BIOSTATISTICS (BIO401)

CURRENT FINAL TERM PAPER FALL 2023(16-FEB)

In regression equation Y= a+bx y is called is.

Y = the variable that you are trying to predict (dependent variable).

In regression equation Y on the X is

In a regression equation, where Y is the dependent variable and X is the independent variable, Y is commonly referred to as the "predicted" or "dependent" variable. It represents the variable that you are trying to predict or explain based on the values of the independent variable, X. The equation Y = a + bX represents the relationship between the dependent variable (Y) and the independent variable (X) in a linear regression model.

What is regressor and regressand?

Regressor: It's a variable that is used to predict or explain the value of another variable.

Regressand: It's the variable that is being predicted or explained by the regressor.

What is Excess Kurtosis?

Excess kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals 3. Therefore, the excess kurtosis is found using the formula below

$$\text{Excess kurtosis} = \text{Kurtosis} - 3$$

Pearson's first & second Coefficient?

$$\text{Pearson's First Coefficient} = \frac{Mean - Mode}{Standard\ deviation}$$

ove formula gives you Pearson's first coefficient. Division by the standard deviatior
u scale down the difference between mode and mean. This will scale down their valu
: of -1 to 1. Now understand the below relationship between mode, mean and median.

$$Mode = 3(Median) - 2(Mean)$$

uting this in Pearson's first coefficient gives us Pearson's second coefficient an
a for skewness:

$$\text{Pearson's Second Coefficient} = \frac{3(Mean - Median)}{Standard\ deviation}$$

Difference between SPSS and excel

SPSS:

- Statistical software designed for advanced data analysis.
- Specialized for statistical functions and data management.
- Ideal for complex statistical analyses and research.
- Has a steeper learning curve.

Excel:

- Spreadsheet software with basic statistical functions.
- General-purpose tool for data manipulation and analysis.
- Suitable for simple statistical tasks and general data handling.
- User-friendly interface, widely used for diverse purposes.

Marginal Error:

This refer to the error or discrepancy associated with a specific marginal distribution or marginal effect in statistical analysis. For instance, in regression analysis, the marginal effect of a predictor variable on the outcome variable might have an associated error or uncertainty.

Margin of Error

The margin of error is a statistic expressing the amount of random sampling error in the results of a survey. The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a census of the entire population.

When to use Partial Correlation?

You should use Partial Correlation in the following scenario:

1) You want to know the relationship between two variables
2) Your variables of interest are continuous
3) You have covariates

Assumptions of Partial correlation

The assumptions for Pearson Correlation include:

1) Continuous
2) Normally Distributed
3) Linearity
4) No Outliers
5) Similar Spread Across Range
6) Covariate(s)

Assumptions for Bernoulli Trials

The three assumptions for Bernoulli trials are:

1) Each trial has two possible outcomes: Success or Failure. We are interested in the number of Successes X (X = 0, 1, 2, 3,...).
2) The probability of Success (and of Failure) is constant for each trial; a "Success" is denoted by the letter p and "Failure" is q = 1 − p.
3) Each trial is independent; The outcome of previous trials has no influence on any subsequent trials.

Difference between qualitative and quantitative variables give examples.

Types of Variable

1. Quantitative Variable:

it is the one that can be measured and expressed numerically.

1) ***Discrete variable:*** it is the ones that have gap or jumps in the possible values. These gaps indicate the absence of values between particular possible values of our variable. It is also known as countable values. ***E.g:*** no of pages in book, teeth per child in a school, children in a family.
2) ***Continuous variable:*** It is the ones that take on each and every possible value within an interval. Don't possess gaps in their possible values. ***E.g;*** height of a person, weight of a person.

2. Qualitative variable

It is the one that can not be measure in a numerical form. These are non-numeric variable where each possible value is a category of a variable that's why they are also known as categorical variable or attribute. ***E.g.*** brand of PC, hair color, marital status, category of different types of animals etc.

Write down the types of sample.

There are two possible ways to ensure that the selected sample is representative.

- Random sample or probability sample:

The selection of units in the sample from a population is governed by the laws of chance or probability. The probability of selection of a unit can be equal as well as unequal.

- Non-random sample or purposive sample:

The selection of units in the sample from the population is not governed by the probability laws. ***For example,*** the units are selected on the basis of the personal judgment of the surveyor. The persons volunteering to take some medical test or to drink a new type of coffee also constitute the sample on non-random laws.

Types Of Probability Distribution

1) Simple random sampling

2) Stratified sampling
3) Systematic sampling
4) Cluster (area) sampling
5) Multistage sampling

What are Point Estimators? Write down its characteristics.

