Statistics Assessment

Question 1

Using statistical analysis investigate whether there is evidence that the incidence of cancer

in these 2 cohorts is related to tumor stage at diagnosis, tumor size at diagnosis, age of

patient or recorded ethnicity of patient. Discuss your results and any significant association found. Explain the nature of any significant association.

Answer

For each question you need to start by stating which test is appropriate and why to answer this question. Then you need to test the assumptions for that test. Then state the null hypothesis for that test and level of statistical significance you are using. Only then do you run the test. Once the test is run, you then discuss the numbers generated from the test and state whether the null hypothesis is proven or disproven and continue to answer the question posed.

Quyestion 1 - You have presented nothing useful in answering question 1. You haven’t chosen or run a test. Go back to the tutorials 1-6 and decide in first place which test is appropriate and then do the test(s) in the same way as we did in the tutorial class.

By keeping in mind the above question researcher has made this hypothesis given below;

State null hypothesis

**H1: There is a significant difference between age, ethnicity, tumor stage and tumor size of patients on the basis of area.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Included | | Excluded | | Total | |
| N | Percent | N | Percent | N | Percent |
| Areas \* Age Non Metropolitan | 384 | 9.5% | 3679 | 90.5% | 4063 | 100.0% |
| Areas \* Metropolitan Age | 3835 | 94.4% | 228 | 5.6% | 4063 | 100.0% |
| Areas \* Race Non-Metropolitan | 402 | 9.9% | 3661 | 90.1% | 4063 | 100.0% |
| Areas \* Race Metropolitan | 3801 | 93.6% | 262 | 6.4% | 4063 | 100.0% |
| Areas \* Metropolitan Tumor stage at diagnosis | 3697 | 91.0% | 366 | 9.0% | 4063 | 100.0% |
| Areas \* Non-Metropolitan Tumor stage at diagnosis | 366 | 9.0% | 3697 | 91.0% | 4063 | 100.0% |
| Areas \* Non-Metropolitan Tumor size at diagnosis | 339 | 8.3% | 3724 | 91.7% | 4063 | 100.0% |
| Areas \* Metropolitan Tumor size at diagnosis | 3265 | 80.4% | 798 | 19.6% | 4063 | 100.0% |

The above table indicates the total number of sample and percentage. The researcher firstly applied descriptive statistics to measure the number of participants and percentage according to the variables in both cohorts, Metropolitan and non-metropolitan. It’s indicates the total number of respondents in each group, excluded and included as well as their percentage. The below tables also portrays the mean and standard deviations according to different variables on the basis of area.

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Age Non Metropolitan** | | | |
| Areas | | | |
| Age Non Metropolitan | Mean | N | Std. Deviation |
| Below 50 | 2.0000 | 31 | .00000 |
| 51-70 | 2.0000 | 178 | .00000 |
| 71-80 | 2.0000 | 96 | .00000 |
| Above 80 | 1.7722 | 79 | .42212 |
| Total | 1.9531 | 384 | .21165 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Metropolitan Age** | | | |
| Areas | | | |
| Metropolitan Age | Mean | N | Std. Deviation |
| Below 50 | 1.7641 | 479 | .42501 |
| 51-70 | 1.0000 | 1758 | .00000 |
| 71-80 | 1.0000 | 1016 | .00000 |
| Above 80 | 1.0000 | 582 | .00000 |
| Total | 1.0954 | 3835 | .29386 |

The above table shows the Mean and standard deviations of samples on the basis of age of both groups, metropolitan and non-metropolitan.

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Race Non-Metropolitan** | | | |
| Areas | | | |
| Race Non-Metropolitan | Mean | N | Std. Deviation |
| Black | 2.0000 | 19 | .00000 |
| White | 1.9202 | 376 | .27132 |
| Other | 1.0000 | 7 | .00000 |
| Total | 1.9080 | 402 | .28944 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Race Metropolitan** | | | |
| Areas | | | |
| Race Metropolitan | Mean | N | Std. Deviation |
| Black | 1.7722 | 474 | .41989 |
| White | 1.0000 | 3146 | .00000 |
| Other | 1.0000 | 181 | .00000 |
| Total | 1.0963 | 3801 | .29503 |

The above table shows the Mean and standard deviations of samples on the basis of race of both groups, metropolitan and non-metropolitan.

