International Journal of Academic Pedagogical Research (IJAPR)
ISSN: 2643-9123
Vol. 6 Issue 7, July - 2022, Pages: 84-122

TEACH YOURSELF SPSS AND STATA

Kamugisha NELSON[1], Friday Christopher[2], Ndyamuhaki Milton[3]

[1] Bachelor of Economics &Statistics, Kyambogo University, Uganda.
[2] Assistant Lecturer/PhD Student, Department of Political and Administrative Studies, Kampala International University, Email: fridaychristopher@rocketmail.com
[3] Physics/Maths Teacher, Kirima Community Secondary School, Kanungu district, South Western Uganda

Abstract: *SPSS is the statistical package for social scientists. The SPSS Corporation initial created the SPSS software system package within the early 1980's and has recently discharged version Twenty two. It is one of the usually used applied mathematical packages. The package is simple to use, its familiarity to several applied math consultants and its practicality. Statistics is that the body of mathematical techniques or processes for gathering, describing organizing and decoding numerical information. Since analysis usually yields such quantitative information, statistics could be a basic tool of measure and analysis. The researcher who uses statistics is bothered with additional than the manipulation of knowledge, applied math ways goes back to basic functions of study. SPSS is used all-purpose survey analysis package, and therefore easy to understand. It helps in analyzing large volumes of quantitative data into manageable forms.*

Keywords: SPSS, Social Sciences, STATA

ACKNOWLEDGEMENT

We appreciate the Library staffs of Kampala International University and Kyambogo University in the resource centre, library and documentation centre for their utmost cooperation during the compilation of this book.

Special thanks go to the lecturers of Kampala International University and the graduates of Bachelor of Economics and Statistics-Kyambogo University for their assistance rendered to us in accessing all the relevant information needed for the success of this study.

Finally, heartfelt thanks go to our dear parents, brothers and sisters, beloved wives and children for being there for us during the period we were in pursuit of this great asset. May God bless you abundantly, above all we thank God for bringing us all this far.

.

Contents

Introduction

SPSS stands for Statistical Package for the Social Scientists. It is general statistical software tailored to the needs of social scientists. Compared to other software, it is more intuitive and easier to learn; the trade-off is less flexibility and fewer options in advanced Statistics than some other statistical software like S-Plus, R and SAS. SPSS is good for organizing and analyzing data. The reader needs to first install SPSS software on his/her computer. It is an extensive package with facilities for data entry, data Manipulation and statistical analysis. It has modules for survey analysis, graphical display and time series.

Basic structure of SPSS

There are two different windows in SPSS

1st – Data Editor Window – this consists mainly of data view and variable view.

Data view

The Data view displays data for each variable.

Variable view

Variable view displays the specifications for each variable in a SPSS data set and is used for creating and modifying the variables, it is where variables are designed from.

Saving contents of data window

- Click on file
- Select save or save as
- Assign a filename and click save

2nd – Output viewer Window

This displays all your results after a command is processed/ shows results of data analysis

Saving contents of output window

- Click on file
- Select save or save as
- Assign a filename and click save
- An extension .spo by default will be assigned
- You must save the data editor window and output viewer window separately.
- Ensure that you save both windows if you want to save your changes in data or analysis.

Note: Ensure that you know your file location (where you are saving work) each time you are saving.

In the Data view:

Rows are cases

Columns are variables

In the Variable view:

Rows define the variables

Variables. A variable is any characteristic, number, or quantity that can be measured or counted such as age, sex, income and country.

Data input

SPSS presents the data in two views: data and variable Data view Looks like Excel. Each row is a “case”, e.g. person. Each column is an “attribute” or variable, e.g. height, age, gender, place of birth. Data

Types of variables

Quantitative/continuous variables

These have values that describe a measurable quantity such as a number like 'how many' or 'how much ‘such as height, age, time and weight.

Qualitative/categorical variables.

These have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category’. These variables are usually coded with values assigned to different categories in that variable E.g. Education level (1- primary, 2-secondary and 3-tertiary), Gender (1- male and 2- female) and place of residence (1-urban and 2-rural).

Getting started with SPSS

There are two ways to launch the SPSS program. One is to simply click on the SPSS icon shown in red letters on your desktop. If you cannot find the icon, you can click Start on the bottom of your screen, then Program Files, and then SPSS for windows, When the SPSS window launches, a dialogue box will pop up as shown

Click on SPSS 15.0 for windows

You have several choices; you can either type in data, or open an existing file.

Click on type in data Click ok Note. Click on type data when you have not yet entered any data and open an existing file if you have some data already entered and saved in SPSS.

The screen appears as below

1 2 Arrow 1 (data view) is where data is entered and in arrow 2 (variable view) is where we design variables/ define variable names from. In the variable view, the window appears as below

For each variable supply the following information, name, type, width, decimal, places, value, missing value code and measures

Name >>> type a short name. It should not be more than 8 characters and it should not have spaces. Each name should be unique e.g gender, district Type>>> it refers to the specific kind of the data being entered. Some common examples of Types are: Numeric (numbers), Date (dates) and String (letters, numbers and characters with, or without, spaces).

Click ok Width >>>type the variable width e.g 2(width is adjustable) Decimal>>>type the places of decimal e.g 0 or 1 or2 Labels >>>describe the name in details e.g if name is gender, its label can be gender of the student or if the variable name is age, then its label can be age of the respondent Value >>>should be specified every time for categorical variables to ease data entry e.g 1- male and 2-female Missing value code- >>>codes which are not of interest in data analysis e.g if you have sex and codes are 1- male or 2-female but a respondents says he is both Measures >>>Measure is the nature of the kind of data that is being entered. Measure has three options: Scale (measurable data such as age and income), Ordinal (ordered data such as education level (primary, secondary and tertiary)) and Nominal (data with no apparent linear sequence/ random such as place of residence and countries).

Entering and editing data.

Once you have defined all your variables, proceed to type your data in the data view as you would do in a spreadsheet. Place the cursor in the cell where you want to enter data and then type the value you want and press the Enter key. This will place the value in the cell and move down to the cell below. You can also move from cell to cell using the arrow keys, the Tab key or the mouse. Example Enter the following data in SPSS Age Major Year 19 1 2 22 2 4 21 3 3 18 1 1 20 3 2

Where; 1= art, 2=statistics and 3=economics

The variable view would look like this

The data view would appear as follows

- Saving Data .
- To save a dataset; Navigate to File >>> Save As…,
- Browse to locate your file and type in the desired filename.

Opening Data

To open a file in SPSS, navigate to File >>> Open >>> Data. Browse for the desired dataset, select and click the open button; If the dataset is not visible make sure that the directory is correct and the proper file format (SPSS (*.sav), Excel (*.xls, *.xlsx, *.xlsm) etc…) is selected.

DATA MANIPULATION

Merging Data Files To add cases-ensure that the variable names are the same in the both files

- Open the first data set
- Click data>>> merge files>>> add cases
- Browse/click on the file you want to add
- Click on open
- Click ok

Example

Enter the following data in SPSS. Save the file as student 1

Age Major Year 19 1 2 22 2 4 21 3 3

Enter the following data in SPPS. Save the file as student 2

Age Major Year 18 1 1 20 3 2

Merge the two files

To add variables-ensure that the new variables we are adding are different from those we already have

- Open the first data set
- Click data>>> merge files>>> add variables
- Browse/click on the file you want to add
- Click on open
- Click ok

Exercise

Enter the following data in SPSS. Save the file as student 1

Age Major Year 19 1 2 22 2 4 21 3 3

Enter the following data in SPSS. Save the file as student 3

District Sex Arua 1 Gulu 2 Mbale 1 Where 1=male and 2=female Merge the two files

Selecting a subset of data You can select a subset of data and analyze it for example; what is the average age of female respondents? You have to first select females before computing their average age.

