### **Hong Kong Baptist University**

### Department of Computer Science

*COMP 7990 Principles and Practices of data analytics (2022-23)*

*Assignment 2*

Name: Zhang Chi SID:\_\_22407987\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 1 – Generate descriptive statistics**

1. Download and unzip the file **Assignment2-datafile.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. Show age distribution in histogram. Paste the screenshots in the table below.

|  |  |
| --- | --- |
| **Descriptives** | **Age distribution in histogram** |
| | Descriptives | | | | | --- | --- | --- | --- | |  | | **Age** | | | N |  | 78 |  | | Mean |  | 38.2 |  | | Median |  | 38.0 |  | | Mode |  | 36.0 |  | | Standard deviation |  | 9.82 |  | | Variance |  | 96.3 |  | | Range |  | 44 |  | | Minimum |  | 15 |  | | Maximum |  | 59 |  | | 25th percentile |  | 31.3 |  | | 50th percentile |  | 38.0 |  | | 75th percentile |  | 45.8 |  | |  | | | | |  |

1. Split the ages by different diets. And **save** the **diet.omv** file.

| Descriptives | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | **Diet** | | **Age** | |
| N |  | 1 |  | 24 |  |
|  |  | 2 |  | 27 |  |
|  |  | 3 |  | 27 |  |
| Mean |  | 1 |  | 39.9 |  |
|  |  | 2 |  | 38.0 |  |
|  |  | 3 |  | 36.8 |  |
| Median |  | 1 |  | 39.5 |  |
|  |  | 2 |  | 38 |  |
|  |  | 3 |  | 35 |  |
| Mode |  | 1 |  | 49.0 |  |
|  |  | 2 |  | 36.0 |  |
|  |  | 3 |  | 34.0 |  |
| Standard deviation |  | 1 |  | 9.73 |  |
|  |  | 2 |  | 9.51 |  |
|  |  | 3 |  | 10.3 |  |
| Variance |  | 1 |  | 94.6 |  |
|  |  | 2 |  | 90.5 |  |
|  |  | 3 |  | 106 |  |
| Range |  | 1 |  | 38 |  |
|  |  | 2 |  | 38 |  |
|  |  | 3 |  | 38 |  |
| Minimum |  | 1 |  | 21 |  |
|  |  | 2 |  | 15 |  |
|  |  | 3 |  | 19 |  |
| Maximum |  | 1 |  | 59 |  |
|  |  | 2 |  | 53 |  |
|  |  | 3 |  | 57 |  |
| 25th percentile |  | 1 |  | 35.0 |  |
|  |  | 2 |  | 31.5 |  |
|  |  | 3 |  | 30.0 |  |
| 50th percentile |  | 1 |  | 39.5 |  |
|  |  | 2 |  | 38.0 |  |
|  |  | 3 |  | 35.0 |  |
| 75th percentile |  | 1 |  | 47.5 |  |
|  |  | 2 |  | 43.5 |  |
|  |  | 3 |  | 45.0 |  |
|  | | | | | |

图片包含 图表

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.

| Group Descriptives | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Group** | | **N** | | **Mean** | | **Median** | | **SD** | | **SE** | |
| Height |  | Female |  | 33 |  | 175 |  | 176 |  | 9.99 |  | 1.74 |  |
|  | | Male |  | 45 |  | 169 |  | 170 |  | 9.31 |  | 1.39 |  |
|  | | | | | | | | | | | | | |

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)?

| Normality Test (Shapiro-Wilk) | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | **W** | | | | | | **p** | | | |
| Height | |  | 0.963 | | | |  | | 0.025 | | |  |
| Note. A low p-value suggests a violation of the assumption of normality | | | | | | | | | | | | |
|  | | | | | | | | | | | | |
| Homogeneity of Variances Test (Levene's) | | | | | | | | | | | | | | |
|  | | **F** | | | **df** | | | **df2** | | | **p** | | | |
| Height |  | 2.35 | |  | 1 |  | | 76 | |  | 0.129 | | |  |
| Note. A low p-value suggests a violation of the assumption of equal variances | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | |

