A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Helen Cristina Silva Costa |
| *Student Number* | 2024142 |
| *Module Title* | Statistical Techniques for Data Analysis |
| *Assessment Title* | CA2 |
| *Assessment Due Date* | 25th-Apr-2024 |
| *Date of Submission* | 19th-May-2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced. **I acknowledge the use of Grammarly for the purpose of improving spelling and grammar.**

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution**.**

# Task 1:

Load the dataset Q1.csv. It contains the exam scores (in percentages) of a sample of 50 students from a Dublin secondary school.

For task 1, we want to test the hypothesis that the teacher's suspicion that the average is below the national average of 70%. So, in this case, we want to analyze two averages of the same population.

The first step is to analyze if it is different from the population average and for this, it is important to choose an appropriate statistical test.

Defining hypothesis:

1. State the Null Hypothesis: Mean score = 70%

2. Alternative Hypothesis: Mean score < 70%

Before choosing the appropriate hypothesis test, it is necessary to test the normality of the data distribution. In some statistical tests, if the distribution of data, errors or differences is not normal, it is necessary to do some kind of transformation or use a non-parametric version (Mello, 2012). Verifying normality of distribution is a requirement for running several of the well-known statistical tests such as Student’s t-test and ANOVA (Biostats, n.d.).

A screenshot of a graph

Description automatically generated

Figure 1: Histogram of df['exam\_score']

We have 4 main Normality Tests

1. Shapiro-Wilk;
2. Kolmogorov-Smirnov;
3. Lilliefors;
4. Anderson-Darlin
   1. Shapiro-Wilk Shapiro-Wilk test is a hypothesis test that evaluates whether a data set is normally distributed. It evaluates data from a sample with the null hypothesis that the data set is normally distributed. A large p-value indicates the data set is normally distributed, a low p-value indicates that it isn’t normally distributed(Malato, 2023).
   2. Kolmogorov-Smirnov This test compares the ECDF (empirical cumulative distribution function) of your sample data with the distribution expected if the data were normal. If this observed difference is adequately large, the test will reject the null hypothesis of population normality. If the p-value of this test is less than your chosen α, you can reject your null hypothesis and conclude that the population is nonnormal (support.minitab.com, n.d.).
   3. Lilliefors The Lilliefors test is a variant of the Kolmogorov-Smirnov test that is specifically designed to test normality. It evaluates whether the data comes from a normal distribution by comparing the empirical distribution function of the data with the expected normal cumulative distribution function (RCODER, 2023).
   4. The Anderson-Darling test has a similar objective to that of the Kolmogorov-Smirnov test, but it is more powerful. This is so since all the data values are considered, not just the one that produces the maximum difference. Also, more weight is given to the tails of the distribution being fitted. Generally, this test should be used instead of the Kolmogorov-Smirnov test (Real-statistics.com, 2024).

The Shapiro–Wilk test is more appropriate method for small sample sizes (<50 samples) although it can also be handling on larger sample size while Kolmogorov–Smirnov test is used for n ≥50. For both of the above tests, null hypothesis states that data are taken from normal distributed population. When P > 0.05, null hypothesis accepted and data are called as normally distributed (Mishra et al., 2019).

The Shapiro-Wilk test is a test of normality. It is used to determine whether or not a sample comes from a normal distribution(Zach, 2020).

A screenshot of a computer code

Description automatically generated

Figure 2: Shapiro test.

Since the p-value is not less than 0.05, we fail to reject the null hypothesis, So the data comes from a normal distribution. With that we can choose the test of hypothesis.

First, let's calculate the average separately.

A screenshot of a computer

Description automatically generated

Figure 3: Mean of df['exam\_score'].

The sample mean is different from 70, but this is not enough evidence to prove that the sample mean is lower than the national average of 70!

Use a one-sample t test to compare a sample mean to a reference value. It allows you to determine whether the population mean differs from the reference value. The reference value is usually highly relevant to the subject area (Frost, n.d.).

Considering that we have a normal distribution and population standard deviation is unknown a one-sample t-test can be used.

the level of significance α = 0.05.

