Jerome Gomes

jeromegomes89@hotmail.com

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Descriptive statistics

[Document subtitle]

Table of Contents

[1 Measure of Central Tendency 3](#_Toc506485046)

[1.1 Mean 3](#_Toc506485047)

[1.1.1 Arithmetic Mean 3](#_Toc506485048)

[1.1.2 Geometric Mean 3](#_Toc506485049)

[1.1.3 Weighted Mean 3](#_Toc506485050)

[1.1.4 Trimmed Mean 3](#_Toc506485051)

[1.2 Median 3](#_Toc506485052)

[1.3 Mode 3](#_Toc506485053)

[1.3.1 Uni-modal 3](#_Toc506485054)

[1.3.2 Bi-modal 3](#_Toc506485055)

[1.3.3 Multi-modal 3](#_Toc506485056)

[1.4 Comparison Between Mean, Median & Mode 3](#_Toc506485057)

[1.5 Percentiles 3](#_Toc506485058)

[1.6 Quartiles 3](#_Toc506485059)

[2 Measure of Variability 4](#_Toc506485060)

[2.1 Range 4](#_Toc506485061)

[2.2 Interquartile Range (IQR) 4](#_Toc506485062)

[2.3 Variance 4](#_Toc506485063)

[2.4 Standard Deviation (S.D.) 4](#_Toc506485064)

[2.5 Coefficient of Variation (C.V.) 4](#_Toc506485065)

[3 Measure of Association between Variables 4](#_Toc506485066)

[3.1 Covariance 4](#_Toc506485067)

[3.2 Correlation of Coefficient (Pearson’s Product Moment Correlation Coefficient) 5](#_Toc506485068)

[3.3 Correlation of Determination 5](#_Toc506485069)

[4 Measure of Distribution (Shape) & Relative Location 5](#_Toc506485070)

[4.1 Histogram 5](#_Toc506485071)

[4.1.1 Skewness 5](#_Toc506485072)

[4.1.2 Kurtoisis 5](#_Toc506485073)

[4.2 Chebyshev’s Theorem 5](#_Toc506485074)

[4.3 Empirical Rule 5](#_Toc506485075)

[4.4 Z - Score 6](#_Toc506485076)

[4.5 Box Plot 6](#_Toc506485077)

[5 Graphical Descriptive 6](#_Toc506485078)

[5.1 Bar Graph 6](#_Toc506485079)

[5.2 Pie Chart 6](#_Toc506485080)

[5.3 Ogive 6](#_Toc506485081)

[5.4 Scatter Plot 6](#_Toc506485082)

[5.5 Kernel density plots 6](#_Toc506485083)

[5.6 Quantile - Quantile plots 6](#_Toc506485084)

[5.7 Line Graph 6](#_Toc506485085)

[6 Common Types of Variables 6](#_Toc506485086)

[7 Data Consolidation 7](#_Toc506485087)

[7.1 Merging Data 7](#_Toc506485088)

[7.2 Splitting Data 7](#_Toc506485089)

[7.3 Removing Unnecessary variables / Observations 7](#_Toc506485090)

[8 Variable Treatment 7](#_Toc506485091)

[8.1 Outlier Treatment 7](#_Toc506485092)

[8.2 Missing Value Treatment 8](#_Toc506485093)

[8.3 Derived Variables & Adding New Variables 8](#_Toc506485094)

[9 Dimensionality Reduction Techniques 8](#_Toc506485095)

[9.1 Weight of Evidence & Information Value 8](#_Toc506485096)

[9.2 Principal Component Analysis / Factor Analysis 10](#_Toc506485097)

[9.3 Variable Classing 10](#_Toc506485098)

[9.4 Variable Inflation Factor (VIF) - Multicollinearity Check 10](#_Toc506485099)

[9.5 Correlation Analysis / High Correlation Filter 10](#_Toc506485100)

[9.6 Backward Feature Elimination 10](#_Toc506485101)

[9.7 Forward Feature Construction 10](#_Toc506485102)

