ALPHA R PROJECT SERIES A

INTRODUCTION TO STATISTICS AND PROBABILITY

Wilfred Mambina

7/12/2020

INTRODUCTION

Alpha R project is a project that consist statistics and probability knowledge using R software with well summarized explanations and illustrations.This project targets the individuals who have pasion in data analytics,data science and machine learning using R.Therefore the project will consist of materials,tricks and illustrations that are suitable for beginners in this field.Remember R software is commonly applied in data science,data analytics and machine learning field hence a need for mastering it. The project will consist of for Series:

1: Series A; Introduction to Statistics & Probability.

2: Series B; Get started with R for beginners

3: Advanced level in R

4: Developing of R package

OBJECTIVES

1. To demonstrate competence with R software both in basics and advanced level, which includes: Downloading and installation of latest R studio. Finding and installation of r packages. Loading and working with data.

2. To explain the basics in Statistics and Probability.

3. To carryout data visualization & Interpretation of the output using R.

TOPIC 1: INTRODUCTION TO STATISTICS AND PROBABILITY

Terminologies

Distribution-This is a framework describing data patterns that includes location, spread and shape.

Population-This is a group of all physical or geographical characeristics of interest to the researcher.

Sample-this is the subset of units that have been selected.

Sample size-described as the number of units in the sample.

Sample frame-the set of units that have a non-zero probability (chance) of being selected.

Estimator-this is a stichastic function of the samplwe data with the aim to come close to the estimant.

Statistic-this is a numeric or nominal characteristics that describes a sample.

Variable-this is a population characteristic that we are interested in study.

Parameter-this is a numeric or norminal characteristics of a population.

Target population-is the total finite population about which we require information.

Study population-this is the basic finite set of individual you wish to study.

Data -this is the information obtained by way of observation or experiment/ these are raw facts that are collected, organized and summarized in a meaningful way to understand a given situation.

Applied Statistics: this is a science that involves collection of data, organization, summarizing, analyzing and interpretation of information to gain knowledge from the available data.

Aspects of data

1. **Accuracy of the measurement**: refers to how close the experimental result is to the true or actual value in the population of interest and it's normally expressed in terms of error. 2**.** **Precision: this** is how close the experimental measurements are to each other and its usually expressed in terms of deviation i.e. standard deviation, range etc.

Measurements of scales

This is a crucial concept for distinguishing between different types of variables

1. Nominal scale

Also referred to as categorical variable. This is a scale in which there is no particular relationship between the different possibilities, for this kind of variables it doesn't make sense that one is superior compared to one another. *Example* Eye colour-blue,brown etc Gender-Male/Female Religion-Muslim,Christian Fruits-Orange,Mango,Lemon Binary digits-(0,1)

Explanation;Suppose I was doing a research on how pple commute to and from work.One

variable I would have to measure would be ,what jind of transportation people use to get to work? Is it train, bus, car, bicycle etc.

2. Ordinal Scale This is a scale in which there is a natural, meaningful way to order the different possibilities but you can't do anything else.There is an inherent ranking in the existing categories of a variable *Example* Salary scale Police rank Disease status Professionalism

3. Interval scale this is a scale in which the differnces between the numbers are interpretable but the variable doesn't have a natural zero value. An example;Measuring temperature in degrees celsius.For instance if it was 25 degrees celsius yesterday and 28 degrees celsius today then the 3 degrees celsius difference between them is meaningful. However, note that the zero degrees celsius doesn't mean no temperature at all! It actually mean the temperature at which water freezes. It's unacceptable for one to multiply and divude temperatures.For instance its wrong to say 30 degrees celsius is three times as hot as 10 degrees celsius.

Ratio scale In this case zero really zero (meaningful) and its ok to multiply and divide. *For example* a good psychological example is the response time (RT) in a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question because it's an indicator of how difficult the task is.

Assessing the reliability of a measurement

The **reliability of a measure** tells you how precisely you are measuring something whereas validity of a measure tells you how accurate the measure is. **Reliabity**; refers to the repeatability or consistency of your measurement

For example

1. The measurement of my weight by means of a “bathroom scale” is very relable.If I step on and off the scales over and over again it will keep giving me the same answer.

2. Measuring my intelligence by means of asking my parents is very unreliable.Some days they tell I’m a bit thick and other days she tells me I’m a complete moron.

Measure of reliability

1. **Test-retest reliability**

It relates to consistency overtime if we repeat the measurement at later date, do we get the same answer?