A screenshot of a computer program

Description automatically generated

Figure 4: One sample T-test.

If the p-value is < α, we reject the H0, if the p-value is > α then we fail to reject the H0.

Note in this case p-value is great than significance level α, so we fail to reject the null-hypothesis. The difference between the sample mean and the population mean value is not statistically significant.

# Task 2:

Load the diamonds dataset and print the first 5 rows.

A screenshot of a computer

Description automatically generated

Figure 5: Diamonds dataset.

To Create a new binary variable in the data frame called “colourless” which records 1 in rows with colourless diamonds and 0 otherwise, we’ll use the lambda function.

The apply function will apply each element of the series to the lambda function. The lambda function will then return a value for each element based on the expression you passed to it (Abayomi Ogunbiyi, 2022). In our case, the expression was x: 1 if x in ['D', 'E', 'F'] each element.

A screenshot of a computer

Description automatically generated

Figure 6: colourless

To perform an appropriate hypothesis test to determine whether there is any association between the clarity of a diamond and whether it is colourless or not. First, we had to use a contingency table.

A contingency table is a type of table that summarizes the relationship between two categorical variables. To create a contingency table in Python, we can use the pandas.crosstab() function (Zach, 2021).

A screenshot of a computer code

Description automatically generated

Figure 7: contingency table.

Our goal is to verified if we have any association between the clarity of a diamond and whether it is colourless or not. A significance

level of alpha = 0.01.

Common Tests for Association

1. The chi square test for association: is used to determine if there is any association between two categorical variables. It is really a hypothesis test of independence. The null hypothesis is that the two variables are not associated, i.e., independent. The alternate hypothesis is that the two variables are associated (SPC for Excel, n.d.).

A Chi-Square test of independence uses the following null and alternative hypotheses:

H0: (null hypothesis) The two variables are independent.

H1: (alternative hypothesis) The two variables are not independent. (i.e. they are associated) (Zach, 2021b).

3. The Cochran-Mantel-Haenszel (CMH) Test: use when you have data from 2×2 tables that you've repeated at different times or locations. It will tell you whether you have a consistent difference in proportions across the repeats.The most common situation is that you have multiple 2×2 tables of independence; you're analyzing the kind of experiment that you'd analyze with a test of independence, and you've done the experiment multiple times or at multiple locations. There are three nominal variables: the two variables of the 2×2 test of independence, and the third nominal variable that identifies the repeats (McDonald, 2014).

4. Fisher’s Exact Test: is grounded in the core statistical principle of contingency table analysis, which focuses on determining whether there is a non-random association between two categorical variables. This test is particularly suited for 2×2 contingency tables, which represent the simplest form of data classification with two variables that can each take on two different values. Compared to other statistical tests like the Chi-square test, Fisher’s test is uniquely reliable when sample sizes are small and when the data contains cells in the contingency table with expected frequencies less than 5. In such cases, Fisher’s Exact Test is preferred due to its exact nature, not relying on any approximation or large sample assumptions(Easily, 2024).

5. Goodman Kruska’s Gamma: s a non-parametric statistical measure identifying the correlation between two ordinal variables. It determines the direction and strength of the relationship between these data pairs to predict the future trend or values. The paired data sets should comprise ordinal variables. Ordinal variables possess categories or levels with natural order but lack specific numerical values. The paired variables should exhibit a monotonic connection whereby a rise in one variable results in an apparent increase or decrease in the rank of another variable (Choubey, 2024).

For our test, we'll perform the CHI square test because the definition is exactly what we want to analyze whether the two categorical variables have any association, and instead of the Fisch test, the CHI square can be used for data with large frequency numbers.

A screenshot of a computer

Description automatically generated

Chi-square test, a p-value that is less than significance level indicates there is sufficient evidence to conclude that a relationship exists between the categorical variables.

* Find and interpret 90% confidence intervals for both the mean price of colourless diamonds and the mean price of non-colourless diamonds.