[9.8 Wald Chi-Square 10](#_Toc506485103)

[9.9 Variable Clustering Using Proc Varclus 10](#_Toc506485104)

[9.10 Random Forests / Ensemble Trees 10](#_Toc506485105)

DESCRIPTIVE STATISTICS

# Measure of Central Tendency

## Mean

### Arithmetic Mean

### Geometric Mean

These are situation when the arithmetic mean neither the median is a better measure for central tendency this happens when the variable is a “rate” – irrespective of growth or de-growth percentage or “rate of change”. Eg. Value of investment over a period of time.

, *i* = (1,2,3, ………,n) , Geometric mean of the returns , , ………,

### Weighted Mean

### Trimmed Mean

It is obtained by deleting a fixed percentage of the smallest and the largest values and computing mean for the rest. Eg. 5% trimmed mean means removing the largest 5% and smallest 5% of the data.

## Median

## Mode

### Uni-modal

### Bi-modal

### Multi-modal

## Comparison Between Mean, Median & Mode

It is better to use median than mean when there are extreme values. Median is not as sensitive to extreme values as the median divides the data into equal parts after sorting the data in ascending or descending order.

## Percentiles

Percentiles divide the data in to 100 parts

## Quartiles

Divides the data into 4 parts

# Measure of Variability

## Range

Rage is the simplest measure of variance as it subtracts the smallest value from the largest.

## Interquartile Range (IQR)

Is the value of 75th percentile minus value of 25th percentile i.e.

## Variance

Also denoted by (sigma square) is sum of the difference between the values of each observation from the mean ()

|  |  |
| --- | --- |
| For Sample | For Population |
|  |  |

## Standard Deviation (S.D.)

When we calculate variation we square the difference between & to nullify the cancellation affect so in standard deviation we square root the variance.

|  |  |
| --- | --- |
| For Sample | For Population |
|  |  |

## Coefficient of Variation (C.V.)

CV tells us how large the standard deviation is in relation to the mean, in other words CV tells us that the sample standard deviation is “so” % of the value of sample mean. CV is useful statistics for comparing the variability of variables that have different SD and different means. SD vary because of the magnitude of the variables.

# Measure of Association between Variables

## Covariance

Covariance tells us the nature of relation between variables, it cannot determine the strength of association between the variables because of the magnitude (units) of the variables.

|  |  |
| --- | --- |
| For Sample | For Population |
|  |  |

## Correlation of Coefficient

It is covariance divided by the SD of variables, we do so to nullify the magnitude (units) component and compare/ judge the derived figure which we couldn’t for covariance. Coefficient of Correlation has a lower limit of -1 and upper limit of +1. -1 signifies negative relation and + 1 signifies positive relation as we move closer to 0 the strength of the relation minimizes.

## Correlation of Determination

The drawback of coefficient of correlation is that except for these values -1, 0 & 1 we cannot interpret the correlation. Squaring the coefficient of correlation we can express it as a percentage i.e.

# Measure of Distribution (Shape) & Relative Location

## Histogram

### Skewness

It is the shape of the distribution i.e. whether the slope of the distribution is on the right – Positive Skewness or the slope of the distribution is on the left – Negative Skewness. For positively skewed distribution mean is greater than the median (mean>median) and for negatively skewed distribution median is greater than the mean (median>mean).

### Kurtoisis

Kurtosis is the height of distribution.

## Chebyshev’s Theorem

Chebyshev’s Theorem enables us to make statements about proportions of data values. One of the advantage of Cheby’s Theorem is that it applies to any set of data regardless of distribution shape.

* At least .75 or 75% of the data values must be with in z = 2 Standard deviation.
* At least .89 or 89% of the data values must be with in z = 3 Standard deviation.
* At least .94 or 94% of the data values must be with in z = 4 Standard deviation.