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Metropolitan Tumor stage at diagnosis** | | | |
| Areas | | | |
| Metropolitan Tumor stage at diagnosis | Mean | N | Std. Deviation |
| Localized | 1.1993 | 1836 | .39962 |
| Regional | 1.0000 | 1568 | .00000 |
| Distant | 1.0000 | 293 | .00000 |
| Total | 1.0990 | 3697 | .29870 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Non-Metropolitan Tumor stage at diagnosis** | | | |
| Areas | | | |
| Non-Metropolitan Tumor stage at diagnosis | Mean | N | Std. Deviation |
| Localized | 2.0000 | 183 | .00000 |
| Regional | 2.0000 | 154 | .00000 |
| Distant | 2.0000 | 29 | .00000 |
| Total | 2.0000 | 366 | .00000 |

The above table shows the Mean and standard deviations of samples on the basis of tumor stage at diagnosis of both groups, metropolitan and non-metropolitan.

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Non-Metropolitan Tumor size at diagnosis** | | | |
| Areas | | | |
| Non-Metropolitan Tumor size at diagnosis | Mean | N | Std. Deviation |
| <2cm | 2.0000 | 156 | .00000 |
| 2-5cm | 2.0000 | 169 | .00000 |
| >5cm | 2.0000 | 14 | .00000 |
| Total | 2.0000 | 339 | .00000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Areas \* Metropolitan Tumor size at diagnosis** | | | |
| Areas | | | |
| Metropolitan Tumor size at diagnosis | Mean | N | Std. Deviation |
| <2cm | 1.2355 | 1554 | .42446 |
| 2-5cm | 1.0000 | 1587 | .00000 |
| >5cm | 1.0000 | 124 | .00000 |
| Total | 1.1121 | 3265 | .31554 |

The above table shows the Mean and standard deviations of samples on the basis of tumor size of both groups, metropolitan and non-metropolitan.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Paired Samples Test** | | | | | | | | | | |
|  | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
| Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Pair 1 | Age Non Metropolitan - Metropolitan Age | 1.58073 | .90473 | .04617 | 1.48995 | 1.67151 | 34.238 | 383 | .000 |
| Pair 2 | Race Non-Metropolitan - Race Metropolitan | .97015 | .25287 | .01261 | .94536 | .99494 | 76.922 | 401 | .000 |
| Pair 3 | Metropolitan Tumor stage at diagnosis - Non-Metropolitan Tumor stage at diagnosis | -.57923 | .63505 | .03319 | -.64451 | -.51396 | -17.450 | 365 | .000 |
| Pair 4 | Non-Metropolitan Tumor size at diagnosis - Metropolitan Tumor size at diagnosis | .58112 | .57182 | .03106 | .52003 | .64221 | 18.711 | 338 | .000 |

The above tables indicates the descriptive statistics applied to all variables. Researcher tries to identify the frequency and percentage difference between variables, which includes, age, stage and size of tumor and race. There is a significant difference found between and within variables on the basis of areas, which were metropolitan and non-metropolitan. The reason could be the heterogeneous sample size. That was not same in both the groups.

Moreover, by keeping in mind the objectives of this question, researcher identified the difference between all variables according to the area. Researcher applied paired sample t-test. Researcher used this test because it is appropriate, as it measures the difference between two different measurements. There was a significant difference and association found between them. That paired differences are independent of each other. The results are highly significant, among all variables. However, the outcomes revealed the alternative hypothesis proven or null hypothesis disapprove. As null hypothesis was made to measure the difference among variables on the basis of areas. Researcher applied paired sample t-test and found pair 1 (*M*=1.58, *SD*=.904) while in pair 2 (*M*=.970, *SD*=.252). However, in pair 3 (*M*=-.579, *SD*=.635), thus pair 4 shows (*M*=.581 *SD*=.571).The alternative hypothesis approved as it seems to be a clear difference among variables, age, size and stage of tumor and ethnicity among both groups.

Question 2

The data in file Q2\_Nov19.xls shows the drop of systolic blood

pressure for each of the patients within each treatment group, following treatment. Using

statistical analysis what can be said about the differences observed for the different drug

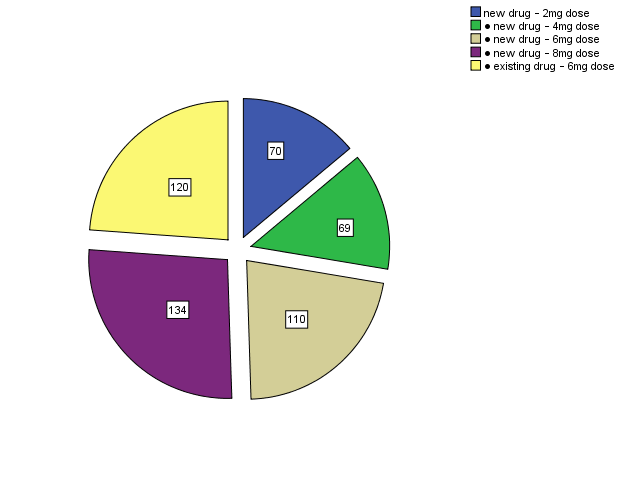
treatments.

