Statistical Analysis

Student’s Name

Institutional Affiliation

Course

Instructor’s Name

Date

**1. Briefly describe dataset (1) in the Word document that contains the interpretation of the results (the sentences we gave as a description are enough, the purpose of this is just to make the interpretations whole)! Give the scale of measurement for each variable!**

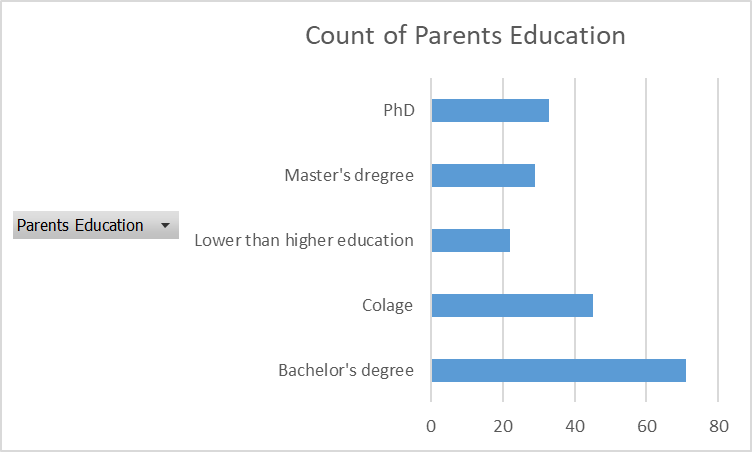
Dataset 1 has 8 variables with 200 rows of data. The names of the variables are, math, history, average, income field, location, parent’s education, and Parent’s position. Math represents students’ math score, history represents students’ history scores, field represents the field that a student is specializing in and average represents students’ high school average score. The income variable is Family income per person in the student's family in thousand francs, parent’s education is representing the highest level of education of parent, and parent’s position is the job position of a parent. The list below indicates the scale of measurement of each variable:

Math: scale

1. History: Scale
2. Average: Scale
3. Income: Scale
4. Field: Nominal
5. Location: Nominal
6. Parents Education: Ordinal
7. Parents Position: Ordinal

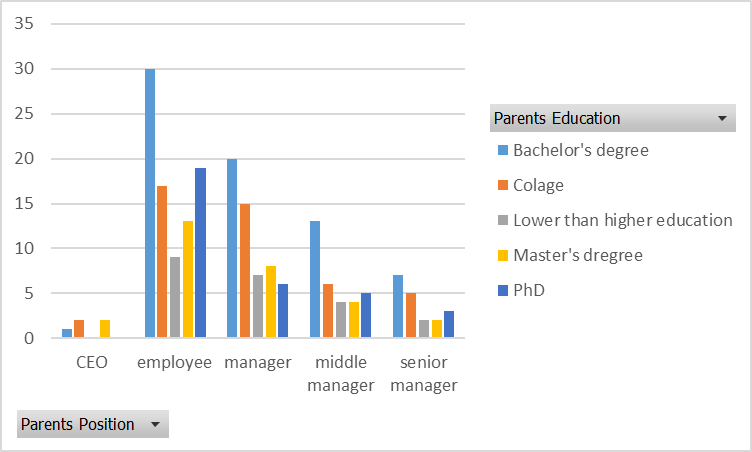
**2. Prepare appropriate tables and charts for the frequency distribution analysis of variable (2). Interpret the results!**

|  |  |
| --- | --- |
| **Parents Education** | **Count of Parents Education** |
| Bachelor's degree | 71 |
| Colage | 45 |
| Lower than higher education | 22 |
| Master's dregree | 29 |
| PhD | 33 |