2. **Inter-rater reliability** this relates to consistency across people.If someone else repeats the measurements.

3. **Parallel forms reliability** it relates to consistency across theoretically equivalent measurements.

4. Internal consistency reliability

Types of validity

1. **Internal validity** this is the extent to which you are able to draw the correct conclusions about the causal relationships between variables.

2.**External validity** This is the extent to you expect to see the same pattern of results in real life

3.**Construct validity** This is basucally a question of whether you are measuring what you want to be measuring.

4. **Face validity** Refers to whether or not a measure looks like it's doing what it's supposed to, nothing else.

5. **Ecological validity** the entire set up of the study should closely approximate the real world scenario that is being investigated.

Kurtosis This is a statistical measure that is used to describe the degree to which scores cluster in the tails or the peak of a frequency distribution.

Types of kurtosis 1.**Mesokkurtic**; these are distributions that are moderate in the breadth and curves with a medium peaked height.

2. **Leptokurtic;** More values in the distribution tails and more values close to the mean i.e sharply peaked with heavy tails.

3.**Platykurtic;**Fewer values in the tails and fewer close to the mean i.e.the curve has a flat peak and has more dispersed scores with lighter tails.

NOTE: What does it mean when the kurtosis is zero? When the kurtosis is equal to 0, the distribution is mesocratic, this means the kurtosis is the same as the normal distribution.

Types of Skewness

1. **Symmetrical distribution**; In this case the value of mean, median and mode coincide (mean=median=mode).The spread of the frequencies is the same on both sides of the centre point of the curve.

2. **Positively skewed distribution**; the value of mean is maximum and that of mode least, the median lies in between the two (mean>median>mode)

3. **Negatively skewed**; the value of mode is maximum and that of the mean is least, the median lies between the two (mean<median<mode)

The Meaning and interpretation of Confidence interval this is the probability that a population parameter will fall between two set valuess for a certain proportion of times. For example, the interpretation of a 95% confidence interval is that if you repeat an experiment (collect a sample) repeatedly, then 95% of the computed 95% confidence interval will contain the parameter value you are measuring.

Basic of a Test statistic

A test statistic is a random variable that is calculated from sample data and used in hypothesis test. Test statistics can be used to:

1. Determine whether to reject the Null hypothesis.

2. Calculate the P-Value.

3. Compare the data with what is expected under the Null hypothesis

4. Measure the degree of agreement between a sample of data and the Null hypothesis.

Basics of P-values and their interpretation

The P-value is a number (decimal value) between 0 and 1 that indicates the probability of the observed result.Its interpretations is as shown below:

1. A small P-Value (<0.05) indicates strong evidence against the Null hypothesis, so you reject the null hypothesis.

2. A large P-Value (>0.05) indicates weak evidence against the Null hypothesis, so you fail to reject the Null hypothesis.

3. P-value very close to the cut off (0.05) are considered to be marginal (could go either way)

NOTE: Always report the P-Value so that the readers can draw the conclusions.

Other Terms Used in data analytics

1**. History effects** this refers to the possibility that the specific events may occur during the study itself that might influence the outcome.

2. **Maturational effects** this is a fundamentally a bout change over time.

3. **Homogenous attrition** in this case the attrition effect is the same for all groups, treatments and conditions.

4. Regression to the mean this refers to any situation where you select data based on an extreme value on some measure.

5. Data Fabrication This is asituation where people make up the data.

6. **Hoaxes** this is often a joke and many of them are intended to be discovered.

7. **The Placebo effect** this is asituation where the mere fact of being treated causes an improvement in outcomes.

8. **Non-parametric test** this is a test that does not have an assumption about an underlying distribution of the data, such as the assumption that the data is normally distributed.

How to test the Normality in your data

This can be done by inspecting a frequency bar graph and checking how well it follows a bell curve vizually by creating a QQ-plot and inspecting vizually how close the data lay to the diagonal line by performing a Shapiro Wilk test, which has the Null hypothesis that the data is normally distributed.When the P-Value of the Shapiro test is above 0.05, you can say the data is likely to be normally distributed.

ANALYSIS OF VARIANCE (ANOVA) This is a statistical tool for comparing three or more population (Group) means for equality.

Hypothesis H_0:$\mu_1=\mu_2=\mu_3$ vs H_a: μ1=! μ2=! μ3

HYPOTHESIS TESTING

This is a statistical approach for evaluating statements, claims or assumptions about population characteristics of interest to the researcher. *Statistical hypothesis*: this is a statement about the problem distributionof X. Testing of hypothesiis is arule for deciding whether to accept or reject a hypothesis.

