### Hong Kong Baptist University

Department of Computer Science

*COMP 7990 Principles and Practices of data analytics (2021-22)*

Name: \_\_\_\_\_\_Zhu Hengxi\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SID:\_\_\_\_\_\_\_\_21405824\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 1 – Generate descriptive statistics**

1. Download and unzip the file **Assignment1-dataset.zip**. Open the file **diet.csv** using Jamovi, save the file as **diet.omv**. This data set contains information on 78 people using one of three diets.
2. Change the measure types for the following attributes:

|  |  |
| --- | --- |
| **Attribute** | **Measure Type** |
| Person | ID |
| Age | Continuous |
| Height | Continuous |
| pre.weight | Continuous |

1. Show a statistical summary for the variable **age**, including the sample size, percentiles (25,50,75), mean, median, mode, standard deviation, variance, range, maximum and minimum. Paste the snapshot of the output below.

表格

描述已自动生成

1. Plot the age distribution in a histogram.

图表, 直方图

描述已自动生成

1. Split the result in c) by gender.

图表, 直方图

描述已自动生成

**Exercise 2 – Independent samples t-test**

1. Independent samples t-test compares the means of two independent groups to determine whether there is a statistically significant difference between the means of an attribute. Let us perform an **independent samples t-test** to compare the means of height for different genders by using **diet.omv**.
2. Show the **Group Descriptives** table.

表格

描述已自动生成

1. Perform the assumption checks on **normality** and **homogeneity (equality of variance)** and paste the results below. Does it violate Normality Test (Shapiro-Wilk test) or Homogeneity of Variances Test (Levene’s test)?

表格

中度可信度描述已自动生成

1. Violate Normality Test? \_\_\_yes\_\_\_\_\_\_\_\_\_\_\_\_\_(Yes/No)
2. Violate Homogeneity of Variances Test? \_\_\_\_\_\_\_No\_\_\_\_\_\_\_\_\_\_(Yes/No)
3. Suggest a way to correct the results if one of the tests above is violated:

\_use Welch’s test\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. In the previous question, if one of the tests is violated, you may use another test to correct it and **capture** the independent sample t test result screenshot including p value, mean difference, effect size etc.

表格

描述已自动生成

1. Report the overall test results.

Overall Test results:

• An independent sample t-test was conducted to determine if there were differences in the means of height between male and female students.

• The means of height for each level of gender were not normally distributed, as assessed by ShapiroWilk test (p < 0.05).

• Homogeneity of variances was violated, as assessed by Levene’s Test for Equality of Variance (p =0.129), so the Welch's t-test was used.

1. Show the **Descriptives plots**.

图表, 箱线图

描述已自动生成

**Exercise 3 – ANOVA**

1. The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent groups. Let us perform **one-way anova** to compare the height of people with 3 different diets by using **diet.omv**.
2. Show the **one-way anova** result table below.

表格

描述已自动生成

1. Perform the assumption checks on **normality** and **homogeneity (equality of variance).** Paste the result in the box below. Determine whether Fisher’s test will be used and capture the results again.

表格

中度可信度描述已自动生成

Homogeneity of variances was met, as assessed by Levene’s Test for Equality of Variance (p > 0.05), so Fisher’s test was used.

表格

描述已自动生成

1. Show the **Group Descriptives** table.

表格

描述已自动生成

1. Show the **Descriptives plots**.

图表, 箱线图

描述已自动生成

1. Perform the **post-hoc test** using **Tukey** and paste the result below. Include the mean difference, p-value in your table.

表格

描述已自动生成

1. Report the test result.

Overall Test results:

• A one-way ANOVA test was conducted to determine if there were significantly differences in the height of people with 3 different diets.

• The height of people with 3 different diets were normally distributed, as assessed by Shapiro-Wilk test (p > 0.05).

• Homogeneity of variances was met, as assessed by Levene’s Test for Equality of Variance (p > 0.05), so Fisher’s test was used.

• There was a significant difference of the height of people at the p < 0.05 for the 3 different diets. F(2, 75)=2.10, p=0.130. The post hoc comparisons using Tukey test indicated that the mean the height of people for diet1 (M=170, SD=10.95), diet2 (M=173, SD=8.90), diet3 (M=167, SD=9.71) was not significantly different from each other.

1. Save the file **diet.omv**

**Exercise 4 – Paired samples t-test**

1. The paired-samples t-test compares the means of two related groups to determine whether there is a statistically significant difference between these means. Let us perform a **paired samples t-test** to compare the weight before and after taking a fitness program by using **diet.omv.** (Assume pre-weight is the weight before taking the fitness program, weight6weeks is the weight after a 6-week fitness program)
2. Show the **paired samples t-test** result table below. Include the mean difference (CI=95%) and effect size in your table.

表格

描述已自动生成

1. Perform the assumption check on **normality** and paste the result below.

表格

描述已自动生成

As assessed by Shapiro-Wilk test (p = 0.802), the normality assumption was violated, so Wilcoxon rank test was used instead.