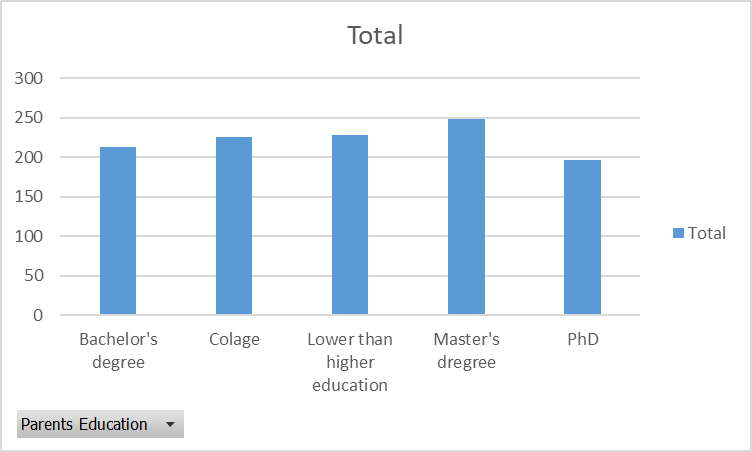
The table and bar graphs shows the distribution of parents on the basis of their education achievement. Majority of the parents have a bachelor’s degree (71 parents) while minority of parents have lower than higher education (22 parents). 45 parents have a college degree, 2m9 have a master’s degrees, and 33 have a PHD.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Count of Parents Position** | **Column Labels** |  |  |  |  |  |
| **Row Labels** | **Bachelor's degree** | **Colage** | **Lower than higher education** | **Master's dregree** | **PhD** | **Grand Total** |
| CEO | 1 | 2 |  | 2 |  | 5 |
| employee | 30 | 17 | 9 | 13 | 19 | 88 |
| manager | 20 | 15 | 7 | 8 | 6 | 56 |
| middle manager | 13 | 6 | 4 | 4 | 5 | 32 |
| senior manager | 7 | 5 | 2 | 2 | 3 | 19 |
| **Grand Total** | **71** | **45** | **22** | **29** | **33** | **200** |



The table and columns graphs shows parent’s position by their level of education. Most CEOs have either a College degree or a Master’s degree. There are no CEOs who have lower than higher education or a PHD. Majority of employees have bachelor’s degree, while only a few of employees have lower than higher education. Similarly, majority of managers have a bachelor’s degree, while only a few managers have lower than a higher education. Most of the middle managers have a bachelor’s degree while the least number of them have college or lower than high school education, and this scenario is similar to senior managers.

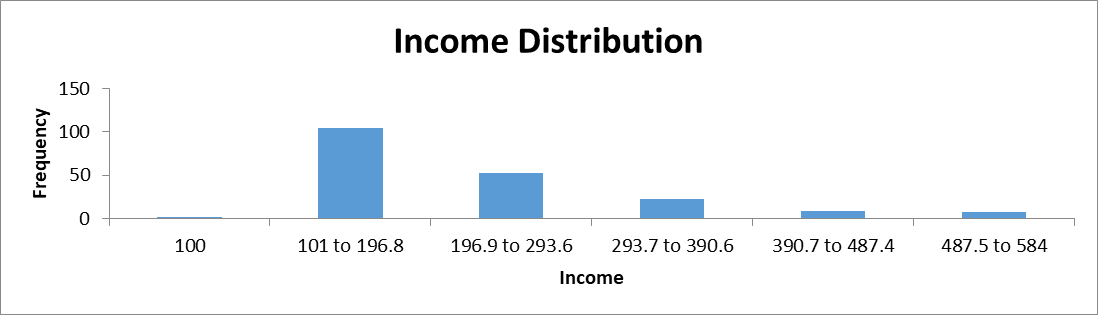
|  |  |
| --- | --- |
| **Row Labels** | **Average of Income** |
| Bachelor's degree | 212.3661972 |
| Colage | 226.1333333 |
| Lower than higher education | 228.4090909 |
| Master's dregree | 248.2068966 |
| PhD | 196.7575758 |
| **Grand Total** | **219.85** |



The table and graph above shows the distribution of average income per parents’ education. Parents who have a bachelor’s degree earn the least average income. The parents’ who have a master’s degree earn the highest average income.