Steps involved

1. Hypothesis formulation

2. Computing test statistic.

3. Decision rule.

4. Conclusion.

The Null & Alternative hypothesis The Null hypothesis is denoted by H_0 and it's the one that is normally tested whereas Alternative hypothesis is denoted by H_a. H_0 is rejected on the basis of the random sample from X.Then some other hypothesis must be true.

NOTE*: A Hypothesis is said to be simple if it specify the population completely otherwise it's composite.*

Decision rule

Given H_0: X **is contained** in rejection area vs H_1: X **contained in** acceptance area, then the decision rule is to reject H_0 if X is contained in rejection region and accept if x is contained in acceptance region.

Errors in decision

1. **Type I Error (α)**

This is an error committed when H_0 is rejected when it oughts to be accepted.

I.e. Reject H_0/H_0 true

Accept H_1 when H_1 is false

2. Type II Error (β)

This error is committed by accepting H_0 when it ought to be rejected.

i.e. Accept H_0/H_0 is false

Reject H_1/H_1 is true

1-β is the power of the test and α is the level of significance or the size of the critical region. An ideal test will be used to minimize the probability of Type I and Type II errors.

Regression Analysis

Regression analysisis a statistical tool for establishing relationship between variables.It involves studying the nature of relationship to explain the difference between the variable. Applied regression analysis provides a framework where regression analysis is used to solve various real-life problems.

Nature of relation

1. **Functional/deterministic relation** this is the perfect relation where there exist a one-to-one correspondence between variables.In this case the data points fall directly on the line of relationship.

2**. Statistical relation** this is the relation in which the observations (data points) do not all fall directly on the line of relationship.

Simple Linear Regression Model

This is a representation of a regression situation involving only two variables. The model is in the form $Y_i=\beta_0+ \beta_1 x_1+ \varepsilon$ I For all i=1,2,3,4…….,n where Y_i=the ith response β_0 & β_1=Unknown parameters ε=Error term The error term accounts for any other possible factor that might have been ommitted from the model but is likely to influence the outcome of interest. β_0 is the intercept parameter(occur when x=0).The intercept coefficient is a value at which the line of relationship cuts the y-axis.It signifies the outcome in the absence of the effect of the preditor value.Its only relevant in the model when the value zero is in the scope of the regression model.

Scope of the regression this is the range of all possible value of x in the model.i.e the intercept coefficient is only meaningful when zero is within the scope of the regression. β1 is the slope parameter, it is used to quantify the effect of X on Y.The sign (+/-) attached to the slope parameter describes the nature of relationship between Y & X

Characteristics of a regressin model

1. **The tendency** the mean of the probability distribution of the response variable vary in a systematic fashion with the levels of the predictor variable.

2. **The scattering** the data points around the line of statistical relation.

Residual It is used to measure the changes (variation) in Y that has not been explained by the regression model.

Assumptions of Regression Model

1. Linearity

The assumption states that the response variable is a linear function of the predictor variable i.e there exists a straight line relationship between X and Y.

2. Independence

It states that the error term are uncorrelated from one response to another.

3. Homoscadesticity

The error term have constant variance.

4. Normality

It states that the error terms are normally distributed with mean 0 and variance sigma squared.

Multicollinearitythis is a situation where the predictor variable is related with another predictor variable in a regression model. **Heteroscadasticity**this is when the value is not constant

OutliersThese are the extreme observations that diverges from an overall pattern in a sample.

Types of Outliers

- Univariate
- Multivariate

Data validation methods used in data analytics

1. Data screening

2. Data verification

Steps invlolved in analytics projects

1. Definition of the problem.

2. Data exploration

3. Data preparation

4. Modeling

5. Implementation and tracking

Time series

This is a collection of observations made subsequentially in time.

Approaches to time series

1. **Autocorrelation function (ACF)** this is the analysis in the time domain.

2. **Spectral density function (SDF)** this is the analysis in the frequency domain.It depends on the spectral density function that describes how variation in a series may be accounted by variation and different frequencies.

Type of Variation

1. Decomposing the variables in a series into trend. 2. Seasonal variation. 3. Cyclic variation. 4. Error variation.

Stationary time series a time series is said to be stationary if it has no systematic change in mean (no trend), if it has no systematic change in variance and if strictly there are no period variation.

SERIES A END