Point estimators are functions that are used to find an approximate value of a population parameter from random samples of the population. They use the sample data of a population to calculate a point estimate or a statistic that serves as the best estimate of an unknown parameter of a population.

- **Properties of Point Estimators**

The following are the main characteristics of point estimators:

1. Bias

The bias of a point estimator is defined as the difference between the expected value of the estimator and the value of the parameter being estimated. When the estimated value of the parameter and the value of the parameter being estimated are equal, the estimator is considered unbiased.

Also, the closer the expected value of a parameter is to the value of the parameter being measured, the lesser the bias is.

2. Consistency

Consistency tells us how close the point estimator stays to the value of the parameter as it increases in size. The point estimator requires a large sample size for it to be more consistent and accurate.You can also check if a point estimator is consistent by looking at its corresponding expected value and variance. For the point estimator to be consistent, the expected value should move toward the true value of the parameter.

3. Most efficient or unbiased

The most efficient point estimator is the one with the smallest variance of all the unbiased and consistent estimators. The variance measures the level of dispersion from the estimate, and the smallest variance should vary the least from one sample to the other. Generally, the efficiency of the estimator depends on the distribution of the population. For example, in a normal distribution, the mean is considered more efficient than the median, but the same does not apply in asymmetrical distributions.

Differentiate b/w biased and random error.

Biased errors are consistent and push measurements consistently in one direction, while random errors are unpredictable variations that can occur in measurements or observations.

Write Type of Relationship and Level of measurements.

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Partial Correlation

Partial Correlation is used to understand the strength of the relationship between two variables while accounting for the effects of one or more other variables. Your variables of interest should be

continuous, be normally distributed, be linearly related, and be outlier free. In addition, your variables should have a similar spread across their individual ranges.

Partial correlation explains the correlation between two continuous variables (let's say X1 and X2) holding X3 constant for both X1 and X2.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Steps for Hypothesis Testing for ρ/How to test correlation with hypothesis?

- **Step 1: Hypotheses**

First, we specify the null and alternative hypotheses:

Null hypothesis H0: ρ=0

Alternative hypothesis HA: ρ≠0 or :ρ<0 or :ρ>0

- **Step 2: Test Statistic**

Second, we calculate the value of the test statistic using the following formula:

$$\text{Test statistic:} \quad t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- **Step 3: P-Value**

Third, we use the resulting test statistic to calculate the P-value.The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.

- **Step 4: Decision**

Finally, we make a decision:If the P-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y.

"If the P-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y."

Probability distribution types & function

1. Normal Distribution:

- ***Function***: dnorm(x, mean, sd) - calculates the probability density function (PDF) for a given value x, mean, and standard deviation.

2. Binomial Distribution:

- ***Function***: dbinom(x, size, prob) - calculates the probability mass function (PMF) for a given number of successes x, number of trials size, and probability of success prob.

3. Poisson Distribution:

- ***Function***: dpois(x, lambda) - calculates the probability mass function (PMF) for a given number of events x and the average rate of events lambda.

4. Exponential Distribution:

- ***Function***: dexp(x, rate) - calculates the probability density function (PDF) for a given value x and the rate parameter.

5. Hypergeometric Distribution:

- ***Function***: dhyper(x, m, n, k) - calculates the probability mass function (PMF) for a given number of successes x, population size m, number of successes in population n, and sample size k.

SPSS

What is SPSS?

☐Originally it is an acronym of Statistical Package for the Social Science but now it stands for Statistical Product and Service Solutions ☐One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions

SPSS is a Windows based program that can be used to perform data entry and analysis and to create tables and graphs. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. SPSS is commonly used in the Social Sciences and in the business world, so familiarity with this program should serve you well in the future. SPSS is updated often. This document was written around an earlier version, but the differences should not cause any problems.

SPSS is widely used for?

SPSS is widely used by the market researchers, health researchers, survey companies, Government, Education researchers, marketing organization's and data miner's for statistical analysis. In short SPSS is widely used statistical software in research community.

The Four Windows of SPSS

1) Data editor
2) Output viewer
3) Syntax editor
4) Script window

1. Data Editor:

The data view is used to store and show your data. It is much like an ordinary spreadsheet although in general the data is structured so that rows are cases and the columns are for the different variables that relate to each case. *Spreadsheet-like system for defining, entering, editing, and displaying data. Extension of the saved file will be* **"sav."**

2. Output Viewer:

This window is used to show the results that have been output from your data analysis. Depending on the analysis that you are carrying out this may include the Chart Editor Window or Pivot Table Window. *Displays output and errors. Extension of the saved file will be* **"spv."**

3. Syntax Editor:

This window shows the underlying commands that have executed your data analysis. If you are a confident coder this is where you can amend the code, or write your own from scratch, and then run your own custom analysis on your data set. *Text editor for syntax composition. Extension of the saved file will be* **"sps."**