You can select a subset by using the:

- Boolean operator and (symbolized &). The & symbol tell SPSS that a case has to meet two specific criteria to be included in the analysis
- The Boolean operator OR (symbolized by |). The | symbol tell SPSS that it should include a case if it meets either of the two criteria.

selecting a subset of data;

- Click on Data
- Select cases
- Select radio button if condition is satisfied

- Click if command button
- Enter condition for selection of subset of data e.g sex =2
- Click continue
- Click ok Example

Age Major Year Sex District 19 1 2 1 Arua 22 2 4 2 Gulu 21 3 3 1 Mbale For years, 2=fresher, 3=junior and 4=senior

Analyze statistics for female junior

- Click on Data
- Select cases
- Select radio button if condition is satisfied
- Click if command button
- Type(sex =2) & (year=3) in the popup window
- Click continue
- Click ok Going back to original/main data set
- Click on Data
- Select cases
- Click reset
- Click ok

Sorting data

- Click on Data >>> sort cases
- Select sorting variables
- Select sorting order
- Click ok

Data transformation

Computing New Variable

Example 1 .

Using the example above, compute the age in months

- Click on Transform >>> Compute Variable
- In the "Target Variable" box , enter the name of the variable that is being computed
- click on the "Type & Label" button to define the characteristics of the new variable
- Now enter the formula for the computed variable in the numeric expression field using any combination of the variables on the left, the calculator, the functions or just typing an expression with a keyboard(Age*12).
- Click OK to compute the variable.

Example 2

Calculate the student's year of birth, based on their age

Recoding Variables: This is done to change a variable into another basing on required transformation.

Example1

1. Change the variable major to major recode using the above data where

1->3

2 to 3->4

- When recording variables there are two options, either recoding into the same variable or recoding into a different variable
- It is recommended to almost always recode into a different variable in order to ensure that no recorded data will be lost.

- To recode a variable, go to Transform >>> Recode into Different Variables…,
- in leftmost column of the popped up box, select the variable(s) to be recoded and
- click the arrow
- The variable(s) will appear in the middle, Numeric Variable -- >Output Variable, box.
- Enter the name of the new, recoded, variable in the "Name" text box, found enter a label if desired
- Click the "Change" button.
- This will replace the question mark with the new variable's name, showing that the variable on the left of the "-->" will be recoded into the variable on the right of the "-->".
- click on the "Old and New Values"
- A new box will pop up with three sections: "Old Value", "New Value" and "Old --> New".
- Select range in the "Old Value" section

- Enter in the corresponding value(s)/range(s).
- Next select and enter the type and value of how the old values should be recoded in the "New Value" section.
- Click the "Add" button in the "Old --> New" section
- To modify an already existing recode: select the recode in the "Old --> New" box, make the desired changes to the old and/or new value(s) and hit the "Change" button.
- Once the desired recodes are inputted and displayed in the "Old --> New" box.
- Click Continue and then
- Click Ok to finish the recode process.
- The recoded variable will be the last variable in the dataset.

Example 2. Change age into age group such that: 18-19->1 20-22->2

DATA ANALYSIS.

It involves major 5 steps

1. Enter your data in the data editor

2. Select a procedure from the menu

3. Select variables from the analysis

4. Examine the results in the output widow

5. Interpret the results in the word document

NOTE. Before any data analysis is done, first identify whether the variable(s) is/are quantitative or categorical.

It includes univariate analysis, bivariate analysis and multivariate analysis.

For univariate analysis, a single variable is analyzed at a time e.g.

What is the average age of the students?

Bivariate- two variables

Does income of the respondent depend on age?

Multivariate – more than two variables

Does income of the respondent depend on age, education level and sex of the respondent?

UNIVARIATE ANALYSIS

Descriptive statistics.

These are computed only for quantitative/continuous variables such as age, height and weight.

Procedure

- Click on analyze>>> descriptive statistics>>> descriptive
- select variables from the LH box into the RH box
- The user can specify the particular statistics required by selecting 'options' or 'statistics' button
- Click OK
- Interpret the results i.e mean, median, mode, frequency, quartile, sum, variance standard deviations, minimum, maximum, range, kurtosis and skewness.

Example

Using the GSS93 subset data, what is the average and minimum age of the respondents?

Click on analyse >>descriptive statistics>>descriptive

Select age from LH box to RH box

Click ok and the output will appear as below

Descriptive Statistics

1495 18 89 46.23 17.418 303.386 1495

Age of Respondent Valid N (list wise) Minimum Maximum Mean Std. Deviation Variance

The average age of the respondents is 46.23 and the minimum age is 18.

Frequency distribution tables.

These are done for categorical/qualitative variables such as sex, marital status and age group e.g. what percentage of respondents are males and what percentage are females?

Procedure

- Click on Analyze>>> descriptive statistics>>> frequencies
- select variables from the LH box into the RH box
- Additional statistics can be selected by clicking on 'statistics' button
- charts like histogram can be selected by clicking on Charts
- Click OK

Example Use the data below in SPSS windows, what percentage of the respondents are married and what percentage are divorced?

- Click on analyse >> descriptive statistics >> frequencies
- Select the variable marital status to RH box
- Click ok.

Frequency Percent Valid Percent

Cumulative Percent

This shows that 53% of the respondents are married and 14.2% of the respondents are divorced.

NOTE. Always use a column of valid percent as that of percent may involve influence of missing values.

Graphing

A pie chart.

This is done for categorical/qualitative variable such as sex and education level.

Example

For the data given above, using the pie chart, what percentage of respondents are males and which percentage are males?

- Click on analyse>>descriptive statistics,>> frequencies >>
- select the variable sex from LH to RH box and then charts
- select pie chart

The pie chart will be produced

To have percentages on pie chart, double click in the pie chart, then chart editor >> show data labels

Under properties, cross percent up to displayed box and click on apply. Ensure it is only percent under displayed box.

Percentages will appear, copy the pie chart and take it word for interpretation.

In case the pie chart comes when it is halved, right click and edit picture.

Bar Chart

- Click on Graphs>>> legacy Dialogs >>> bar>>> simple>>>define
- Select the categorical variables to be charted
- Click OK

Box Plot

- Click on Graphs>>> legacy Dialogs >>>Box plot>>> simple
- Select summaries of separate variables
- Define
- Select the continuous variables to be charted
- Click OK

Histogram

- Click on Graphs>>> legacy Dialogs >>> Histogram
- Select the variable(s)
- Select display normal curve
- Click OK

Line Graph

- Graph>>> legacy Dialogs>>> line>>> simple
- Select values of individual cases
- Define
- Select the Y and X-axis variables
- Click OK

MULTIPLE RESPONSES/OPEN ENDED QUESTIONS.

This is the case where each respondent gives one or more than one answer to a particular question e.g. what are reasons for high children drop outs in some parts of Uganda?

Method 1/Dichotomies method

Steps

- Data .entry

Best candidate Rigge delection Dictator only

- Click on Analyze

Respondents Response 1 1 2 2,3 3 2,3,4 4 3,4 5 2,3,4 6 2,4 7 3,2 22

- Multiple responses
- Define sets
- Move the desired variables from set definition box to variable set box
- Click on dichotomies counted values
- Put 1
- Go to name- reasons
- Label-why Kagame won
- Click on add
- Close
- Go to analyze
- Multiple responses
- Frequencies

The output appears as below

Interpretation

The above analysis shows that the main reasons for Kagame's win were that he rigged the election and that He is a dictator supported by 33.3% of the responses in either cases. This is followed by reason that he was the only candidate supported by 26.7% of the responses. The other minor reason was that he was the best candidate supported by only 6.7% of the responses.

Method 2/categories method

Steps

- Data entry
- Click on Analyze
- Multiple responses
- Define sets
- Move the desired variables from set definition box to variable set box
- Click on categorical values
- Put 1 through 4 (it depends on the number of reasons you have)
- Go to name- reasons
- Label-why Kagame won
- Click on add
- Close
- Go to analyze
- Multiple responses
- Frequencies

Interpretation is similar as in method 1.

BIVARIATE ANALYSIS

Cross tabulations.

These are only done for categorical variables.