表格

描述已自动生成

1. Show the **Descriptives table**.

表格

描述已自动生成

1. Show the **Descriptives plots**.

图表, 箱线图

描述已自动生成

1. Report the overall test result.

Overall Test results:

• A paired sample t-test was conducted to determine if there were statistically significant differences in the weight before and after taking a fitness program.

• As assessed by Shapiro-Wilk test (p = 0.802), the normality assumption was violated, so Wilcoxon rank test was used instead.

• Wilcoxon rank test showed that there was a significant difference (p <0.001) in the median weight before and after taking a fitness program.

• the median weight before taking a fitness program was 72.0, compared to 69.0 after taking a fitness program. These results support that taking a fitness program made a difference on the weight.

1. Save the file **diet.omv**

**Exercise 5 – Correlation and linear regression**

1. Simple linear regression is useful for finding relationship between two continuous variables. Open the file **weather.csv** using Jamovi and save the file as **weather.omv**
2. Show the **scatterplot** of Humidity (Y-axis) vs Temperature (X-axis). Include the linear regression line in your scatterplot. (Hint: **scatr** module must be installed first)

图表, 散点图

描述已自动生成

1. Find the **Pearson correlation coefficient** betweenTemperature and Humidity. Show the **correlation matrix** table below. Include the p-values and flag significant correlations in your table.

表格

描述已自动生成

The correlation coefficient **r = \_\_\_\_-0.700\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**, this means that the two variables are \_\_\_negatively\_\_\_\_\_\_\_\_(positively/negatively) correlated.

1. Use **linear regression** to build a model to predict the humidity by using temperature. Show the **Model Fit Measures table** with overall model test (**R**, **R2** ,**p value** and the **F test**).

表格

描述已自动生成

1. Show the **Model Coefficients** table.

表格

描述已自动生成

1. The formula for predicting humidity is:

Humidity = 1.0481 – 0.0252(Temperature)

1. **Report the overall results** and save the file **weather.omv**

Overall Test results:

• A simple linear regression was carried out to predict the humidity based on the temperature.

• A significant regression equation was found (F (1, 718) = 690, p < 0.001), with an R\*2 of 0.490. The humidity is equal to 1.0481 – 0.0252(Temperature) when the temperature is measured in degree Celsius.

• Humidity decreased by 0.0252 for each degree increase of temperature.

**Exercise 6 – Logistic regression**

1. Logistic Regression is a regression technique that is used when we have a **categorical outcome**.

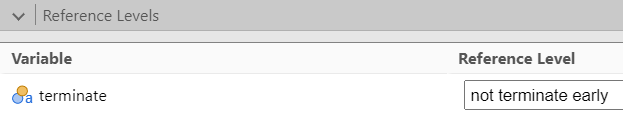
Open the **terminate.csv**, save it as **terminate.omv**. We are going to use binomial logistic regression to determine the likehood of early termination from counseling in samples at a community mental health center.

1. Change the **Measure Type** of the attributes **avoidanceOfDisclosure** and **symptomSeverity** as **continuous**.
2. Use **binomial logistic regression** to build a model. Use **avoidanceOfDisclosure** and **symptomSeverity** as predictors. Place the predictors in different blocks using **Model Builder**. Show the **Model Fit Measures table** by clicking **Overall model test**.

表格

描述已自动生成

1. Set the **References Levels** of the predictors as follow.



1. Show the **Model Coefficients table**. Include the **odds ratio** in your table.

表格

描述已自动生成

1. Perform the assumption check on **collinearity statistics** and paste the results below.

表格

描述已自动生成

1. VIF is below 10? \_\_\_\_\_\_\_\_\_yes\_\_\_\_\_\_\_\_\_\_\_(Yes/No)
2. Tolerance is greater than 0.2? \_\_\_\_\_\_\_yes\_\_\_\_\_\_\_\_\_\_\_\_\_(Yes/No)
3. Under the **Prediction** option, show the **Classification table** of the prediction below.

表格

描述已自动生成

1. **Report the overall result** and **save** the file **terminate.omv**.

Overall Test results:

• A binomial logistic regression was carried out to ascertain the effects of avoidanceOfDisclosure and symptomSeverity that the likehood of early termination from counseling in samples at a community mental health center.

• The logistic regression model was statistically significant, χ2(2) = 26.2, p < 0.001. The model explained 42.3% (McFadden’s R^2) of the variance in decision and correctly classified 82.2% of cases.

• An increase in avoidanceOfDisclosure was associated with an increase in the likelihood of early termination from counseling in samples at a community mental health center.

• An increase in symptomSeverity was associated with a decrease in the likelihood of early termination from counseling in samples at a community mental health center.

• [Model: -1.139 + 0.356(avoidanceOfDisclosure) – 0.317(symptomSeverity)]

**Assignment Submission**

Submit the file **lab1-assignment-ans.docx, diet.omv, weather.omv** and **terminate.omv** to bulearning website