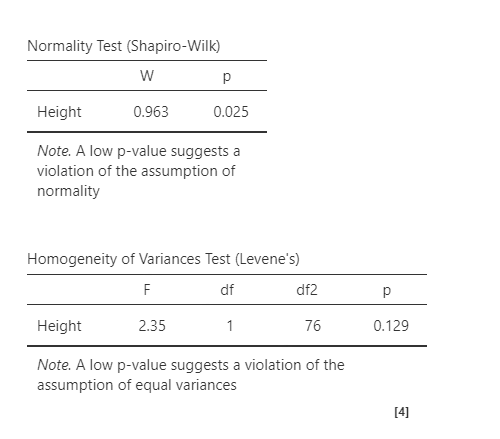
1. Does it violate Normality Test? \_\_\_\_Yes\_\_\_\_\_\_\_\_\_\_\_\_\_(Yes/No)
2. Does it violate Homogeneity of Variances Test? \_\_No\_\_\_\_\_\_\_\_\_\_\_\_(Yes/No)
3. Suggest a way to correct the results if one of the tests above is violated:

\_\_\_using Mann-Whitney U’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.

| Independent Samples T-Test | | | | | | | | | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | **95% Confidence Interval** | | | |  | | | | **95% Confidence Interval** | | | |
|  | |  | | **Statistic** | | **df** | | **p** | | **Mean difference** | | **SE difference** | | **Lower** | | **Upper** | |  | | **Effect Size** | | **Lower** | | **Upper** | |
| Height |  | Student's t |  | 2.70 |  | 76.0 |  | 0.009 |  | 5.94 |  | 2.20 |  | 1.56 |  | 10.3 |  | Cohen's d |  | 0.618 |  | 0.140 |  | 1.09 |  |
|  |  | Mann-Whitney U |  | 480 |  |  | | 0.008 |  | 7.00 |  |  |  | 2.00 |  | 10.0 |  | Rank biserial correlation |  | 0.354 |  |  |  |  |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |

1. Report the **overall test results**.



1. Show the **Descriptives plots**.

图表

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.

| One-Way ANOVA | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | **F** | | **df1** | | **df2** | | **p** | |
| Height |  | Welch's |  | 2.31 |  | 2 |  | 48.7 |  | 0.110 |  |
|  | | Fisher's |  | 2.10 |  | 2 |  | 75 |  | 0.130 |  |
|  | | | | | | | | | | | |

1. Perform the assumption checks on **Normality** **Test** and **Homogeneity Test (for checking equality of variance).** Paste the results in the box below. Determine whether Fisher’s test will be used and capture the results again.

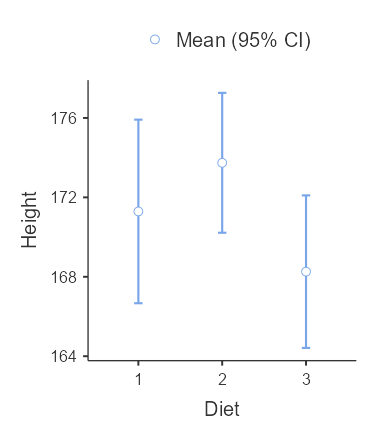
| Normality Test (Shapiro-Wilk) | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | **W** | | | | | | | **p** | |
| Height | | | | |  | | 0.975 | | | | | |  | 0.132 |  |
| Note. A low p-value suggests a violation of the assumption of normality | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Homogeneity of Variances Test (Levene's) | | | | | | | | | | | |
|  | | **F** | | **df1** | | | | **df2** | | **p** | |
| Height |  | 0.114 |  | 2 | |  | | 75 |  | 0.892 |  |
|  | | | | | | | | | | | |

 The p value is 0.130>0.05, so Fisher’s test will be used and capture the results again.

1. Check the option **Descriptives** **table** to show the **Group Descriptives**.

| Group Descriptives | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Diet** | | **N** | | **Mean** | | **SD** | | **SE** | |
| Height |  | 1 |  | 24 |  | 171 |  | 10.95 |  | 2.23 |  |
|  |  | 2 |  | 27 |  | 174 |  | 8.90 |  | 1.71 |  |
|  |  | 3 |  | 27 |  | 168 |  | 9.71 |  | 1.87 |  |
|  | | | | | | | | | | | |

1. Show the **Descriptives plots**.



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

| Tukey Post-Hoc Test – Height | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | **1** | | **2** | | **3** | |
| 1 |  | Mean difference |  | — |  | -2.45 |  | 3.03 |  |
|  |  | p-value |  | — |  | 0.650 |  | 0.518 |  |
| 2 |  | Mean difference |  |  |  | — |  | 5.48 |  |
|  |  | p-value |  |  |  | — |  | 0.108 |  |
| 3 |  | Mean difference |  |  |  |  |  | — |  |
|  |  | p-value |  |  |  |  |  | — |  |
|  | | | | | | | | | |

1. Report the test result.

A one-way ANOVA test was conducted to determine if there were significant differences in Height between three different diets.