A confidence interval for a mean is a range of values that is likely to contain a population mean with a certain level of confidence. What to do that? The reason is we want to capture our uncertainty when estimating a population, in our case price mean of colourless and non\_colourless (Zach, 2020a).

A screenshot of a computer code

Description automatically generated

Figure 8: confidence intervals.

There is only a 10% chance that the true mean price lies outside of the 90% confidence interval. That is, there’s only a 10% chance that the true mean price of colourless is greater than 3373.6593252499883 or less than 3301.85947692509.

There is only a 10% chance that the true mean price lies outside of the 90% confidence interval. That is, there’s only a 10% chance that the true mean price of non-colourless is greater than 4533.679900742729 or less than 4448.7802444452245.

Task 3: The PlantGrowth dataset.

Comparing the trt2 groups has the highest mean and lowest standard deviation, followed by the control group with a mean of 5.032 and a standard deviation of 0.583, thus being the lowest mean of the trt1 group with 4.661, but with the highest standard deviation of 0.793.

A screenshot of a computer

Description automatically generated

Figure 9: summary statistic.

The ctrl treatment showed the greatest variability in scores and may be skewed to the left. In comparison, the trt1 treatment had the lowest average weight, while the trt2 treatment had the highest average weight among the three.

A screenshot of a graph

Description automatically generated

Figure 10: Box plot.

* Conduct an appropriate hypothesis test to see if there is evidence of a difference between the three means (that is, the control and the two treatments). Use a significance level of alpha = 0.05.

ANOVA has two types of tests: one-way and two way (can be with or without replication). one-way between group used to test two groups to see if there's a difference between them. one way has one independent variable affecting a dependent variable. the one way tells that at least two groups were different from each other. But it won't tell you which groups were different. Two ways has two independent variables. two way is used when you have one quantitative outcome, and you have two categorical explanatory variables Two way without replication used when you have one group an you're double-testing that same group. Two-way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing (Statistics How To, n.d.).

The purpose of a one-way ANOVA (analysis of variance) test is to determine the existence of a statistically significant difference among the means of three or more populations. The test uses variances to help determine if the population means are equal or not.

The following assumptions are required to use a one-way ANOVA test:

Each population from which a sample is taken is normally distributed. All samples are randomly selected and independently taken from the populations. The populations are assumed to have equal variances. The population data is numerical (interval or ratio level) (Watts, 2022).

Hypothesis:

H0 = All means are equal.

HA = At least one population mean is different from the others.

A screenshot of a computer program

Description automatically generated

Figure 11: Anova test part 1.

A screenshot of a computer

Description automatically generated

Figure 12: Anova 2.

The p-value is less than significance level (0.05), so reject the null hypothhesis and conclude that not all population means are equal.

In the interval plot, trt1 has the lowest mean and trt2 has the highest. but just with this graph, we cannot determine whether any differences are statistically significant.

A screen shot of a graph

Description automatically generated

Figure 13: Mean and confidence interval for each group.

To determine where the difference between means may lie, I will utilize two methods: the Tukey method and the Dunnett method. Tukey method: ANOVA will tell you if there are differences among the levels of the independent variable, but not which differences are significant. To find how the treatment levels differ from one another, perform a TukeyHSD (Tukey’s Honestly-Significant Difference) post-hoc test.y) The Tukey test runs pairwise comparisons among each of the groups, and uses a conservative error estimate to find the groups which are statistically different from one ano (scipy.github.io, n.d.)th aDs. Dunnett method is used in ANOVA to create confidence intervals for differences between the mean of each factor level and the mean of a control group. If an interval contains zero, then there is no significant difference between the two means under comparison. You specify a family error rate for all comparisons, and Dunnett's method determines the confidence levels for each individual comparison accordingly (Minitab support, n.d.).

A screenshot of a computer code

Description automatically generated

Figure 14: Dunnett test.

The p-value (0.323) corresponding with the comparison between sample 0 (trt1) and control exceeds 0.05. The p-value (0.154) corresponding with the comparison between sample 0 (trt2) and control exceeds 0.05.