Answer

Question 2 - again you need to choose the correct test. Go through the tutorials and decide the correct test for the data in this question. Then proceed to answer the test as described above. Nothing useful presented below.

By keeping in mind the above question and data, researcher applied one sample t-test and state the hypothesis below.

**H2: there is a significant difference in systolic blood pressure level by using different drug dose.**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **One-Sample Statistics** | | | | |
|  | N | Mean | Std. Deviation | Std. Error Mean |
| new drug – 2mg dose | 10 | 7.0000 | 1.82574 | .57735 |
| • new drug – 4mg dose | 10 | 6.9000 | 2.13177 | .67412 |
| • new drug – 6mg dose | 10 | 11.0000 | 2.49444 | .78881 |
| • new drug – 8mg dose | 10 | 13.4000 | 4.50185 | 1.42361 |
| • existing drug – 6mg dose | 10 | 12.0000 | 3.74166 | 1.18322 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **One-Sample Test** | | | | | | |
|  | Test Value = 0 | | | | | |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| new drug – 2mg dose | 12.124 | 9 | .000 | 7.00000 | 5.6939 | 8.3061 |
| • new drug – 4mg dose | 10.235 | 9 | .000 | 6.90000 | 5.3750 | 8.4250 |
| • new drug – 6mg dose | 13.945 | 9 | .000 | 11.00000 | 9.2156 | 12.7844 |
| • new drug – 8mg dose | 9.413 | 9 | .000 | 13.40000 | 10.1796 | 16.6204 |
| • existing drug – 6mg dose | 10.142 | 9 | .000 | 12.00000 | 9.3234 | 14.6766 |

One sample t-test was applied. As a sample is one all are the patients of hypertension, and their systolic blood pressure measurements were taken, the effect of dose was measured between 50 patients. That were divided into 5 groups. 10 in each group. Who were taking different dose of medicine?

Differences observed for the different usage of dose and treatment among patients of hypertension. The difference was significant and can be seen among all groups. But the patients taking high dosage may seems to be better like new drug with 8mg dose. However, researcher used one sample t-test because they want to identify the blood pressure level by use of medicine. Although the patients were taking the different dose of drug but all were the patients of hypertension. So that means it indicates a one sample. The researcher measured the hypothesis and it has approved as alternative hypothesis and null hypothesis was disapproved. Patients are getting better with every dose they were taking but 8 mg dose has significant effect on them (*M*=13.4, *SD*=4.50; t=9.413, p= .000)

Question 3

What can be deduced by statistical analysis about the effect of high versus low protein diet

upon the weight of tumors observed in this study?

Answer

Researcher tries to find out the diet type effect on tumor weight, as the question above indicates that high protein diet helps in reducing tumor weight. By keeping in mind this statement and data researcher state the null hypothesis given below;

What is H3?

**H3: There is a non-significant difference in high/low protein diet on tumor weight.**

|  |  |
| --- | --- |
| **Labels** | **Tumor weight (g)** |
| high protein | 1.709333333 |
| low protein | 1.862222222 |
| **Grand Total** | **2.315894737** |

Researcher firstly, check the variance of both diet types and tumor weight. The above table shows that low protein diet increase the tumor weight while high protein diet enables them in reducing their size.

It is proven by the analysis that, high protein diet may help in lessen the growth of tumor size. Researcher applied Covariance to measure the size of tumor with protein diet according to levels. The results show significant difference between tumor size and high/low protein diet. It has rejected the null hypothesis.

Researcher applied two independent t-test to measure the effect of diet type on patient tumor weight.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Statistics** | | | | | |
|  | Diet Type | N | Mean | Std. Deviation | Std. Error Mean |
| Tumor Weight | Low protein | 10 | 11.5000 | 1.36463 | .43153 |
| High protein | 10 | 9.9600 | 1.30741 | .41344 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | | | |
|  | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
| F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Tumor Weight | Equal variances assumed | .143 | .710 | 2.577 | 18 | .019 | 1.54000 | .59762 | .28444 | 2.79556 |
| Equal variances not assumed |  |  | 2.577 | 17.967 | .019 | 1.54000 | .59762 | .28427 | 2.79573 |

Question 3 – Correct test chosen but no explanation as to why test is appropriate, no assumptions tested, Results misinterpreted.

The above table indicates the results obtained from two independent t-test. Researcher applied this test because there are two types of diet low protein and high protein which they wants to measure the effect on tumor weight. Researcher made the null hypothesis that was approved. There was a non-significant effect/difference found between the low protein/ high protein diet on tumor weight (t=2.577, p= .710). Which has shown a no effect of diet on tumor size.

