Hadi Serag Eldin
INST314

INST314 - Project 1

Section 1

Question 1 -

Control Group: Dogs without prior history of cancer who were not exposed to 2,4-D. Treatment Group: Dogs without prior history of cancer who were exposed to 2,4-D.

Question 2 -

Hypothesis in Words: Exposure to 2,4-D increases the risk of developing cancer in dogs compared to dogs who are not exposed to the herbicide.

Mathematically: H0: Ptreatment = Pcontrol, Ha: Ptreatment > Pcontrol.

Question 3 -

Alpha Level: 0.05. Test Used: Right tailed test. p-value: less than 2.2e-16.

Question 4 -

Results suggest a significant association between exposure to 2,4-D and increased cancer risk in dogs.

Question 5 -

No, this is an observational study that can only show correlation, not causation. Observational studies can detect links, but they cannot conclusively demonstrate causality due to a variety of confounding factors and biases that may have been overlooked. More controlled experiments would be needed to determine causation.

Section 2

Part 1

1. What test are we going to run?

We will use an independent samples t-test to compare the mean G3 grades of students in Mathematics and Portuguese. This test was chosen because it is appropriate for comparing the means of two independent groups with a continuous result.

2. Have we met all assumptions?

Before doing an independent samples t-test, we must ensure that the following assumptions are met: observation independence, normality of the distribution of G3 grades in both groups, and variance homogeneity. Visual examination, normality tests, and an equal variance test will be used to verify these assumptions.

3. Hypotheses and Alpha

Null Hypothesis (H0): There is no difference in the mean G3 grades between students studying Mathematics and those studying Portuguese. (H0: μMath = μPortuguese)
Alternative Hypothesis (Ha): There is a difference in the mean G3 grades between students studying Mathematics and those studying Portuguese. (Ha: μMath ≠ μPortuguese)
Alpha (α): A significance level (alpha) of 0.05 will be used, indicating a 5% chance of incorrectly rejecting the null hypothesis.

4. Results

Initial Overview of Data:
Mathematics: Mean G3 Grade = 10.42, Standard Deviation = 4.58.
Portuguese: Mean G3 Grade = 11.91, Standard Deviation = 3.23.
This suggests that, on average, students taking Portuguese have higher G3 grades compared to students' grades in Mathematics.

Assumption Checks:
Normality Test for Mathematics and Portuguese G3 grades indicated non-normal distributions.
Homogeneity of Variance Test indicated unequal variances between the groups. Given these findings, a Mann-Whitney U test was conducted as a non-parametric alternative.

Mann-Whitney U Test Results:
U statistic = 103715.0, p-value ≈ 1.95 x 10^-7. This p-value is significantly less than the alpha level of 0.05, indicating strong evidence against the null hypothesis.

We reject the null hypothesis, indicating a statistically significant difference in G3 grades between students studying Mathematics and Portuguese. Students taking Portuguese performed better on average.

Part 2

1.

A paired samples t-test will be used to compare the mean G1 and G2 grades among students in the Portuguese dataset. This test was chosen because it compares the means of two similar groups with a continuous outcome.

2.

The paired samples t-test assumptions include observation dependence, normal distribution of differences between G1 and G2, and measurements on an interval or ratio scale.

3.

Null Hypothesis (H0): There is no difference in the mean grades between G1 and G2 for students studying Portuguese (H0: $\mu G1 = \mu G2$).
Alternative Hypothesis (H1): There is a difference in the mean grades between G1 and G2 for students studying Portuguese (H1: $\mu G1 \neq \mu G2$).
Alpha (α): The significance level is set at 0.05, indicating a 5% risk of concluding a difference exists when there is none.

4.

Normality Test for Differences between G1 and G2 grades indicated a non-normal distribution with a p-value < 0.05. However, the paired samples t-test showed a statistically significant difference between G1 and G2 grades (t-test statistic = -2.945, p-value ≈ 0.0033), indicating a difference in mean grades.

We reject the null hypothesis and conclude that there is a statistically significant difference in mean grades between G1 and G2 for students studying Portuguese. The direction of this difference, as well as its consequences for student performance over time, indicate that academic achievement varies across different years. The violation of the normalcy assumption for the differences calls for a non-parametric test, yet the significance of the data supports the conclusion that there is a difference.

Section 3

1 - Measures of central tendency include the mean, weighted mean, median, weighted median, and trimmed mean. The sample is not a measure of central tendency. Wk 6

2 - Definitions: wk1

1 - Probability distribution: Probability of a specific outcome occurring

2 - Population: The entire group that you want to draw conclusions about .

3- Sample: A subset of the population .

4 - Parameter: A quantity that describes a statistical population.

5 - Variable: Characteristics or attributes that can assume different values across individuals or objects .

3 - Variable types:

1 - Dice rolls would be a discrete variable since the outcome is a countable number of dots on the dice surface .

2 - Lyme disease infection rate would be a continuous variable as it represents a rate measured over a continuous range .

3 - Satisfaction survey rankings would be an ordinal variable as they indicate an order preference but the differences between rankings may not be uniform .

4 - Whether or not a dog got parvo is a binary variable, representing a yes/no situation .

5 - IQ scores are continuous variables since they represent a score that can have a wide range of values .

6 - Ethnicity is a categorical variable, as it represents a category to which an individual can belong .

7 - Height is a continuous variable as it can be measured on a continuous scale .

8 - Income is a continuous variable because it can take on any value within an interval and can be measured precisely.

4 - Types of probability sampling include simple random sampling, stratified sampling, and cluster sampling. You would use simple random sampling when you want each member of the population to have an equal chance of being selected. Stratified sampling is used when you want to ensure that specific subgroups of the population are adequately represented within the sample. Cluster sampling is useful when it is impractical to conduct a study with a simple random sample, often due to the population being spread out over a wide area or difficult to access.

5 - Robustness with respect to normality refers to statistical methods that are not sensitive to departures from a normal distribution. Methods that are robust can provide reliable results even when the assumptions of normality are violated.

6 - Statistics can be used incorrectly in several ways, such as biased labeling, unfair comparisons, and misuse of correlation implying causation.

7 - A large sample size is an advantage because it tends to produce more accurate and reliable estimates of population parameters. However, it may not always be practical or possible due to cost, time, or logistical constraints. Researchers may opt for smaller samples due to these limitations or when they have a high-quality sampling method that does not require a large number of observations.

8- Causation can be proved if there is a strong association between the cause and effect of an experiment, if the results are consistent even in different studies and experiments, and if there is strong evidence indicating how the cause leads to the effect.

9 - An assumption of a test could be the expectation that the data follows a normal distribution or that the variances across groups are equal. If you cannot assume equal variances, you would use Welch's t-test instead of the standard Student's t-test to compare the means.

10 - The assumptions of a test often include the normality of data and homogeneity of variances. When these assumptions are not met, the test might not be valid, and the results might not be reliable.

11 - Confidence intervals can replace a t-test when the goal is to estimate a population parameter with a certain degree of confidence, rather than testing a hypothesis about the population parameter. However, a confidence interval does not directly test a hypothesis; it provides a range of values that are likely to contain the parameter of interest.