Example Using GSS93 data, which percentage of the males are black and which percentage of those who are black is females?

- Click on Analyze>>>descriptive statistics>>cross tabs
- Select one categorical variable for a row and another for a column
- Click on cells>> rows>>columns>>total >>continue
- Click ok

10.3% of the males are black and 60.7% of those who are black are females. Exercise Using GSS93 subset data, what percentage of the married are graduates and what percentage of high school respondents are divorced?

Correlation analysis**.**

This is a measure of relationship between two variables. It tells us how strong the correlation between the two variables is. The relationship could be negative (-) or positive (+).

If the correlation coefficient ( ) =1, then there is perfect positive correlation between the

variables and. if it is =-1, then there is perfect negative correlation between the variables.

If >0.5, there is a strong relationship between the variables.

If =0.5, the relationship is moderate.

If <0.5, there is weak relationship between the variables.

If <0, then the relationship is very weak.

26

NOTE

In correlation analysis, we analyze the strength, direction and significance of the relationship.

Direction

If the correlation coefficient is negative, it implies the two variables are moving in the different directions, as one variable increases, another one decreases. If the correlation coefficient is positive, it implies that the two variables are moving in the same direction, as one variable increases, another variable also increases.

Significance of the relationship

If the P-value is less than the level of significance such as 0.05, 0.01, then the relationship is statistically significant otherwise it is insignificant.

Testing for correlation

1. Graphical approach. A scatter plot is used.

The scatter plot illustrates relationship between the variables which can be positive, negative or non-existing.

Procedure

- Click on Graphs>>> legacy Dialogs>>>Scatter>>> simple>>>define
- Select the Y and X-axis variables
- Click OK
- To add the line of best fit, double click in the plot and click on add a reference line from equation.

The following are scatter plots for visual interpretations of types of correlations.

Example

Using the data below, Is there any relationship between the variables?

X 2441.1 3776.3 2476.9 3843.1 2503.7 3760.3 2619.4 3906.6 2746.1 4148.5 2865.8 4279.8

Y 2969.1 4404.5 3052.2 4539.9 3162.4 4718.6 3223.3 4838 3260.4 4877.5 3240.8 4821

The scatter plot appears as below

By looking at scatter plot, there is positive relationship between the variables.

2. Statistical Tests

Pearson correlation coefficient. This is used for quantitative variables such as age and income.

For example.

Is there any significant correlation between age and income of the respondent?

The hypotheses are stated as follows

Ho: there is no significant correlation between age and income of the respondent.

Ha: there is a significant correlation between age and income of the respondent.

X

5000 4800 4600 4400 4200 4000 3800

Y

3400 3200 3000 2800 2600 2400

Steps

- Click on Analyze-correlate-bivariate
- Select the variables from the LH box into the RH box
- Select Flag significant correlations
- Select type of correlation coefficient Pearson
- Click ok
- Interpret the output

Example

Using the existing GSS93 subset data, is there a significant relationship between highest year of school completed and age of the respondent?

Following the above procedure, the output below is obtained.

Ho: there is no significant relationship between highest year of school completed and age of the respondent.

Ha: there is a significant relationship between highest year of school completed and age of the respondent.

Interpretation

The correlation coefficient is -0.259, there is a weak negative relationship between highest year of school completed and age of the respondents. This relationship is statistically significant at 1% level of significance since the P-value (0.000) <0.01 thus the null hypothesis is rejected and conclusion made there is a significant relationship between highest year of school completed and age of the respondent.

(i) Spearman -deals with ranked data.

(ii) Kendall's- categorical variables of some order such as education level.

Note. Criteria for rejection.

If the P- value is less <0.05(level of significance), reject the null hypothesis otherwise fail to reject the null hypothesis.

If >2, then the null hypothesis is rejected by rule of thumb otherwise fail to reject the null hypothesis at a given level of significance. If the confidence interval does not include the hypothesized value of the population parameter, the null hypothesis is rejected otherwise it is not.

Chi- square test.

It is a test of dependence or association between two variables which must both be categorical such as marital status, education level, religion etc. Example Does religion of the respondent depend on marital status? (Using GSS93 subset data)

Procedure

- Click on Analyze>> descriptive statistics>>cross tabs
- Select one variable for a row and another for a column.
- Click statistics >> chi square>> continue Cells>>> row and column percentages>>continue
- Click ok

The output appears as below;

Ho: religion of the respondent does not depend on marital status. Ha: religion of the respondent depends on marital status.

Since the p-value (0.00) <0.05, the null hypothesis is rejected implying that religion of the respondent depends on marital status.

Sample t-tests.

These are used for testing/comparing means Sample t-tests in SPSS

- One sample t- test
- Paired sample t-test
- Independent sample t- test
- ANOVA test

Please always remember that:

- One sample t-test is used to compare the mean of one variable from a target value
- Paired sample t- test is used to compare the mean of two variables for a single group
- Independent sample t- test is used to compare means of two distinct groups of cases e.g alive or dead, on or off, men or women etc
- ANOVA is used for testing several means.

One sample t-test One sample t –test is performed when you want to determine if the mean value of a target variable is different from a hypothesized value

Examples

- A researcher might want to test whether the average age for respondents differs from 52
- A researcher might want to test whether the average mark of students differs from 75

Assumptions for the one sample t-test

- The dependent variable is normally distributed within the population
- The data are independent(scores of one participant are not dependent on scores of others)

Steps

- Click on Analyze-Compare Means- one sample T-test
- Enter the hypothesized test value i.e numeric test value against which each sample mean is compared
- Optionally, you can click options to control the treatment of missing data and the level of confidence interval
- Finally, click ok
- Interpret the output

Example1

Is the average age of respondent equal to 50 (Using GSS93 subset data)

Ho: the average age of respondent=50

Ha: the average age of respondent ≠50

Since the P-value (0.000) <0.05, the null hypothesis is rejected thus the average age of respondent ≠50.

Example2

A study on the physical strength measured in kilograms on 7 subjects before and after a specified training period gave the following results.

Subject Before After Diff

1. 100 115, 2. 110 125, 3. 90 105, 4. 110 130, 5. 125 140, 6 .130 140,

7. 105 125

Is there a difference in the physical strength before and after a specified training period?

- State the hypothesis
- Use t-test to find out if there is a difference in the physical strength before and after a specified training period

Solution

First compute a new variable- the difference between the after value and the before value

Steps

- Click on Transform-compute
- For target variable type diff, for numeric expression type after-before
- Click ok
- Analyse-compare means-one sample test
- Select diff as the test variable and test value to be 0
- Click on option and put 95%
- Under missing value select “exclude cases analysis by analysis”
- Click Continue
- Click Ok

Interpretation of output

Ho: there is no significant mean difference in physical strength before and after a specified training period

Ha: there is a significant mean difference in physical strength before and after a specified training period.

Since the P-value (0.000) <0.05, the null hypothesis is rejected implying that there is a significant mean difference in physical strength before and after a specified training period.

The paired sample t-test.

- The paired samples t-test procedure compares the means of two variables for a single group.
- It computes the difference between values of the two variables for each case and tests whether the average differs from 0
- It is usually used in the matched pairs or case- control study

Steps

- Click on Analyze-Compare means- Samples T-test
- Select a pair of variables, as follows
- Click each of two variables. The first variable appears in the current selection group as Variable 1, and the second appears as variable 2
- After you have selected a pair of variables;
- Click the arrow button to move the pair into the paired variables list
- You may select more pairs of variables
- To remove a pair of variables from the analysis;
- Select a pair in the paired variables list and click the arrow button
- Click Options to control treatment of missing data and the level of the confidence interval

Example3

Using example 2 above, the following results are obtained

Interpretation is similar to that of example 2.

Independent Sample t-test.

This is used for testing means of a variable between two groups of cases e.g. is there a significant difference in income between the male and female respondents?