Question 4

A study was conducted to investigate whether quadriceps muscle strength in men could be explained by the explanatory variables age and/or height. Data was collected for 41 men. The file Q4\_Nov2019.xls details the maximum voluntary contraction of the quadriceps muscle (MVC) for 41 men and is shown along with their height and age.

Using the data contained within this file, determine by statistical analysis which variables have an impact on male quadriceps muscle strength, and the extent of the relationship.

Answer

Researcher wants to examine the two measurements under this question and data set. When there is a word used for relationship, researcher tries to measure the relationship between variables for this we applied pearson correlation, rather than Spiderman correlation which is used for data sets with ranking. Moreover, impact between variables always measured with the test linear regression. That test used to examine the impact of one variable on another. The above question approached to examine the two things one is impact and other relationship between variables. By keeping in mind this data and question researcher stated the two different hypotheses as they has to apply the two different tests for answering the both queries.

H4: There is a significant positive relationship between age, height and MVc of males. What is H4?

You need to approach this question exactly as we did in the tutorial, firstly by choing the correct test and then by doing exactly as we did in the lecture booklet.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
|  | | AGE | Height (cm) | MVC (newton) |
| AGE | Pearson Correlation | 1 | -.338\* | -.417\*\* |
| Sig. (2-tailed) |  | .030 | .007 |
| N | 41 | 41 | 41 |
| Height (cm) | Pearson Correlation | -.338\* | 1 | .419\*\* |
| Sig. (2-tailed) | .030 |  | .006 |
| N | 41 | 41 | 41 |
| MVC (newton) | Pearson Correlation | -.417\*\* | .419\*\* | 1 |
| Sig. (2-tailed) | .007 | .006 |  |
| N | 41 | 41 | 41 |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | |

Researcher applied correlation to measure the connection/association between variables. To determine the extent of relationship, researcher applied Pearson correlation. Researcher identified the significant relation between male quadriceps muscle strength, age and height of males. The relationship was significant but negative in magnitude. It has approved the hypothesis to some extent, for example the hypothesis has been approved as researcher found the significant relation between variables but not on magnitude. Researcher thought that it has positive relation, but found a negative correlation among all variables. That means by increasing in one thing may decrease another or vice versa. There is a significant relation found between age and height of participants (r= -.338\*). The correlation is significant at 0.05 level. However, a significant relationship found between height and Mvc but positive at 0.01 level (r=.419\*\*). Thus the relation was also significant but negative in magnitude at 0.01 level (r= -.417\*\*). The height has a positive association with Mvc of males that means more heighten males has good quadriceps muscle strength.

H4: Age & height has a significant impact on MVC of males quadriceps muscle.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficients** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -465.626 | 460.333 |  | -1.011 | .318 |
| AGE | -3.075 | 1.467 | -.311 | -2.096 | .043 |
| Height (cm) | 5.398 | 2.545 | .314 | 2.121 | .040 |
| a. Dependent Variable: MVC (newton) | | | | | | |

To measure the impact of age and height of males on male quadriceps muscle strength. Researcher applied linear regression. Researcher found a significant impact of age and height on male quadriceps muscle strength. It has proven that male quadriceps muscle strength affected by the age and height of males. This has approved the alternative hypothesis. That showed a significant impact among variables. Age and height is far important and considered good for male quadriceps muscle strength. Age (r=.043) while height with mvc (r=.040).

DON’T do what you have written below. Not relevant here

Here is a summary of tests applied in this assignment, with their description and examples for better understanding of readers.

**Correlation**

It was used to measure the connection/ relationship between variables either they are interlinked or not with each other.

E.g

Relation of two variables like height and weight of body are interlinked.

**One sample t-test**

One sample t-test is a statistical procedure used to examine the mean difference between the sample and the known value of the population mean.

E.g

A group of one sample but there are variations, like students in one classroom belongs to different socio economic background.

**Two independent sample t-test**

The independent t-test, also called the two sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.

E.g

A difference between male and female on the basis of religion. A two defined types of groups differentiate on one particular thing.

**Paired sample t-test**

Paired Sample T-Test. The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations.

E.g

A pairs of friends and they’re differentiated in class presentation on the basis of topic, marks and grades.

**Regression**

Simple regression is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known. ... People use regression on an intuitive level every day.

E.g

Something have an impact on other, like sun may burn the skin.

**Descriptive statistics**

Descriptive statistics are brief descriptivecoefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

E.g

Total number of participants, their percentage, mean an average among them and standard deviation. That basically describes the numbers and values in quantity.

**Covariance**

Covariance measures the directional relationship between the returns on two assets. A positive covariance means that asset returns move together while a negative covariance means they move inversely.

E.g

Two different types of thing covariant on particular item. Like patients level of pain is different but they all are taking medicines accordingly.