Transcript for Statistical analysis presentation

Hello everyone, My presentation is about statistical analysis of the data taken from Health Survey performed in 2011 for England, using R Studio software.

Slide 2:

The Health survey England is a periodic survey on a sample of people to monitor the health trends and estimate the occurrence of risk factors associated with certain health conditions. The recent reports of the Office of National Statistics and National Health Service information centre report that excessive alcohol consumption leads to increased alcohol related hospital admissions, increased alcohol specific deaths, increased cost and use of prescription drugs to treat alcohol dependence and increase in road causalities involving illegal alcohol levels

Slide 3:

The aim of this presentation is to present the statistical analysis results and interpretation of the key findings from the health survey 2011 dataset which was published on 20th December 2012. The dataset is in the format .sav file which can be easily imported in R. The analysis includes descriptive statistics, graphical representation of variables and inferential statistics including hypothesis testing and interpreting results using p-value. R Studio software is used for analysis, as it is free and open source resource and has good data visualisation tools like pie charts, histograms, box plots, scatter plot. Also it has inbuilt functions and libraries providing easy and quick results

Slide 4:

This table shows the proportion of some of the variables in the sample data set.

It is evident that more than three quarters of people in the sample data drink alcohol, more than half where women and nearly a quarter had highest education level

Slide 5:

The bar chart of the variable marital status displays the data graphically, representing the relative frequency of each group in the sample data. From the chart, it is very clear that most of the people in the sample data were married.

Slide 6:

For the three variables, household size, B.M.I. which is the body mass index and age at last birthday, this table shows the descriptive statistics, measures of central tendency such as the mean, median and mode and measures of variability, such as standard deviation range minimum and maximum. It is noted that the statistics mean median mode for the variables household size and age at last birthday are almost similar.

Slide 7:

This is a histogram created for the variable B.M.I. which is useful to examine the normality in the data. Here normality refers to the distribution of data that has equal data values on both sides of the mean. In this histogram long tail in the right side means that the data is skewed to the right.

Slide 8:

Boxplot are used for continuous data variable. Here the age variable is displayed graphically representing the five figure summary statistics such as the minimum,

First quartile, median, third quartile and maximum. Also boxplot shows normality of data about the median.

Slide 9:

To find out which gender drinks more alcohol now-a-days, contingency table with gender against drinking status was created and its significance tested using Chi square test. This test is used because the variables gender and drinking status are in the form of counts. From the contingency table it was noted that more proportion of male drink alcohol compared to female. p-value obtained with chi square test was lesser than 0.005 which shows very highly significant difference that more male drinks alcohol compared to female.

Slide 10:

Similarly to find out which region drinks more alcohol now-a-days, contingency table with different regions in England against drinking status was created and significance tested using Chi square test. From the table it is clear that high proportion of people in South west region drink compared to other regions. p-value obtained was lesser than 0.005 which shows very highly significant proportion of people in South West region drinking alcohol compared to other regions

Slide 11:

Independent two sample t-test was performed to find if there is any statistical difference exists between men and women on their mean height. Null hypothesis defined as their mean height difference equal to zero, while alternative hypothesis defined as their mean difference not equal to zero. p value obtained from the t-test is less than 0.005 shows very highly significant difference in mean heights between

men and women. Hence the null hypothesis is rejected and the alternative hypothesis is failed to be rejected that the difference in mean height between men and women is not equal to zero.

Similarly Independent two sample t-test were performed to find if there is any statistical difference in mean weight between men and women, p value obtained is less than 0.005 shows very highly significant difference in mean weights between men and women. And therefore null hypothesis is rejected.

Slide 12:

Correlation is a statistical measurement of the relationship between two variables. Pearson correlation was calculated between the variables, drinking now-a-days status, total household income, Age and Gender to measure their relationship. From the obtained r values, it was found that positive correlation exists between drinking now-a-days and gender, between drinking now-a-days and age and between drinking now-a-days and total house-hold income. Poor correlation between Total household income and gender

Slide 13:

Wilsnack et al. 2018 and Chaiyasong et al. 2018 had reported that the percentages of high-volume drinking and high-frequency drinking were greater in men than women. With the given health survey data, using the variable drinking now-a-days status, the results showed that more male proportion drinks alcohol compared to female. In addition testing of the variable total units of alcohol/week against gender is required to see which gender drinks high-volume of alcohol.

In conclusion, great care should be taken in choosing the right variable and right statistic test for analysis to avoid misinterpretation of test results and would recommend R as a great tool for statistical analysis and graphical representation of data.

Slide 14:

References for this presentation are listed here.

Slides 15 to 21:

All R screenshots relevant to the analysis shown in this presentation are included here.

Thank you for listening