## Empirical Rule

It is for distribution which are symmetric (i.e. a bell shaped distribution)

* Approximately 68% of the data values will be within 1 standard deviation of the mean.
* Approximately 95% of the data values will be within 2 standard deviation of the mean.
* Approximately 99.9% of the data values will be within 3 standard deviation of the mean.

## Z - Score

Associated with each values in a data set is another value called the z –score which tells it’s (the values) relative location within the particular data set. The z-score is often called standard value and is interpreted as the number of standard observations (the value) is greater or lesser than the mean.

E.g. if = 1.2, it would indicate that is 1.2 standard deviation greater than the mean. For = 0, is equal to mean.

## Box Plot

# Graphical Descriptive

## Bar Graph

## Pie Chart

## Ogive

## Scatter Plot

## Kernel density plots

## Quantile - Quantile plots

## Line Graph

# Common Types of Variables

* **Categorical variable:** variables than can be put into categories. For example, the category “Toothpaste Brands” might contain the variables Colgate and Aquafresh.
* **Confounding variable:** extra variables that have a hidden effect on your experimental results.
* **Control variable:** a factor in an experiment which must be held constant. For example, in an experiment to determine whether light makes plants grow faster, you would have to control for soil quality and water.
* **Dependent variable:** the outcome of an experiment. As we change the values of independent variable, which directly effects the dependent variable, i.e. *y*
* **Independent variable:** sometimes called an **experimental or predictor variable** that is not affected by anything that you, i.e.
* **Discrete variable:** a variable that can only take on a certain number of values. For example, “number of cars in a parking lot” is discrete because a car park can only hold so many cars.
* **Continuous variable:** a variable with infinite number of values, like “time” or “weight”.
* **Nominal variable:** another name for categorical variable.
* **Ordinal variable:** similar to a categorical variable, but there is a clear order. For example, income levels of low, middle, and high could be considered ordinal.
* **Qualitative variable:** a broad category for any variable that can’t be counted (i.e. has no numerical value). Nominal and ordinal variables fall under this umbrella term.
* **Quantitative variable:** A broad category that includes any variable that can be counted, or has a numerical value associated with it. Examples of variables that fall into this category include discrete variables and ratio variables.
* **Random variables** are associated with random processes and give numbers to outcomes of random events.
* **A ranked variable** is an ordinal variable; a variable where every data point can be put in order (1st, 2nd, 3rd, etc.).

# Data Consolidation

## Merging Data

## Splitting Data

Before proceeding with any analysis we need to split the data randomly in to Development sample and Validation sample or Training and Test data set. Development sample is used to build the model while the Validation sample is used to validate the effectiveness of the model. We divide the data set into 60:40, or 70:30 ratio.

## Removing Unnecessary variables / Observations

# Variable Treatment

## Outlier Treatment

An outlier is an observation that diverges or stands out from the majority of the data. Following are the ways to handle Outliers:

* Drop the observations (i.e., the Outliers)
* Persist with the observations (i.e., the Outliers) as they are
* Cap the observations (i.e., the Outliers) at the closest possible actual value - This approach is the generally advised method of handling Outliers; this approach is commonly referred to as “Outlier Capping”

1. **Three Sigma Approach**: This is applicable when the data follows a normal distribution, so we know that 99.73% of observations fall between μ ±3σ. We use this capping to replace observations less than m-3s and replace with the value m-3s & observations greater than m+3s are replaced with value m+3s.
2. **Interquartile Range (IQR) Approach**: All observations that are less than Q1 – 1.5 IQR are Outliers and replaced with the values Q1 – 1.5 IQR. Similarly, all observations greater than Q3 + 1.5 IQR are Outliers and replaced with the values Q1 + 1.5 IQR. The range defined by Inter Quartile Range (IQR) covers nearly μ ± 2.7σ for a Normal distribution. Thus, for a normally distributed variable, this range contains 99.31% of the observations.