**3. Prepare a binned frequency table for variable (3) and show the corresponding graphic chart! Interpret the results!**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Bin* | *Frequency* | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 100 | 2 | 0.0100 | 2 | 0.0100 |
| 101 to 196.8 | 105 | 0.5250 | 107 | 0.5350 |
| 196.9 to 293.6 | 53 | 0.2650 | 160 | 0.8000 |
| 293.7 to 390.6 | 23 | 0.1150 | 183 | 0.9150 |
| 390.7 to 487.4 | 9 | 0.0450 | 192 | 0.9600 |
| 487.5 to 584 | 8 | 0.0400 | 200 | 0.9600 |
|  |  |  |  | 1.0000 |
| Total | 200 |  |  |  |



The table above is a binned frequency for Income variable while the histogram shows the distribution of income. The interval between one data point to the next is 96.8. The histogram shows that income is skewed to the right. The histogram shows that, as the amount of income increases, the number of parents’ decreases, which implies that only a few parents’ earn higher income than the rest.

**4. Prepare the relative frequency, cumulative frequency and cumulative relative frequency columns as well. Interpret row (4) of every column!**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Bin* | *Frequency* | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 100 | 2 | 0.0100 | 2 | 0.0100 |
| 101 to 196.8 | 105 | 0.5250 | 107 | 0.5350 |
| 196.9 to 293.6 | 53 | 0.2650 | 160 | 0.8000 |
| 293.7 to 390.6 | 23 | 0.1150 | 183 | 0.9150 |
| 390.7 to 487.4 | 9 | 0.0450 | 192 | 0.9600 |
| 487.5 to 584 | 8 | 0.0400 | 200 | 0.9600 |
|  |  |  |  | 1.0000 |
| Total | 200 |  |  |  |

The table contains relative frequency, cumulative frequency, and cumulative relative frequency of income distribution. Row 4 value of the relative frequency is 0.1150 or 11.5%, which implies that the fourth row of the relative frequency represents 11.5% of the data. Row four value of the cumulative frequency is 183 which is the total count from the first row. The fourth row value of **cumulative frequency is 0.9150 or 91.5%, which implies that it represents 91.5% of the data.**

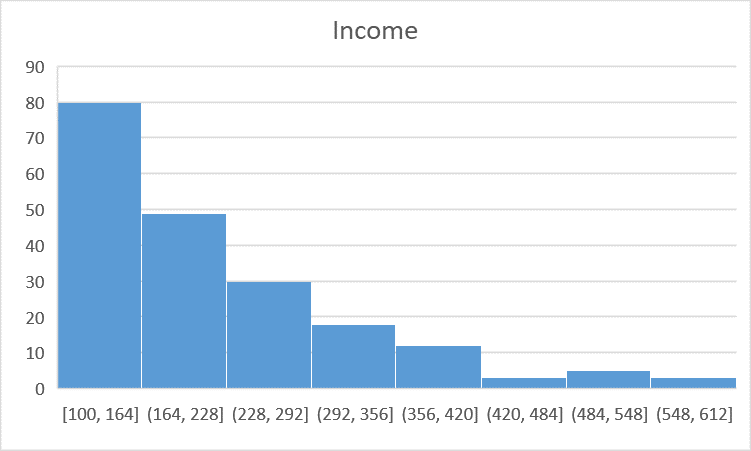
**2nd homework**

**Task J  
Dataset Sz2 - Income**

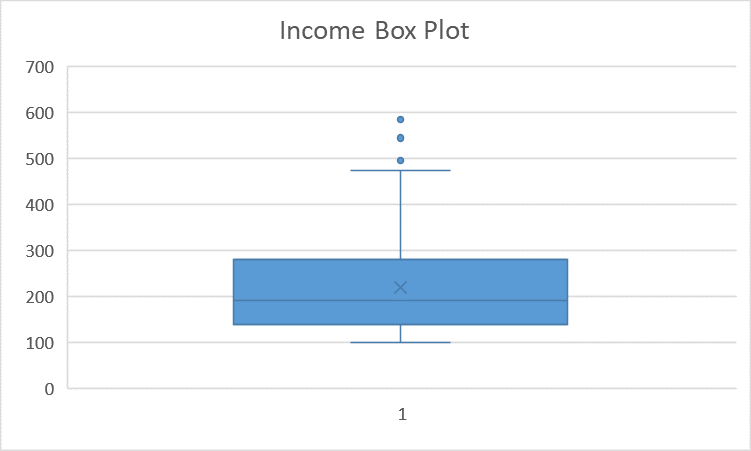
**Perform a descriptive statistical analysis of the variable assigned to your task on the criteria given below as follows::  
1. Using descriptive statistical measures, analyze the variable according to the aspects learned: measures of central tendency, quantiles, measures of dispersion and measures of shape! (With a detailed interpretation!)**

