:

\[
\begin{array}{c|c|c}
| & Y = \text{“Yes”} & N = \text{“Not Yes”} | \\
\hline
\Omega_1 & \ & \text{total in } \Omega_1 \\
\hline
\Omega_2 & \ & \text{total in } \Omega_2 \\
\hline
\text{Total “Yes”} & \text{Total “Not Yes”} & \text{total}
\end{array}
\]

1. Make two tables, the first of which shows the number of people in all of the categories, and the second of which shows the percentage of persons in all of the categories. (For the second, simply divide everything in the first table by the total in the bottom right.)

2. Compute the conditional percentages \( P(\Omega_1 \mid Y) \) and \( P(\Omega_2 \mid Y) \) and explain what these values mean.
Confidence Intervals

Next, you will construct confidence intervals for the true population proportion \( p \), and the true sub-population proportions \( p_1 \) and \( p_2 \).

1. Compute 95% confidence intervals for \( p \), \( p_1 \), and \( p_2 \), where

\[
p \approx \bar{p} \pm \frac{1.96 \sqrt{\bar{p}(1-\bar{p})}}{\sqrt{n}} \quad \text{or} \quad p \approx \bar{p} \pm \frac{1.96 \sqrt{\bar{p}(1-\bar{p})}}{\sqrt{n}} \sqrt{N-n} \quad \text{(for population size } N)\]

Use similar formulas for \( p_1 \) and \( p_2 \) (with \( \bar{p}_1, n_1, N_1 \), and \( \bar{p}_2, n_2, N_2 \)).

2. Do you feel that the margins of error are small enough to pinpoint the true population proportions? Explain.

3. Based on the confidence intervals for \( p_1 \) and \( p_2 \), is it possible for \( p_1 \) and \( p_2 \) to be equal; that is, is there any overlap in their confidence intervals?

4. Find a 95% confidence interval for the difference \( p_1 - p_2 \) and explain the interval in words. Based on this confidence interval, is it now possible for \( p_1 \) and \( p_2 \) to be equal?

Hypothesis Tests

Lastly, you will use the sample data to perform hypothesis tests on the initial personal estimates that you made.

1. Let \( P \) denote your personal estimate for \( p \). (See Initial Estimates Question 2.) Use your data to test the null hypothesis \( H_0: p = P \) with a one-sided alternative (you can use the 1–PropZTest). Explain your conclusion in terms of the \( p \)-value.

2. Test the null hypothesis \( H_0: p_1 = p_2 \) with a one-sided alternative (2–PropZTest). Explain your conclusion in terms of the \( p \)-value. What do you infer now about the proportions being the same for each sub-population?

3. Now let \( P \) denote your personal estimate of the difference \( p_1 - p_2 \). (See Initial Estimates Question 1.) Use your data to test the null hypotheses \( H_0: p_1 - p_2 = P \) with a one-sided alternative. Explain your conclusion in terms of the test statistic and rejection region. (Recall: In this case, if \( P \neq 0 \), then we define the test statistic by

\[
z = \frac{(\bar{p}_1 - \bar{p}_2) - P}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}\]

which follows an approximate standard normal distribution \( Z \sim N(0, 1) \) for large sample sizes.)

State any final conclusions and add any other comments that you like.