Height for each diet was normally distributed, as assessed by the Shapiro-Wilk test (p > 0.05)

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

There was not a significant difference in Height at p > 0.05 for the three diets.

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.

| Paired Samples T-Test | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | **95% Confidence Interval** | | | |  | | | | **95% Confidence Interval** | | | |
|  | |  | |  | | **statistic** | | **df** | | **p** | | **Mean difference** | | **SE difference** | | **Lower** | | **Upper** | |  | | **Effect Size** | | **Lower** | | **Upper** | |
| pre.weight |  | weight6weeks |  | Student's t |  | 13.3 |  | 77.0 |  | < .001 |  | 3.84 |  | 0.289 |  | 3.27 |  | 4.42 |  | Cohen's d |  | 1.51 |  | 1.18 |  | 1.83 |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | | | |

1. Perform the assumption check by selecting **Normality Test** and paste the result below.

| Normality Test (Shapiro-Wilk) | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | |  | | **W** | | **p** | |
| pre.weight |  | - |  | weight6weeks |  | 0.990 |  | 0.802 |  |
| Note. A low p-value suggests a violation of the assumption of normality | | | | | | | | | |
|  | | | | | | | | | |

1. Show the **Descriptives table**.

| Descriptives | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **N** | | **Mean** | | **Median** | | **SD** | | **SE** | |
| pre.weight |  | 78 |  | 72.5 |  | 72.0 |  | 8.72 |  | 0.988 |  |
| weight6weeks |  | 78 |  | 68.7 |  | 69.0 |  | 8.92 |  | 1.011 |  |
|  | | | | | | | | | | | |

1. Show the **Descriptives plots**.

图表, 箱线图

描述已自动生成

1. Report the overall test result.

A paired sample t-test was conducted to determine if there were statistically significant differences in weight at the beginning of the diet and the end of it.

As assessed by the Shapiro-Wilk test (p > 0.05), the normality assumption was not violated.

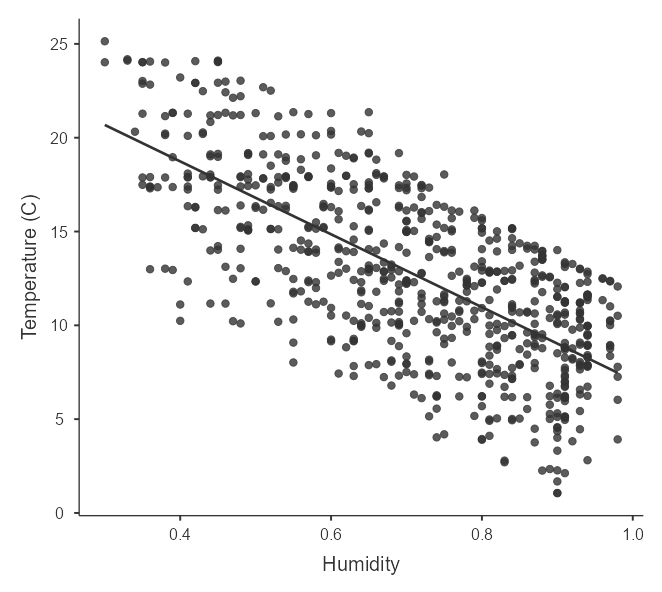
According to the Student’s test, there was a significant difference (p < 0.001) in weight median at the beginning of the diet and at the end.