Assumptions for the independent sample t-test

- The variances of the dependent variable in the two population is equal
- The dependent variable is normally distributed within the population
- The data are independent(scores of one participant are not dependent on scores of others)

The independent sample t-test procedure

- It compares means for two groups of cases
- The subjects should be randomly assigned to two groups so that any difference in the responses is due to the treatment or lack of treatment but not to other factors
- Always ensure that the differences in other factors are not making or enhancing a significant difference in mean.

Example

- The researcher is interested to see if in the population men and women have the same scores in a test
- If there is a difference in the highest year of school completed between the males and females.

Steps

- Click on Analyze-compare means-independent-sample t-test
- Select one or more quantitative test variables, a separate t test is computed for each variable
- Select a single grouping variable
- Click define groups to specify two codes for the groups you want to compare
- Click options to control the treatment of missing data and the level of the confidence

Example

Using the GSS93 subset data, is there any significant difference in highest year of school completed between the male and female respondents?

State the hypotheses

Ho: There is no significant mean difference in the highest year of school completed by sex of the respondents.

Ha: There is a significant mean difference in the highest year of school completed by sex of the respondents.

Procedure

- Click on Analyze-compare means-independent-sample t-test
- Select highest year of school completed for test variable
- Select sex for grouping variable
- Click on define groups-use specified values; put 1 for group 1 and2 for group 2. This is because 1 stands for male 1 and 2 stands for female
- Click Continue
- Click ok

The Results show two sets of test statistics

- Equal variance assumed
- equal variance not assumed
- If the F- statistics is significant (the null hypothesis is rejected), we use row of equal variance not assumed for interpreting the t-test
- If the F-statistics is not significant (the null hypothesis is not rejected), we use the row of equal variances assumed for interpreting the t-test

Note. Levene's test helps to determine which row to use to make a decision of accepting or rejecting the null hypothesis.

Output is interpreted as below

Ho: There is no significant mean difference in the highest year of school completed by sex of the respondents.

Ha: There is a significant mean difference in the highest year of school completed by sex of the respondents.

Note: first use Leven's test to identify which row to use to get the correct P-value as shown below.

Ho: There is no difference in the variance.

Ha: There is a difference in the variance.

Since P-value (0.000) <0.05, the null hypothesis is rejected and conclusion made that there is a difference in the variance thus row for equal variance not assumed is used (second row).

Since the p-value (0.107) >0.05, we fail to reject the null hypothesis and conclude that there is no significant mean difference in highest year of school completed by sex of the respondents.

Analysis of Variance (ANOVA) One Way Analysis of Variance

- The one way ANOVA procedure produces a one way analysis of variance for a quantitative dependent variable by a single factor (independent) variable.
- Analysis of variance is used to test the hypothesis that several means are equal.
- This technique is an extension of the two sample t –test.
- In addition to determining that differences exist among the means, you may want to know which means differ.
- Here we can use post hoc tests which are run after the experiment has been conducted.

Assumptions

- Independent random samples:
- That is, the group being compared are regarded as distinct populations, so samples from such population are said to be independent.
- The population are normally distributed
- The population variances are equal

Steps for a one way Analysis of Variance

- Click on Analyze-Compare Means-One Way ANOVA
- Select one or more dependent variables
- Select a single independent factor variable

Post hoc

This is used to find out which groups are significantly different from each other.

- Click on post hoc and select benferroni test

Which group would you recommend?

Under options, click on mean plots or options >>descriptive>> continue

Steps

- Click on Analyze
- Compare means
- Select the dependent variable (continuous/quantitative)
- Select the independent variable (categorical)
- Options-select mean-continue-ok

Example

Using the GSS93 subset data,

- Is there any significant difference in the highest year of school completed (educ) by religion?
- Which religion would you recommend to a respondent and why?
- Which religions are significantly different in effect?

Steps

- Click on Analyse-Compare means-One way ANOVA
- Dependent list (educ)
- Factor (relig)
- Option
- Select post- hoc–benferroni
- Select options, click Mean plots and descriptive
- Click Continue
- Click ok

Interpretation

Answer to question 1

Ho: there is no significant mean difference in the highest year of school completed by different religions.

Ha: there is a significant mean difference in the highest year of school completed by different religions.

Since the P-value (0.000) < 0.05, the null hypothesis is rejected and conclusion made that there is a significant mean difference in the highest year of school completed by different religions.

Answer to question 2

This has two approaches:

Using Descriptive table and means plot

I would recommend the Jewish religion since on average, it has a higher highest year of school completed than other religions.

Religious Preference

Other, None, Jewish, Catholic, Protestant

…….

Answer to question 3

The protestant and Catholics, protestants and Jewish, protestants and others, Catholics and Jewish, those with no religion and Jewish are significantly different from each other since their respective mean differences are statistically significant (P-values<0.05).

In general, two groups of cases are significantly different from each other if their mean difference is statistically significant (P-value is < level of significance such as 0.05 or 0.01).

Question.

In a biological experiment, 4 concentrations of a certain chemical are used to enhance growth of a certain type of plant over a specified period of time. The following growth data in cm were recorded for the plants that survived.

1. 2. 3. 4. 8.2 7.7 6.9 6.8 8.7 8.4 5.8 7.3 9.4 8.6 7.2 6.3 9.2 8.1 6.8 6.9 9.5 8.0 7.4 7.1 9.6 7.8 6.1 7.0

(i) Is there any significant difference in the average growth of these plants for the different concentrations of the chemical?

(ii) Which concentration would you recommend for the type of a plant and why?

(iii) Which concentrations are significantly different from each other?

Regression analysis.

In regression analysis, two sets of variables are considered

Dependent variable

Independent variables

Regression is important in enabling the predictions of the dependent variable given the values of the independent variable using the formulated regression equation

Simple linear regression.

This is used when the variables in consideration are only two and also quantitative in nature.

It always takes the form of;

$Y=\beta0+\beta1X$ + error term/standard error (Є)

Where y= dependent variable/Exogenous Variable

X= independent variable/Endogenous variables $\beta1$ and $\beta0$ = coefficients of the regression

Note.

- Independent variables (X) are variables that drive or determine other variables or relationships
- Dependent variable (Y) are variables that are caused by or influenced by the independent variables
- $\beta0$ is the intercept when the independent variable is not in place or play or are explicitly zero (when X=0)

- β1 is the change in y affected by the change in X
- Є contains other factors that affect the dependent variable e.g seasonality, the product sold

Steps

- Click on Analyze-Regression- Linear
- In the liner Regression dialog box, select a numerical dependent Variable
- Select one numerical independent variables
- Click Ok
- Interpret the output

Interpretation

Start with

R Square (R2)

It explains the goodness of the model or the ability of the independent variable in explaining the variation of the dependent variables.

- If R squared value is less than 0.5, then the model is not good at all. The independent variables are not the only variables that affect the dependent variable.
- The greater the R square, the better the model. If R square is greater than 0.5, the model is good.

Example

If R squared value= 0.3, then it implies that 30% of the variations in the dependent variable can be explained by the independent variable thus a poor fit.

Then interpret the coefficients.

Use standardized coefficient, if the dependent and independent variables are measured in the same units. For example, Y and X are measured in years.

Use unstandardized coefficient, if the dependent and independent variables are measured in different units. For example, Y is measured in months while x is measured in years.

Example

Wage= β0+ β1 Educ+ Є

Wage is in dollars and education is in number of years

Wage=-0.9 +0.54 Educ

An additional year in school would on average lead to 0.54$ increases in the wage keeping other factors constant.

Without education, the worker would on average earn a negative wage/income of 0.9$

Example

Using the GSS93 subset data, does highest year of school completed depend on age of the respondent?

Ho: Highest year of school completed does not depend on age of the respondent.

Ha: Highest year of school completed depends on age of the respondent.

Procedure

- Click on Analyze-Regression- Linear
- In the liner Regression dialog box, select a dependent Variable highest year of school completed.
- Select an independent variable age.
- Click Ok
- Interpret the output

The R-square value =0.067, this implies that 6.7% of the variations in the highest year of school completed can be explained by age of the respondent hence it is a poor fit.