4. Script Window:

Provides the opportunity to write full-blown programs, in a BASIC-like language. Text editor for syntax composition. *Extension of the saved file will be* **"sbs."**

- **Managing data in SPSS:**

Three key elements to data management processes include:

1. Importing data into SPSS;

2. Labelling data (variables and values); and

3. Sorting and merging data.

Window Sheets

- **There are two sheets in the window:**

1. Data View window: This sheet is visible when you first open the Data Editor and this sheet contains the data.

2. Variable view Window: This sheet contains information about the data set that is stored with the dataset. Click on the tab labeled Variable View.

The variable view contains the variables on your data set, so it defines the properties of your dataset. Each row will define all of the various variables for one set of data. For example, for a numerical piece of data this would show (amongst other things) the number of decimal places that are stored for that piece of data. The variables include - name, type, width, decimals, label, values, missing, columns, align and measure. Ensuring that the 'measure' of your variables is correct is vital. The variable can be Nominal which is for strings of data, Ordinal for data that isn't continuous but can be ranked or ordered or, finally, scale which is used for a variable that is continuous, for example a distance to somewhere.

a) ***Name:*** ☐ The first character of the variable name must be alphabetic ☐ Variable names must be unique, and have to be less than 64 characters. ☐ Spaces are NOT allowed.
b) ***Type:*** ☐ Click on the 'type' box. The two basic types of variables that you will use are numeric and string. This column enables you to specify the type of variable.

c) ***Width:*** ☐Width allows you to determine the number of characters SPSS will allow to be entered for the variable.
d) ***Decimals:*** ☐Number of decimals ☐It has to be less than or equal to 16.
e) ***Label:*** ☐You can specify the details of the variable ☐You can write characters with spaces up to 256 characters.
f) ***Values:*** ☐This is used and to suggest which numbers represent which categories when the variable represents a category.

- **The basic analysis of SPSS that will be introduced in this class**

Frequencies ☐This analysis produces frequency tables showing frequency counts and percentages of the values of individual variables.

Descriptives ☐This analysis shows the maximum, minimum, mean, and standard deviation of the variables

Linear regression analysis ☐Linear Regression estimates the coefficients of the linear equation

How to save the data in SPSS?

To save the data file you created simply click 'file' and click 'save as.' You can save the file in different forms by clicking "Save as type."

In other words When attempting to save files from the SPSS software, it is important to first remember that the only information that is saved is what it in the current window. For example, if the currently displayed window contains the output from an anaylsis (frequency tables, t-test results, graphs, etc.), the only information that will be contained in the resulting file is the output. The information from the Data Editor will not be saved. In order to save the information from the Data Editor, it must be active window by selecting Window at the top of the screen and then selecting the Data Editor. Once the Data Editor has been made the actice window, it can be saved. The following illustrations provide a guide to saving Data and Output files in SPSS.After the data has been entered into the Data Editor, it can be saved by selecting File and Save or Save As.

After selecting Save or Save As, the Save dialog box will appear. The name of the data file can be entered in the box labeled File name and the directory in which the file is saved can be changed by selecting the down arrow on the right side of the box labeled Save in.

Pie Chart

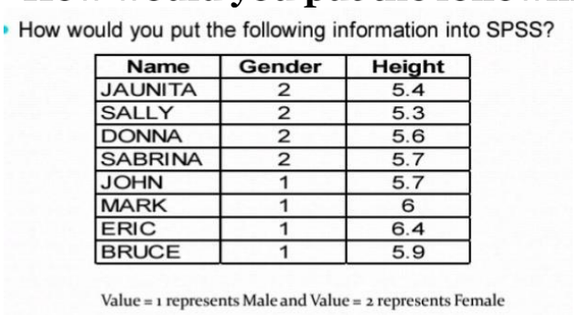
A pie chart is a circular graph that represents data as slices of a pie. Each slice represents a category or a proportion of the whole. The size of each slice is proportional to the quantity it represents. Pie charts are commonly used to show the distribution or composition of different categories within a dataset. They are visually appealing and make it easy to understand the relative proportions of each category.

How to insert pie chart in SPSS?

Here are the steps to create a pie chart in SPSS:

- Open your SPSS dataset.
- Go to "Analyze" > "Descriptive Statistics" > "Frequencies".
- Select the variable you want for the pie chart.
- Click "Charts" and choose "Pie" in the "Chart Builder" dialog box.
- Drag and drop the variable into the "Pie Chart" area.
- Customize the chart if needed by clicking on the "Options" button and click "OK" to generate the pie chart.