This means we do not have sufficient evidence to say that there is a difference between the control treatment and the two treatments (trt1 and trt2). Therefore, there is no significant evidence that the means of treatment compared to the control group differ. So, in this case we "Fail to reject the null hypothesis: The differences between the means are not statistically significant."

A screenshot of a computer

Description automatically generated

Figure 15: Tukey test.

According to Tukey's method, the only one that has a statistically evident difference is the comparison between trt1 and trt2 the p-adj (0.012) is less than the alpha value 0.05, so in this case, we can reject the null hypothesis and accept that the 'differences between some of the means are statistically significant'.

# Task4: The trees dataset

The pairplot shows us that we have a good approximately linear relationship between girth and volume. But girth/height and height/volume don't have a good linear approximation.

A screenshot of a graph

Description automatically generated

Figure 16: pairplot of trees.

Fisrt goal: a. Perform a correlation analysis between all numerical variables. Include and comment on the results of hypothesis tests for the population correlation coefficients between all three pairs of variables (you can use the pearsonr function from the scipy.stats library).

The Pearson correlation coefficient [1] measures the linear relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases (Scipy.org, 2019).

Pearson’s correlation coefficients measure only linear relationships. Consequently, if your data contain a curvilinear relationship, the Pearson correlation coefficient will not detect it (Frost, 2018).

ρ = population correlation coefficient

Null hypothesis: There is no linear relationship between the two variables. ρ = 0.

Alternative hypothesis: There is a linear relationship between the two variables. ρ ≠ 0.

A screen shot of a graph

Description automatically generated

Figure 17: Girth and height.

The coefficient R 0.519 fits in a positive relationship and the p-value is less than 0.05 with these results we can reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between Girth (x) and Height (y) because the correlation coefficient is significantly different from zero. When the distance between the points on the drawn line is not very close, it suggests a moderate correlation, as indicated by a correlation coefficient of 0.519.

A screen shot of a graph

Description automatically generated

Figure 18: Volume and height.

The same occurs in height and volume, if there is a positive relationship the p-value is less than 0.05, however we have a moderate correlation, since according to the graph the points are far from each other.

A screen shot of a graph

Description automatically generated

Figure 19: Girth and volume.

The correlation coefficient between 'Girth' and 'Volume' is 0.967, which is higher compared to the other values obtained. The points on the graph are close to each other, forming a perfect line and indicating a very strong linear relationship. A positive correlation implies that as 'Girth' increases, so does 'Volume'. Therefore, we can reject the null hypothesis since the p-value is less than 0.05 and the correlation coefficient is different from zero.

* Multiple linear regression

Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. It is an important regression algorithm that models the linear relationship between a single dependent continuous variable and more than one independent variable. It uses two or more independent variables to predict a dependent variable by fitting a best linear relationship (Yadav, 2021).

A screenshot of a computer program

Description automatically generated

Figure 20: model 1 Girth + height.

Results

Model 1 (Girt + Height): The R-squared is 0.948, which is our coefficient of determination, which indicates that 94.8 of volume can be explained by Girth and Height, and looking at the p-value of both, we can say that the two are quite significant to predict volume, but girth is more important since its p-value is equal to zero, i.e. less than the alpha significance level of .05.

A screenshot of a computer

Description automatically generated

Figure 21: model 2 Height and volume.

Model 2 (Height): The R-squared is 0.358, which indicates that 35.8% of volume can be explained by height, and looking at the p-value, we can say that the height is statistically significant to predict volume. However, it only predicts 35.8% of volume, so the model fit is poor compared to Model 1.

A screenshot of a computer

Description automatically generated

Figure 22: model 3 Girth and volume.

Results

Model 3 (Girt): The R-squared is 0.935, which indicates that 93.5% of volume can be explained by girth, and looking at the p-value, we can say that the girth is highly statistically significant (p-value=0.000) to predict volume.

Among the three linear regression models, the ones that best estimate volume are models 1 and 3. Model 1 explains 94.8% of the volume, being one of those that would indicate estimating the volume. However, model 3 would also be a great option for this since it uses only one predictor variable and still has an accuracy of 93.5%, which is very good.