Coefficient

Note that since both variables are measured in the same units, standard coefficients are used.

The coefficient (-0.259) shows that a unit increase in age of the respondent would on average lead to 0.259 decreases in highest year of school completed keeping other factors constant. Since P-value (0.000) < 0.05, the null hypothesis is rejected and conclusion made that, the highest year of school completed depends on age of the respondent.

Multiple linear regression Under this regression, there are more than one independent variable. Both the dependent variable and independent variables are quantitative.

Model specification

If one wants to predict a salesperson's total yearly sales (Dependent variable) from the independent variables such as years of Education(Ed), age and Years of experience

S=f(Ed, A, Ex)

$S= \beta 0+ \beta 1\ Ed+ \beta 2A+ \beta 3Ex+ Є$

Note.

For such a regression, state individual hypotheses.

Example

Use the data below to answer the questions that follow.

Consumption Income Price Age 88.9 57.5 91.7 26 88.9 59.3 92 36 89.1 62 93.1 25 88.7 56.3 90.1 12 88 52.7 82.3 36 85.9 44.4 76.3 25 86 43.8 78.3 14 87.1 47.8 84.3 52 85.4 52.1 88.1 63 88.5 58 88 45

(i) Regress consumption on income, price and age

(ii) Specify the model.

(iii) Interpret all your results.

Ho1: consumption does not depend on income

Ha1: consumption does not depend on income

Ho2: consumption does not depend on price

Ha2: consumption depends on price

Ho3: consumption does not depend on age

Ha3: consumption depends on age

Procedure

- Click on Analyze
- Regression >>linear
- Select the dependent variable and independent variables into RH dialog box as shown below.
- Click ok

The R –square = 0.822, this implies that 82.2% of the variations in consumption can be explained by income, price and age hence it is a good fit.

(ii) Model Consumption=80.074+0.233*income-0.047*price-0.025*age.

(iii) Interpretation When income, price and age of the respondent are equal to zero, then the average consumption is 80.074.

A unit increase in income would on average lead to 0.233 increases in consumption keeping other factors constant .This is statistically insignificant since the P-value (0.07)>0.05, thus we fail to reject the null hypothesis Ho1 and conclude that consumption does not depend on income.

A unit increase in price would on average lead to 0.047 decreases in consumption keeping other factors constant .This is statistically insignificant since the P-value (0.699)>0.05, thus we fail to reject the null hypothesis Ho2 and conclude that consumption does not depend on price.

A unit increase in age would on average lead to 0.025 decreases in consumption keeping other factors constant constant .This is statistically insignificant since the P-value(0. 158)>0.05, thus we fail to reject the null hypothesis Ho3 and conclude that consumption does not depend on age.

Dummy Variable regression In situations where the dependent variable Y depends on qualitative variables as explanatory variables e.g. gender, race, colour, education, religion, profession, nationality, wars.

Attributes of such unquantifiable variables are measured using dummy variables which takes a on the value of 1 to indicate the presence of unquantifiable variable and 0 to indicate the absence.

Dummies Are variables created from other or another variable e.g from Categorical or Groupings

Examples

Categorical

- Sex

1 - Male

2 – Female

- Religion 1-Protestant 2-Catholic 3-Moslem

Grouping

Age

- 15-30 -youth or group 1
- 31-45- Adult or group 2
- 45 and above – group 3

Note

- When introducing “k” dummies, the regressions are estimated without a general constant term or coefficient
- To run it with the constant, always drop one of the dummies

Question

Does the wage of a worker in an organization depend on sex?

$w = \beta 0 + \beta 1\ S$

Ho: Wage does not depend on sex

Ha: Wage depends on sex

Please note that sex is a categorical variable and so we first create dummies for it

Creating a Dummy for Male

$W = \beta 0 + \beta 1\ dmale + Є$

Steps

- Click on Transform
- Record into different variables
- Select sex and put it in the numeric variable in the Output box
- Give output variable name- Dmale
- Type appropriate label- dummy for male
- Click on change
- Click on Old and new values
- The box of recode into different variable box: old and new values,
- Under old value, Click on values and put 1
- Under new values, click on values and put 1
- Click Add
- Repeat the last 3 steps
- under old value, click on values and put 2
- under new values , click on values and put 0
- Click Add
- Click Continue-ok
- The new variable/column Dmale will be generated

Creating a Dummy for female

$W = \beta 0 + \beta 1\ dfemale + Є$

Steps

- Click on Transform
- Record into different variables
- Click on reset (VERY IMPORTANT)
- Select sex and put it in the numeric variable in the Output box
- Give output variable name- dFemale
- Type appropriate label- dummy for female
- Click on change
- Click on Old and new values
- The box of recode into different variable box: old and new values,
- Under old value, Click on values and put 1
- Under new values, click on values and put 0
- Click Add
- Repeat the last 3 steps
- under old value, click on values and put 2
- under new values , click on values and put 1
- Click Add
- Click Continue-ok
- The new variable/column Dfemale will be generated

Example

Does the wage of a worker in an organization depend on sex? (1=male, 2= Female)

wage($) 22 19 18 21.7 18.5 21 20.5 17 17.5 21.2

Sex 1 2 2 1 2 1 1 2 2 1

Create dummies first

Then follow the steps

- Analyze-Regression- Linear
- In the liner Regression dialog box, select a quantitative dependent Variable
- Select one of the dummies (either dmale or dfemale)

Note. When using dummies in a regression, one of the dummies is dropped.

- Click Ok
- Interpret the output

Ho: wage does not depend on sex. Ha: wage depends on sex

Model

Unstandardized Coefficients

……..

The mean wage of the female workers would be $18.

The coefficient for Dmale (3.230) shows that the males are more likely to earn a higher average wage than females. This is statistically significant at 5% level of significance since the P-value (0.00) <0.05.

Work out

Run the regression using the female dummy (dfemale).

Note 2

A regression model containing independent variables that are exclusively dummy or qualitative in nature are called Analysis of Variance (ANOVA) models

- Does the wage of the worker depend on sex?

$W= \beta 0+ \beta 1\ S+Є$

Where dummies variables are used in a model with quantitative independent variables, they are called Analysis of Covariance (ANCOVA) models

- Does the wage of a worker depend on the year of experience and sex?

Example for ANCOVA model

Using the GSS93 subset data, does highest year of school completed depend on age, sex and Number of brothers and sisters (sibs) of the respondent?

Hypotheses

Ho1: highest year of school completed does not depend on age of the respondent

Ha1: highest year of school completed depends on age of the respondent

Ho2: highest year of school completed does not depend on sex of the respondent

Ha2: highest year of school completed depends on sex of the respondent

Ho3: highest year of school completed does not depend on number of brothers and sisters of the respondent

Ha3: highest year of school completed depends on number of brothers and sisters of the respondent.

Steps

- Click on Analyze
- Regression >>linear
- Select the dependent variable and independent variables into RH dialog box as shown below
- Click ok

The output below is obtained

The R-square value=0.125, this implies that 12.5% of the variations in the highest year of school completed can be explained by age, sex and number of brothers and sisters of the respondent thus a poor

Model

Highest year of school completed=15.708 -0.040*age -0.246*sibs+0.190*dmale

Interpretation

A unit increase in age of respondent would on average lead to 0.04 decreases in highest year of school completed keeping other factors constant. This is statistically significant since the P-value (0.000) <0.05, thus the null hypothesis (Ho1) is rejected implying that highest year of school completed depends on age of the respondent.

A unit increase in the number of brothers and sisters would on average lead to 0.246 decreases in highest year of school completed keeping other factors constant. This is statistically significant since the P-value (0.000) <0.05 thus the null hypothesis Ho3 is rejected implying that highest year of school completed depends on number of brothers and sisters of the respondent

The coefficient of dummy for male (0.190) shows that the male respondents are more likely to have a higher average highest year of school completed than the female respondents keeping other factors constant. This is statistically insignificant since the P-value (0.209)>0.05.