|  |  |
| --- | --- |
| *Income* | |
|  |  |
| Mean | 219.85 |
| Standard Error | 7.529526 |
| Median | 192 |
| Mode | 132 |
| Standard Deviation | 106.4836 |
| Sample Variance | 11338.75 |
| Kurtosis | 1.400475 |
| Skewness | 1.306455 |
| Range | 484 |
| Minimum | 100 |
| Maximum | 584 |
| Sum | 43970 |
| Count | 200 |
| Q1 | 139 |
| Q2 | 192 |
| Q3 | 276.5 |

The table above shows the descriptive statistics of the income variable. The mean income is 218.85 while the median income is 192. The first, second, and third quartiles are 139, 192, and 276.5 respectively. The standard deviation is 104.4836, kurtosis is 1.400475, and skewness is 1.306455.   
2. Use Excel's built-in option (!) to create a diagram characterizing the distribution of the variable and interpret it! How do the findings of the descriptive statistical measures calculated in the previous task appear on the diagram?



The histogram shows that income is skewed to the right.   
**3. Determine the values appearing on the box plot! Prepare and interpret the box plot! Discuss how the findings of task 1 appear on the box plot!**



The Box plot shows that the data has outliers. The minimum value is 100, first quartile is approximately 150, the median approximately 190. The mean value is approximately 220, and the third quartile is approximately 290. The plot shows that the maximum value is approximately 490, which implies that any values above 490 are outliers.

**3rd homework**

1. **In this task, work with the database you worked with in Homework #2 ! Consider your database as a random sample! Choose a quantitative variable that has at least 10-20 different unique values! Give a 95% confidence interval for the mean of the variable. Interpret the result in the context of the dataset!**

|  |  |
| --- | --- |
| *Math* | |
|  |  |
| Mean | 64.415 |
| Standard Error | 1.397814 |
| Median | 64.5 |
| Mode | 78 |
| Standard Deviation | 19.76807 |
| Sample Variance | 390.7767 |
| Kurtosis | -0.74134 |
| Skewness | -0.15808 |
| Range | 74 |
| Minimum | 26 |
| Maximum | 100 |
| Sum | 12883 |
| Count | 200 |
| Confidence Level(95.0%) | 2.756428 |
|  |  |
| Confidence Level | 95% |
| Margin Error | 2.739665 |
| Lower Limit | 61.67534 |
| Upper Limit | 67.15466 |

The table above shows the descriptive statistics and confidence intervals of the mathv scores. The confidence interval for math scores is 61.667534 and 67.15466. The confidence intervals imply that the average math score of student ranges between 61.67534 and 67.15466.   
**2. Search on the Internet for the results of a public opinion poll on the proportion of supporters of a politician! One good datasource could be the site**[**https://www.270towin.com/polls/latest-2020-presidential-election-polls/**](https://www.270towin.com/polls/latest-2020-presidential-election-polls/)**. For example, here you can see poll data on the proportion of supporters of Donald Trump in the USA. However, any other similar data source will do. Enter the link to your chosen data source! Complete the 95% confidence interval estimate for the population proportion. Determine the margin of error where you use the maximum theoretical amount of the standard error and also the margin of error when you calcuate the standard error based on the sample proportion (p)! Interpret the result in the context of your chosen data!**

|  |  |
| --- | --- |
| Total Number of Supporters (Donald Trump) | 3574 |
| Total weighted sample size | 8765 |
| Proportion (p) | 0.407758129 |
| Standard error | 0.00524898 |
| Margin of error | 0.010288001 |
| Confidence Interval | 0.397470128 |
| Confidence Interval | 0.413007109 |

The table above shows the confidence interval from an opinion poll data about Donald trump election. The poll had 8765 voters where 3574 voted for trump. The confidence interval of Donald trump votes is 0.3974, and 0.413. This implies that we can be 95% confident that the true proportion of Donald Trump supporters is between 39.74% and 41.13%.