## Missing Value Treatment

Some of the thumb rule for handling missing values are:

* If a variable has missing values for more than 8-10% of observations, we drop the variable from our analysis.
* If the variable is a discrete variable or an ordinal variable, we impute the missing values by the median of the available observations.
* If the variable is a continuous variable, we impute the missing values by the mean of the available observations.
* If a variable is a continuous variable, but we know from experience that it is clearly not normally distributed, we impute the missing values by the median of the available observations.
* If a variable is a categorical variable, we impute the missing values by the mode of the available observations.

## Derived Variables & Adding New Variables

Some time we need to add some extra columns e.g. ratio column, bucketing column to the data frame for extracting more information.

We can get a better relation between the dependent and independent values by transforming the independent variable (i.e. we take the log, exponential), also known as Box-Cox Transformation were we consider the following transformations on our predictor variables:

* Log
* Exponential
* Square
* Square Root
* Inverse
* Inverse Log
* Inverse Exponential

# Dimensionality Reduction Techniques

Finding key variables for analysis

<https://www.linkedin.com/pulse/variable-reduction-art-well-science-shailendra-s/>

<https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html>

<https://www.casact.org/pubs/forum/06wforum/06w93.pdf>

<http://www.askanalytics.in/p/varred4.html>

<https://www.lexjansen.com/nesug/nesug11/sa/sa08.pdf>

## Weight of Evidence & Information Value

The weight of evidence (WOE) and information value (IV) provide a great framework for exploratory analysis and variable screening for binary classifiers. Binary classification models are perhaps the most common use-case in predictive analytics. The reason is that many key client actions across a wide range of industries are binary in nature, such as defaulting on a loan, clicking on an ad, or terminating a subscription. WOE and IV play two distinct roles when analysing data: WOE describes the relationship between a predictive variable and a binary target variable &IV measures the strength of that relationship.

Weight of evidence (WoE) can be used to find similar groups and reduce data dimensions vertically by grouping similar groups.

Information value (IV) of a predictor variable is a non-parametric measure that calculates the level of information contained in the predictor variable about the target or dependent variable. It captures linear as well as nonlinear relationships, works for both quantitative and qualitative variables, not affected by outliers and can reduce thousands of variables to 100 to 150 variables. It works for supervised learning and not for unsupervised learning. However there is a major drawback to this method for continuous target variables IV might not be well defined. Higher the IV of a predictor variable, greater is the power of the variable to predict the target variable .Thus, we should select only those variables that have high IVs. In real-world problems, any IV greater than 0.2 can be considered good. However, this may vary from case to case. Usually, we decide how many variables we want to take to the next level. Then, we select the top few variables in descending order of their IVs.

Few exceptions to IV are:

* In a rare instance, if the observations in a group happen to be either all good or all bad, then the calculation of WoE is not possible. To avoid such a problem, we introduce a small correction factor in the formula for “%Good” and “%Bad”, i.e. %Bad = (Bad + 0.005) / (Total Bad + 0.005) & %Good = (Good + 0.005) / (Total Good + 0.005).
* A hypothetical scenario, with respect to the variable “Channel”, wherein all the Bad connections arise from the “C” channel. In such a scenario, the variable “Channel” is a strong predictor for the target variable, and IV of the variable “Channel” is high.
* Consider another hypothetical scenario, with respect to the variable “Channel”, wherein the Churn Rate is the same across all the channels. In such a scenario, the variable “Channel” does not contain any meaningful information about the target variable, and the IV of the variable “Channel” is zero



Channel, Age, Gender, Demography, Month, Employment status etc.

<https://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/>

<http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/>

<http://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html>

<http://r-statistics.co/Information-Value-With-R.html>

https://blogs.oracle.com/r/computing-weight-of-evidence-woe-and-information-value-iv

## Principal Component Analysis / Factor Analysis

Orthogonal

## Variable Classing

## Variable Inflation Factor (VIF) - Multicollinearity Check

## Correlation Analysis / High Correlation Filter

## Backward Feature Elimination

## Forward Feature Construction

## Wald Chi-Square

## Variable Clustering Using Proc Varclus

## Random Forests / Ensemble Trees