2. STATA

Introduction STATA is a powerful statistical and econometric software package with smart data-management facilities, a wide array of up-to-date statistical techniques, and an excellent system for producing publication quality graphs. Stata is fast and easy to use.

Stata is a multi-purpose statistical package to help you explore, summarize and analyze datasets.

Note To be conversant with Stata, use the knowledge of SPSS covered since interpretation of Output is similar.

- Getting started with Stata
- Click on Start>.> all programs>> stata>> intercooled stata 8 Or
- Double Click on the Stata icon on your desktop or other location.

Stata User-interface The Stata user-interface consists of the following elements:

Results window .

All outputs appear in this window. Only graphics will appear in a separate window. This is the command line where commands are entered for execution. Variables window All variables in the currently open dataset will appear here. By clicking on a variable its

name can be transferred to the command window. Review window previously used commands are listed here and can be transferred to the command window by clicking on them. Buttons .

The most important button functions are the following: Open (use):

Opens a new data file.

Save:

Saves the current data file.

• Print results:

Prints the content of the results window.

New Viewer:

Opens a new viewer window, e.g. to open log-files. New Do-file Editor: Opens a new instance of the do-file editor (same as do edit).

Data Editor:

Opens the data editor window (same as edit).

Data Browser:

Opens the data browser (same as browse).

Break: Allows to cancel currently running calculations.

Menu. Almost all commands can be called from the menu. However, we do not recommend to learn Stata using the menu commands since the command line will give the user much better control and allows for a much faster and more exact working process.

Getting data into STATA Entering data directly /typing data from the keyboard .

- Use Data Editor
- Don't type the var names in the first row, first type some values and change the default name
- To change the default var names, double click on the relevant column to open a pop-up dialog box
- Under name ,type the required name e.g age
- Under label ,type in the desired label name e.g the age of the respondents
- Enter the other values as you keep on pressing the entre button each time a new value is entered.

Excel to Stata

(Copying-and-pasting)

- Open the Excel file
- High light and copy data in the excel file
- paste data into the Stata's"Data editor "which you can open by clicking on the icon Press ctrl v to paste the data
- close the data editor by pressing the "X"button the upper right corner of the editor Importing (import *.csv)
- Open the excel file
- File- save as
- Under save as type, select csv(comma separated value)
- You will get the following message, click ok, yes
- To save only the active sheet, click ok
- To keep this format, which leaves out any incompatible features , click yes
- In stata, Click on File>>import-ASCII data created by spreadsheet
- Click on browse to find the file and ok

Saving the file

- Click on File- save as- file name Opening the data Open stata>>Click on file>>open >> find work>>click open Saving output(logging) File >>log>> begin>> file name In the save as type box , change to >> save.

Saving output can be closed or suspended. File>>log>>close or suspend Saving the Stata commands This is done by using the do file editor Looking at the Data File

- Either use Data
- Data browser or the corresponding icon on the toolbar.
- You can scroll to look at any part of the dataset.
- Note: the browser does not allow editing of data. Or
- Type browse in the command window& press enter to look at the data file

To edit data, use the Data editor Or type edit in the command window>>enter.

Note. Stata is case sensitive so use lower case in typing commands.

Listing Type list in the command window & enter This command lists values of variables in data set. Eg. List age. Listing can have conditions attached such as list age if it is greater than 50, list income if sex female, list data from the first to the 20th, list income if sex is male and age is greater than 40 etc

Use the command. List if age>50 List income if sex==2 List in 1/20 List income if sex==1 & age >40

Sorting (Using commands) Type Sort var eg sort age

For descending order Type gsort-var such gsort - income For ascending order. Type gsort + income

Describe General information about the dataset can be retrieved with describe. The command displays the number of observations, number of variables, the size of the dataset, and lists all variables together with basic information (such as storage type, etc)

Codebook The codebook command delivers information about one or more variables, such as storage type, range, number of unique values, and number of missing values. Type codebook and press enter

Summarize The most important descriptive statistics for numerical variables are delivered with the summarize command: e.g. summarize age or sum age To get detailed descriptive statistics, use the command: summarize var, detail tabulate One-way frequency tables for categorical variables can be drawn with the tabulate command: ie tab var name such as tab sex Two-way cross-tables for two categorical variables can be drawn with another version of tabulate: tabulate varname1 varname2 eg tabulate race sex

Renaming variables a) Menu Data- variable utilities-rename variables Under existing variable name, select the variable name you want to change Under new variable name, type the new name

b) Command rename [old name] [new name]

Examples

rename last name last

Rename student status s

Rename sex gender

Variable name BEFORE Variable name AFTER Last name Last Student Status Sex Gender

Add/changing variable labels

a) menu

- Data- labels-label variables
- Under variable- select the variable desired

- Under new variable label- type in the variable name(according to the question)
- This option is used for data imported or copied from excel or ms access

b) Command

Label variable [var name] "Text"

Examples

Label variable last name "last "

Label variable status "status: married or single"

Label variable sex " gender of the respond

Variable label Before Variable label After Last name Last Status : married or single

Defining value labels

- Click on Data
- Labels
- Label values and notes
- Define or modify value labels
- Click Define
- Under the define new label dialogue box, type in the name of the variable value you want to define e.g gender or marital status
- Click OK
- Under the add value dialogue box, type in the value and the text
- E.g 1 in the value box and male in the text box

Note: Defining labels is not the same like creating variables

Assigning labels

- Click on Data
- Labels
- Label values and notes
- Assign value labels to variable
- Under the Variable box, select the desired variable l.e gender
- Under the value label, select the desired variable l.e gender
- Click Ok

Creating new variables

- Click on Data
- Create or change variables
- create new variable
- Under the new variable name, type the desired name e.g age2
- Click create
- Under the expression builder, type in the desired expression e.g age *12
- Assuming you wanted to create a new variable age2 , for age in months
- Click ok, ok or submit

Generating new variables using the Command

Operators and Expressions

The following table shows the standard arithmetic, logical and relational operators you may use in expressions:

Arithmetic

Logical

Relational + add ! not (also ~) == equal - subtract | or != not equal (also ~=) * multiply & and < less than / divide <= less than or equal ^ raise to power > greater than + string concatenation >= greater than or equal

Steps

- generate or (gen for short),
- type in the command
- generate [newvar] = [expression]
- generate age2= age*12

Recoding a variable using command

Example

Recode age into 3 groups using the cancer file given below

…….

Steps

- gen age group = 0
- replace age group =1 if age › =47 & age <=50
- replace age group =2 if age ›=51& age < = 60
- replace age group =3 if age ›=61& age <= 67
- The new variable is called 'agegroup1', created within the recode command
- Run the frequency of agroup1

Merging data files

- You merge when you want to add more variables to an existing dataset.
- Both files must be in Stata format

Using the menu

- Click on Data
- Combine datasets
- Append datasets
- Browse to select the desired file
- Click Ok

Note. Appending data is adding rows and merging data is adding columns

DATA ANALYSIS

UNIVARIATE ANALYSIS

Descriptive statistics.

This includes mean, kurtosis, minimum etc. it is only for quantitative / continuous variables. Using the employee data, compute the average salary and standard deviation Command. Summarize/sum salary Or Click on statistics>>summaries, tables and test>> summary statistics>> summary statistics>> In the variable dialog box, click on salary>>display additional statistics >>ok Hence the average salary is 34419.57 and standard deviation is 17075.66

Frequency distribution table.

It is used for categorical variables.

Command. Tab var eg tab gender

Example.(using employee data that is in SPSS windows). Note. You can easily get work already entered in SPSS into stata by changing the saving type from SPSS into stata version as shown below. Close the file and then open it from Stata What percentage of the respondents are males and what percentage are females?

45.57% of the respondents are females and 54.43% of the respondents are males.

NOTE To take the output to word document for interpretation, print screen>> paint>>paste>>select the output to interpret.

Graphing data

Pie chart .