How would you put the following information into SPSS?


How would you put the following information into SPSS?

Name	Gender	Height
JAUNITA	2	5.4
SALLY	2	5.3
DONNA	2	5.6
SABRINA	2	5.7
JOHN	1	5.7
MARK	1	6
ERIC	1	6.4
BRUCE	1	5.9

Value = 1 represents Male and Value = 2 represents Female

- Open SPSS and create a new data file or open an existing one.
- Click on the "Variable View" tab.
- Create variables for Name, Gender, and Height.
- Assign a value of 1 for Male and 2 for Female in the Gender variable.
- Enter the corresponding values for each person's Name, Gender, and Height.
- Save the data file.
- Use SPSS's statistical tools to analyze the data.

How to add variance and kurtosis in SPSS?

- ☐Click ‘Analyze,’ ‘Descriptive statistics,’ then click ‘Descriptives…’
- ☐Click ‘Educational level’ and ‘Beginning Salary,’ and put it into the variable box.
- ☐Click Options
- ☐The options allows you to analyze other descriptive statistics besides the mean and Std.
- ☐Click ‘variance’ and ‘kurtosis’
- ☐Finally click ‘Continue’
- ☐Finally Click OK in the Descriptives box. You will be able to see the result of the analysis.

State the shape of distribution:

X is the frequency distribution mean is 32 and median is 30,tell the shape:

X is the frequency distribution mean is 30 and median is 32,tell the shape:

X is the frequency distribution mean is 30 and median is 30,tell the shape:

- **Mean is 32, Median is 30:**

This suggests a negatively skewed distribution, where the tail on the left side is longer or fatter than the right side.

- **Mean is 30, Median is 32:**

This indicates a positively skewed distribution, where the tail on the right side is longer or fatter than the left side.

- **Mean is 30, Median is 30:**

In this case, the distribution is approximately symmetric, as the mean and median are equal. The shape is close to being normal or bell-shaped.

Aik yah short question tha statement di thi aur pocha tha ky shape of distribution batao statement main tha if data of median is 30 = 32 the shape…. If median is 32=32 then shape …. If median 32= 30 then distribution shape…

“Less than” (30 < 32) suggests a negative skew.

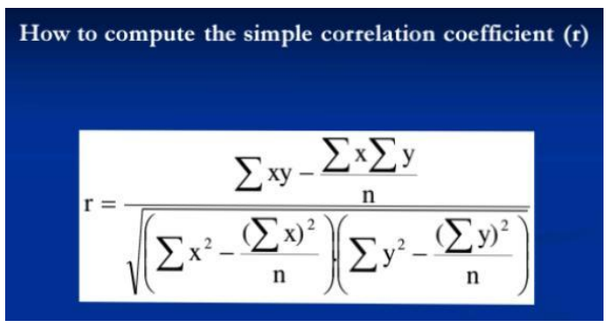
“Equal to” (32 = 32) indicates symmetry.

“Greater than” (32 = 30) suggests a positive skew.

How to Find the Correlation?

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

How to find Simple Correlation Coefficient?



A club ha 5 members. how many ways to select 3 members as president, secretory and …?

A club consists of four members. How many ways are there of selecting three officers: president, secretary and treasurer? It is evident that the order, in which 3 officers are to be chosen, is of significance. Thus there are 4 choices for the first office, 3 choices for the second office, and 2 choices for the third office. Hence the total number of ways in which the three offices can be filled is $4 \times 3 \times 2 = 24$.

The same result is obtained by applying the rule of permutations:

$$
\begin{aligned}
{}^{4}P_{3} &= \frac{4!}{(4-3)!} \\
&= 4 \times 3 \times 2 \\
&= 24
\end{aligned}
$$

Let the four members be, A, B, C and D. Then a tree diagram which provides an organized way of listing the possible arrangements, for this example, is given below:

ek m tha k three persons a, b, and d and suppose they want to take photo

Suppose that there are three persons A, B & D, and that they wish to have a photograph taken.

The total number of ways in which they can be seated on three chairs (placed side by side) is

$${}^{3}P_{3} = 3! = 6$$

Or 1 tha k committee of 3 person 10 group hu to kitny possible ways hn groups bnany k

EXAMPLE

A three-person committee is to be formed out of a group of ten persons. In how many ways can this be done?

Since the order in which the three persons of the committee are chosen, is unimportant, it is therefore an example of a problem involving combinations. Thus the desired number of combinations is

$$
\begin{aligned}
\binom{n}{r} &= \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!\,7!} \\
&= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \\
&= 120
\end{aligned}
$$

In other words, there are one hundred and twenty different ways of forming a three-person committee out of a group of only ten persons!

Prepared By KAINAT ALVI