The median weight at the beginning of the diet is 72.0, compared to 69.0 after the diet. These results support that the diets 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)



1. By using **Correlation Matrix**, find the **Pearson correlation coefficient** between **Temperature** and **Humidity**. Show the **correlation matrix** table below. Include the p-values and flag significant correlations in your table.

| Correlation Matrix | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | **Temperature (C)** | | **Humidity** | |
| Temperature (C) |  | Pearson's r |  | — |  |  |  |
|  |  | p-value |  | — |  |  |  |
| Humidity |  | Pearson's r |  | -0.700 |  | — |  |
|  |  | p-value |  | < .001 |  | — |  |
|  | | | | | | | |

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**).

| Model Fit Measures | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | **Overall Model Test** | | | | | | | |
| **Model** | | **R** | | **R²** | | **F** | | **df1** | | **df2** | | **p** | |
| 1 |  | 0.700 |  | 0.490 |  | 690 |  | 1 |  | 718 |  | < .001 |  |
|  | | | | | | | | | | | | | |

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

| Model Coefficients - Humidity | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | | **Estimate** | | **SE** | | **t** | | **p** | |
| Intercept |  | 1.0306 |  | 0.0131 |  | 78.9 |  | < .001 |  |
| Temperature (C) |  | -0.0252 |  | 9.61e-4 |  | -26.3 |  | < .001 |  |
|  | | | | | | | | | |

1. The formula for predicting humidity is:

Humidity = 1.0306 - 0.0252 \* (Temperature)

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

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

A significant regression equation was found (F = 690, p < .001), with an R2 = 0.490. The humidity is equal to 1.0306 - 0.0252 \* (temperature) when the temperature is measured in Celsius.

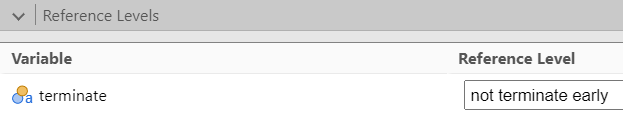
The humidity decrease by 0.0252 for each degree decrease of temperature.

**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.
2. Change the **Measure Type** of the attributes **avoidanceOfDisclosure** and **symptomSeverity** as **continuous**.
3. 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**.

| Model Fit Measures | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | **Overall Model Test** | | | | | |
| **Model** | | **Deviance** | | **AIC** | | **R²McF** | | **χ²** | | **df** | | **p** | |
| 1 |  | 45.1 |  | 49.1 |  | 0.270 |  | 16.7 |  | 1 |  | < .001 |  |
| 2 |  | 35.7 |  | 41.7 |  | 0.423 |  | 26.2 |  | 2 |  | < .001 |  |
|  | | | | | | | | | | | | | |

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



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

| Model Coefficients - terminate | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | | **Estimate** | | **SE** | | **Z** | | **p** | | **Odds ratio** | |
| Intercept |  | -1.139 |  | 1.500 |  | -0.759 |  | 0.448 |  | 0.320 |  |
| avoidanceOfDisclosure |  | 0.356 |  | 0.118 |  | 3.025 |  | 0.002 |  | 1.427 |  |
| symptomSeverity |  | -0.317 |  | 0.123 |  | -2.566 |  | 0.010 |  | 0.729 |  |
| Note. Estimates represent the log odds of "terminate = terminate early" vs. "terminate = not terminate early" | | | | | | | | | | | |
|  | | | | | | | | | | | |

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

| Collinearity Statistics | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | **VIF** | | **Tolerance** | |
| avoidanceOfDisclosure |  | 1.39 |  | 0.720 |  |
| symptomSeverity |  | 1.39 |  | 0.720 |  |
|  | | | | | |

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.

| Classification Table – … | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |  | |
| **Observed** | | **not terminate early** | | **terminate early** | | **% Correct** | |
| not terminate early |  | 21 |  | 4 |  | 84.0 |  |
| terminate early |  | 4 |  | 16 |  | 80.0 |  |
| Note. The cut-off value is set to 0.5 | | | | | | | |
|  | | | | | | | |

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

A binomial logistic regression was carried out to ascertain the effects of avoidance of disclosure and symptom severity to determine the likelihood of early termination from counseling in samples at a community mental health center.

The logistic regression model was statistically significant, X2 (2) = 26.2, p < .001. The model explained 42.3% of the variance in the decision and correctly classified 82.2% of cases.

An increase in avoidance of disclosure score was associated with an increase in the likelihood of termination.

An increase in symptom severity score was associated with a decrease in the likelihood of termination.

[Model: -1.139 + 0.356 \* (avoidanceOfDisclosure) - 0.317 \* (symptomSeverity)]

**Assignment Submission**

Submit the following files to [buelearning](https://buelearning.hkbu.edu.hk/) website:

* lab2-assignment-ans.docx
* diet.omv
* weather.omv
* terminate.omv