With the help of a pie chart, which percent of the respondents are managers and which percent are clerks? Click on Graphics>>pie chart Select the variable Slices are distinct values of a variable>>insert the variable

To insert percentages, click on labels>> in the dialog box of label, insert percent

Click ok Copy the pie chart

37.67% of the respondents are managers and 54.26% of the respondents are clerks.

Scatter plot It is for two quantitative variables eg produce a scatter showing a relationship between Education level (educ) and salary of employees. Command

If no line of best fit is required, the command is

graph twoway scatter educ salary or scatter educ salary

Histogram For quantitative variable like salary

Command histogram salary or graph twoway histogram salary

For a qualitative variable like jobcat Command histogram jobcat, discrete frequency

BIVARIATE ANALYSIS

Cross tabulations.

These are for two categorical variables.

For example. What percentage of the female employees are managers and what percentage of the clerks are males? tab gender jobcat, row column

4.63% of the female employees are managers and 43.25% of the clerks are males.

Chi square test.

For two categorical variables e.g.

Does job category depend on the gender of the employee? Command tab jobcat gender, chi2 Ho: job category does not depend on gender of the employee. Ha: job category depends on gender of the employee.

Since the p-value (0.000) <0.05, the null hypothesis is rejected implying that job category depends on gender of the employee.

Correlation analysis Is there any significant correlation between education level (years) and salary of the employee? Command pwcorr educ salary, sig Ho: there is no significant correlation between education level and salary of the employee. Ha: there is a significant correlation between education level and salary of the employee.

The correlation coefficient 0.6606 shows a strong positive correlation between education level and salary of the employee. This means that as education level increases, the salary of the employee increases, the relationship is significant at 5% level of significance Since the p-value (0.00) <0.05, thus the null hypothesis is rejected and conclusion made that there is a significant correlation between education level and salary of the employee.

Sample t- tests

- One sample t- test
- Paired sample t-test
- Independent sample t- test
- ANOVA test

Overview

- One sample t-test is used to compare the mean one variable
- Paired sample t- test is used to compare the mean of two variables for a single group
- Independent sample t- test is used to compare means of two groups of cases
- t-test is used for testing single mean and ANOVA is used for testing several means.

One sample t-test One sample t –test is performed when you want to determine if the mean value of a target variable is different from a hypothesized value

Command

ttest varname == # for example ttest age == 20

or

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- One sample mean comparison test
- Choose the variable name
- Chose the hypothesis mean
- Choose the confidence interval. (Preferred C.I is 95%)
- Click ok

Example 1

Using the employee data set, is the average education level in years (educ) of the employees equal to 15

Ho: the average education level in years (educ) =15

Ha: the average education level in years (educ) ≠ 15

Command t test==15

Since the P-value (0.000) <0.05, the null hypothesis is rejected and we conclude that the average education level in years (educ) ≠ 15

Example2

The data below shows the sugar levels of 11 runners before and after the marathon race.

After 29.6 25.1 15.5 29.6 24.1 37.8 20.2 21.9 14.2 34.6 46.2

Before 4.3 4.6 5.2 5.2 6.6 7.2 8.4 9 10.4 14 17.8

Question .is there a difference in the sugar levels before and after the marathon race?

State the hypothesis

Use t-test to find out if there is a difference in the sugar levels before and after the marathon race.

Solution

First compute a new variable- the difference between the after value and the before value

Using the command

ttest diff == 0

Using the menu

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- One sample mean comparison test
- Select diff
- Set the hypothesis mean to zero
- Click ok

Ho: there is no significant mean difference in sugar levels before and after the marathon race.

Ha: there is a significant mean difference in sugar levels before and after the marathon race.

Since the p value (0.000) is less than 0.05, we reject the null hypothesis and conclude there is a significant mean difference in sugar levels before and after the marathon race.

Note.

Conclusion is the same using the confidence interval and t-value

Looking at the confidence interval, it does not include 0, so we reject the null hypothesis

Looking at the t value, t=7.4602, is greater than 2, we reject the null hypothesis

The paired sample t-test /Two-sample mean comparison test Command

ttest var name1 == var name2

Data is assumed to be paired,

For example ttest after == before

or

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- two sample mean comparison test (unpaired)OR
- mean sample test , paired test
- Select the first variable
- Select the second variable
- Select the confidence interval. (Preferred C.I is 95%)
- Click ok

Example

A group of 10 students were pretested before and after a tutorial and gave the following achievement scores

After 17 16 21 10 10 14 20 22 14 12 Before 14 12 20 8 11 15 17 18 9 7

Is there evidence of any improvement in achievement scores after the tutorial period?

- State the hypothesis
- Use t-test to find out whether there is evidence of an improvement in achievement scores after the tutorial period.

Using the command

ttest after == before

Using the menu

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- mean sample test , paired test
- Select after as the first variable
- Select before as the second variable
- Click ok

Solution

H0: There is no significant mean difference between the after and the before achievement scores

Ha: There is a significant mean difference between the after and the before achievement scores.

Since the p value (0.0031) is less than 0.05, we reject the null hypothesis and conclude that there is a significant mean difference between the after and the before achievement scores. There is evidence of an improvement in achievement scores after the tutorial period since the average achievement score after a tutorial period is greater than the average score before a tutorial period.

Interpretation is similar for confidence interval and t-value.

Looking at the confidence interval, it does not include 0, so we reject the null hypothesis

Looking at the t value, t=3.5553, is greater than 2, we reject the null hypothesis

Exercise

1. Using the example 1, use the paired t test, to show that there is no mean difference in the achievement scores before and after the tutorial period

2. A medical study was conducted to compare the difference in effectiveness of two particular drugs in lowering cholesterol levels. Drug X was given to 8 persons randomly selected while Drug Y was given to the other 8 individuals. After a specified amount of time each person's cholesterol levels was measured again. Use the paired t test, to show that there is mean difference in the effectiveness of the two drugs, X and Y , to lower cholesterol?

Drug X Drug Y 29 26 32 27 31 28 32 27 32 30 29 26 31 33 30 36

Independent Sample t-test

The independent sample t-test procedure

- It compares means for two groups of cases
- The subjects should be randomly assigned to two groups so that any difference in the responses is due to the treatment or lack of treatment but not to other factors
- Always ensure that the differences in other factors are not making or enhancing a significant difference in mean.

command

ttest var name, by(var)

eg ttest age, by(sex)

Menu steps

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- group mean comparison test
- select the desired variable
- Select the group variable
- Choose the confidence interval. (Preferred C.I is 95%)
- Click ok

Example

A question was asked to 20 respondents whether they used internet. 10 respondents said that they used internet while 10 said they never used internet.

Use internet-internet use

0-NO (Respondents who do not use internet)

1-yes (Respondents who use internet)

The 20 respondents were asked how many hours they spent watching TV

The data below shows hours spent watching TV by internet users and non-internet users. Assuming equal variance, is there is a difference in hours spent watching TV by the two groups?

Hours spent watching TV- TVhours

State the hypotheses

Ho: There is no significant mean difference in hours spent watching TV by the internet use

Ha: There is a significant mean difference in hours spent watching TV by the internet use

Command

ttest tv hours , by(interne tuse)

Using the Menu

- Click on Statistics-Summaries tables, test-Classical test of hypothesis
- group mean comparison test
- select TV hours as the test variable name
- Select internet use as the group variable name
- Click ok

Internet users 5, 12, 6, 8, 11, 14, 14, 14, 4, 3

Non internet users 8, 16, 7, 10, 10, 15, 17, 18, 9, 7

Since the P-value (0.1995) > 0.05, we fail to reject the null hypothesis and conclude that there is no significant mean difference in hours spent watching TV by the internet use.

Interpretation is similar for confidence interval and t-value.

- Looking at the confidence interval, it includes 0, so we fail to reject the null hypothesis
- Looking at the t value, t=1.3318, is less than 2, we fail to reject the null hypothesis

Exercise

1. The residents of Orange City complain that traffic speeding fines given in their city are higher than traffic speeding fines that are given in nearby Deland. Independent random samples of the amount paid by residents for speeding tickets in each of two cities over the last three months were obtained. These amounts were,

Orange City 100, 125, 135, 128, 140, 142, 128, 137, 156, 142

Deland 95, 87, 100, 75, 110, 105, 85, 95

Assuming an equal population Variance, is there a difference in the mean cost of speeding tickets in these two cities?