**4th homework**

***Task J  
Dataset Sz2 - Average score + Location***

**Perform statistical testing of the relationship between the two variables given below as follows:  
1.    Prepare and interpret the table containing the means and standard deviations of the numerical variable in the sub-populations defined by the nominal variable!**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Abroad | Capital | Country |
| Mean | 4.0439 | 4.0322 | 4.05389 |
| Standard Deviation | 0.3822 | 0.3417 | 0.40739 |

Students living in abroad, Capital, and country have mean average score of 4.0439, 4.0322, and 4.05389. Students in country have the highest average scores while those in capital have the least average scores. The standard deviation of average score in abroad, capital, and country is 0.3822, 0.3417, and 0.40739 respectively.   
**2.    Determine the within, between, and total standard deviations and interpret them in the context of the dataset!**

|  |  |
| --- | --- |
| Total Standard Deviation | 0.496470315 |
| Abroad Standard Deviation | 0.3822164 |
| Capital Standard Deviation | 0.341686531 |
| Country Standard Deviation | 0.407389666 |
| Abroad and capital standard devaition | 0.364732183 |
| Abroad and country standard deviation | 0.386335343 |
| Capital and Country Standard Deviation | 0.356340923 |

The table above shows the within, between and total standard deviation of average score. The total standard deviation is 0.4965, which implies that average scores are relatively to each other. Abroad, Capital, and country standard deviations are 0.38221, 0.3427, and 0.4074 respectively. The standard deviations are smaller implying that the average scores in each category is relatively dispersed. The standard deviation of between capital and abroad is 0.3647, between abroad and country, and capital and country are 0.365, 0.386, and 0.356 respectively indicting that the average score between the groups are relatively dispersed.   
**3.    Determine the strength of the relationship between the two variables!**



The p value of the anova result is 0.9730, which is not significant. The p-value indicates that there is no difference in average scores between the groups. Similarly, the F-value is smaller than F critical thus supporting the alternative hypothesis.   
**4.    Consider your dataset as a sample and test the existence of the relationship in the population in the way you have learned (on 5% significance level)! Interpret the result together with the value given in Task 3!**

|  |  |
| --- | --- |
| **Confidence Interval** |  |
| Population mean | 4.1692 |
| Confidence Interval | 5% |
| Margin Error | 0.002201 |
| Lower Limit | 4.166998631 |
| Lower Limit | 4.1692 |

The table shows the confidence intervals of average scores. The confidence interval is 4.166998631, and, 4.1692. therefore, we can be 5% confident that average scores are between 4.166998631 and 4.1692.

**5th homework**

***Task J  
1.Dataset Sz2 - Parents Position + Location  
2.Dataset Sz2 - History Score+Math Score; History Score+Income***

**Perform the testing of the relationship between the two variables given below in the 1st line of your task as follows:  
1.    Prepare the cross table (contingency table) and then the percentage distribution according to one of the variables. How does the relationship between the two variables appear in this latter table?**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Count of Parents Position** | **Column Labels** |  |  |  |  |  |
| **Row Labels** | **CEO** | **employee** | **manager** | **middle manager** | **senior manager** | **Grand Total** |
| abroad | 3 | 30 | 16 | 13 | 5 | 67 |
| captial | 1 | 50 | 30 | 18 | 13 | 112 |
| country | 1 | 8 | 10 | 1 | 1 | 21 |
| **Grand Total** | **5** | **88** | **56** | **32** | **19** | **200** |
|  |  |  |  |  |  |  |
| Percentage Disttribution | |  |  |  |  |  |
|  | **CEO** | **employee** | **manager** | **middle manager** | **senior manager** | **Grand Total** |
| abroad | 60.00% | 34.09% | 28.57% | 40.63% | 26.32% | 33.50% |
| captial | 20.00% | 56.82% | 53.57% | 56.25% | 68.42% | 56.00% |
| country | 20.00% | 9.09% | 17.86% | 3.13% | 5.26% | 10.50% |