Analysis of variance (ANOVA) One Way Analysis of Variance

- The one way ANOVA procedure produces a one way analysis of variance for a quantitative dependent variable by a single factor (independent) variable.
- Analysis of variance is used to test the hypothesis that several means are equal.
- This technique is an extension of the two sample t –test
- In addition to determining that differences exist among the means, you may want to know which means differ .
- Here we can use post hoc tests which are run after the experiment has been conducted.

Example

Solution

Using employee data,

- Is there any significant difference in salary for the three employment categories?
- Which employment category (job cat) would you recommend for an employee and why?
- Which employment categories are significantly different from each other?

Using the command

- One way salary job cat
- One way salary job cat , tabulate
- One way hour salary job cat , bonferroni

Using the menu

- Click on Statistics-Linear model and related-ANOVA
- One- way ANOVA
- Select salary as the responsible variable
- Select job cat as the factor variables
- Under the multiple comparison test, Select bonferroni test

Answer to question 1

Ho: there is no significant mean difference in salary by employment category.

Ha: there is a significant mean difference in salary by employment category.

Since the P-value (0.0000) < 0.05, we reject the null hypothesis and conclude that there is a significant mean difference in salary by employment category.

Answer to question 2

I would recommend an employee to be a manager since managers have a higher average salary than others.

Answer to question 3

Managers and clerks are significantly different from each other since their mean difference is statistically significant (P-value (0.000) <0.05). Managers and custodians are significantly different from each other since their mean difference is statistically significant (P-value (0.000) <0.05).

Exercise

a) Using the cancer file show that there is a difference between study time and drug types

b) Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. The table below gives the performance in percentages.

Teaching techniques 1 2 3 4 65 75 59 94 87 69 78 89 73 83 67 80 79 81 62 88 81 72 83 85 69 79 76 82 70 90 80 78

(i)Is there any significant difference in performance for the different teaching techniques?

(ii)Which teaching technique would you recommend and why?

(iii)Which teaching techniques are significantly different from each other?

Regression analysis

- We use regression to estimate the unknown effect of changing one variable over another.
- When we run a regression we assume linear relationship between two variables (i.e. X and Y).
- Technically, it estimates how much Y changes when X changes one unit.
- Before running a regression it is recommended to have a clear idea of what you are trying to estimate (i.e. which are your dependents and independent variables).
- A regression makes sense only if there is a sound theory behind it.

Simple linear regression

command

- type: regress [dependent variable] [independent variable]
- both dependent variable and independent variable is continuous/quantitative e,g. Height (y) and age (x)
- Regress y x Using the menu
- Click on Statistic
- Linear models and related
- Liner regression
- Select the dependent variable
- Select the independent variable
- Click Ok/submit Run a regression model given by, $Y = \beta 0 + \beta 1X$ and Interpret your results, where y=study time and x = age
- Study time refers to months to death/ or end of experiment
- Age refers to age of the patient at the time of experiment

Example

Using employee data set,

Does employee salary depend on education level (years)?

Ho: employee salary does not depend on education level (years).

Ha: employee salary depends on education level (years).

Interpretation

The R –squared value=0.4363, this implies that 43.63% of the variations in the salary can be explained by education level (years) hence it is a poor fit.

If the year of education is equal to zero, the average salary would be-18331.18.

An increase in education level by one year would on average lead to 3909.907 increases in salary of the employee keeping other factors constant.

Multiple linear regression Command

type: regress [dependent variable] [independent variable(s)]

Both dependent variable and independent variables are continuous/quantitative e.g expenditure(y) and income earned(x1), prices of the commodity(x2) , commodity type (x3) .

Type: regress y x1 x2 x3

Using the menu

- Click on Statistics
- Linear models and related
- Liner regression
- Select the dependent variable
- Select the independent variables
- Click Ok/submit

Interpret the results.

Example

Does employee salary depend on education level in years (educ) and job time?

Ho1: employee salary does not depend on education level in years (educ)

Ha1: employee salary depends on education level in years (educ)

Ho2: employee salary does not depend on job time.

Ha2: employee salary depends on job time.

The interpretation is similar to that done in SPSS.

Dummy Variable regression Dependent variable is continuous and independent variables are categorical e.g. education, religion, place of residence and gender and agegroups

Creating Dummy

You can create dummy variables by either using recode or using a combination of tab/gen commands:

Using the tab/gen command

- Type : tab variable, gen (variable)
- Using the cancer file in SPSS create dummies for died and drug
- tab drug, gen (drug) or tab drug, gen(drug_dum)
- go to data editor and view all the created dummies
- you will see 3 new columns (drug1, drug2 and drug3) OR (drug_dum1, drug_dum2 and drug_dum3)
- create the label name for each dummy by double clicking on the variables
- when you double click on drug1 or drug_dum1, type in the name i.e dummy for drug type1

Question. Does the study time depend on the drug type?

- Generate the dummies for drug type
- Drop one dummy variable and take it as a comparison group
- Type: drop drug1
- Go to Statistic
- Linear models and related
- Liner regression
- Select the dependent variable (study time)
- Select the independent variables (drug2 drug3)
- Click Ok/submit
- Interpret your results

Ho: study time does not depend on the drug type.

Ha: study time depends on the drug type.

Model Study time= 9 + 5.93*drug2 + 16.36*drug3

If drug2 and grug3=0, the patient on drug type one would on average need 9 months of study time.

The patient on drug type 2 would on average need 14.93 months of study time (9+5.93) ,that is when dgrug3=0

When drug2=0, the patient on drug type 3 would on average need 25.36 months of study time (9+16.36).

Interpretation of coefficients

The coefficient for drug2 (dummy for drug2) is positive implying that patients using drug type2 are more likely on average to need more study time than those using drug1. This is statistically significant at 5% level of significance since the P-value (0.034) <0.05. The interpretation for the coefficient of drug3 (dummy for drug3) is similar to that of coefficient for drug2.

Since the F-probability is statistically significant (0.000 <0.05), the null hypothesis is rejected and conclusion made that study time depends on the drug type.

Exercise Does study time depend on age groups? Using the cancer file create new variable age groups where 47-50=1, 51-60=2 and 61-67=3

USE STATA AND SPPS

Exercises on Simple Regression

A company sets different prices for a particular DVD system in eight different regions of the country. The accompanying table shows the numbers of units sold and the corresponding prices (in hundreds of dollars)

Sales 420 380 350 400 440 380 450 420

Prices 5.5 6.0 6.5 6.0 5.0 6.5 4.5 5.0

a) Find and interpret the coefficient of determination for the regression of DVD system sales on price

b) Estimate the linear regression of sales on price

c) What effect would you expect a $100 increase in price to have on sales

Exercises on Multiple Regressions

Use the data below to develop a liner model that predicts annual profit margins as a function of revenue per deposits and the number of offices.

Savings and loan Association Operation Data

Revenue per Dollar

Number of Office

…….

Exercise on Dummy variables for regression models

The president of investors' ltd wants to determine if there is any evidence of wage discrimination in the salaries of male and females' financial analysts.

Gender Years of experience

Annual Salary 2 5 36730 2 7 40650 2 9 46820 2 10 50149

2 14 59679 2 17 67360 1 5 51535 1 7 62289 1 9 72486 1 10 75022 1 14 93379 1 17 105979 Where 1= male and 2= female

a) Run a regression of annual salary as a function of years of experience and gender

b) Interpret the estimated coefficients

Note. For ANCOVA models, the procedure and interpretation is similar to that previously done in SPSS.

REFERENCES

Kirkpatrick, Lee A., 1958-. (2013). A simple guide to IBM SPSS statistics for versions 20.0 & 21.0. Australia ; Belmont, CA :Wadsworth,