The table shows that majority is CEOs live in Abroad. Majority of employees live in capital while the least number of employees live in country. The highest number of managers are in capital while the least are in country. Similarly, the highest number of senior managers are found in capital while the least are found in country.

**2.    Prepare the cross table (contingency table) assuming independence! Use a measure to describe the strength of the relationship between the two variables! Interpret the value of this measure!**

|  |  |  |  |
| --- | --- | --- | --- |
| **Row Labels** | **Count of Parents Position** | **Row Labels** | **Count of Location** |
| CEO | 5 | abroad | 67 |
| employee | 88 | captial | 112 |
| manager | 56 | country | 21 |
| middle manager | 32 | **Grand Total** | **200** |
| senior manager | 19 |  |  |
| **Grand Total** | **200** |  |  |

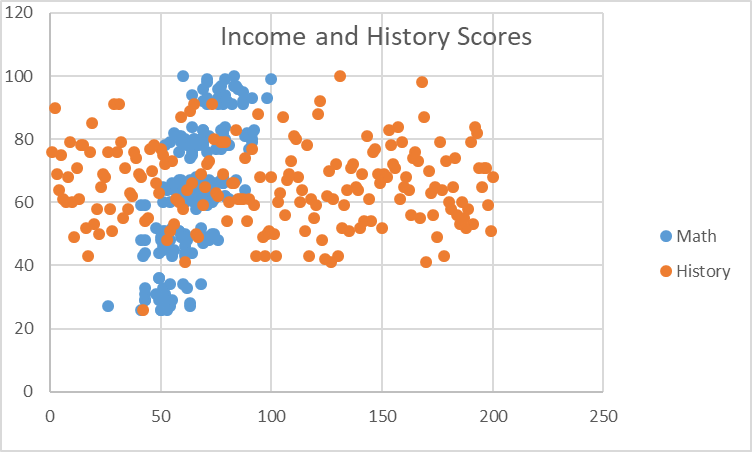
|  |  |  |
| --- | --- | --- |
|  | *Count of Parents Position* | *Count of Location* |
| Count of Parents Position | 1 |  |
| Count of Location | -0.041198827 | 1 |

There are only 5 CEOs, 88 employees, 56 managers, 32 middle managers, and 19 senior managers. 67 participants live in abroad, 112 live in capital and 21 live in country. Count of parents’ position and count of location have a moderately low negative correlation (-0.041198827)  
**3.    Consider your database as a sample and test the existence of the relationship in the unobserved population in the way you have learned! Compare the results with the result of the previous task!  
Perform the testing of the relationship between the two set of variable-pairs given below in the 2nd line of your task as follows:  
1.    Determine the values of the correlation coefficient and the coefficient of determination. Which variable-pair has a stronger relationship?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *History* | | *Math* | |
| History | 1 | |  | |
| Math | 0.677273 | | 1 | |
|  | *History* | *Income* | |
| History | 1 |  | |
| Income | -0.09236 | 1 | |

History and math scores have a moderate positive relationship. History and income have a moderately negative correlation.   
**2.    Interpret the value of the correlation coefficient and the coefficient of determination in the context of the dataset!**

History and math scores have a moderate positive relationship. The correlation implies that if math score increase, also history score increases, or if history score increase also math score increases. When history score increases income decreases.   
**3.    For both variable-pairs, create a scatter (point) diagram and interpret the relationships through this as well!**



The scatter plot shows math scores are concentrated at a particular point while history scores are scattered all over. Besides, history scores are much higher than math scores.

The scatter plot shows the distribution of income per history score. Income is highly concentrated at one point. The scatter plot shows that the data has outliers.