Jerome Gomes

jeromegomes89@hotmail.com OR 8240767360

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

Introduction to Statistics

Table of Contents

[1 Measure of Location 3](#_Toc17287801)

[1.1 Mean 3](#_Toc17287802)

[1.1.1 Arithmetic Mean 3](#_Toc17287803)

[1.1.2 Geometric Mean 3](#_Toc17287804)

[1.1.3 Weighted Mean 3](#_Toc17287805)

[1.1.4 Trimmed Mean 3](#_Toc17287806)

[1.2 Median 4](#_Toc17287807)

[1.3 Mode 4](#_Toc17287808)

[1.3.1 Uni-modal 4](#_Toc17287809)

[1.3.2 Bi-modal 4](#_Toc17287810)

[1.3.3 Multi-modal 4](#_Toc17287811)

[1.4 Comparison Between Mean, Median & Mode 4](#_Toc17287812)

[1.5 Percentiles 4](#_Toc17287813)

[1.6 Quartiles 4](#_Toc17287814)

[2 Measure of Variability 4](#_Toc17287815)

[2.1 Range 4](#_Toc17287816)

[2.2 Interquartile Range (IQR) 5](#_Toc17287817)

[2.3 Variance 5](#_Toc17287818)

[2.4 Standard Deviation (S.D.) 5](#_Toc17287819)

[2.5 Coefficient of Variation (C.V.) 5](#_Toc17287820)

[3 Measure of Association between Variables 5](#_Toc17287821)

[3.1 Covariance 5](#_Toc17287822)

[3.2 Correlation of Coefficient [Pearson Product Moment] 6](#_Toc17287823)

[3.3 Correlation of Determination 6](#_Toc17287824)

[4 Measure of Distribution (Shape) & Relative Location 6](#_Toc17287825)

[4.1 Histogram 6](#_Toc17287826)

[4.1.1 Skewness 6](#_Toc17287827)

[4.1.2 Kurtoisis 6](#_Toc17287828)

[4.2 Z - Score 7](#_Toc17287829)

[4.3 Chebyshev’s Theorem 7](#_Toc17287830)

[4.4 Empirical Rule 7](#_Toc17287831)

[4.5 Box Plot 7](#_Toc17287832)

[5 Graphical Descriptive 8](#_Toc17287833)

[5.1 Bar Graph 8](#_Toc17287834)

[5.2 Pie Chart 8](#_Toc17287835)

[5.3 Ogive 8](#_Toc17287836)

[5.4 Scatter Plot 8](#_Toc17287837)

[5.5 Kernel density plots 8](#_Toc17287838)

[5.6 Quantile - Quantile plots 8](#_Toc17287839)

[5.7 Line Graph 8](#_Toc17287840)

[5.8 Bubble Charts 8](#_Toc17287841)

[5.9 HeatMaps 8](#_Toc17287842)

[6 Data Consolidation 8](#_Toc17287843)

[6.1 Merging Data 8](#_Toc17287844)

[6.2 Splitting Data 8](#_Toc17287845)

[6.3 Removing Unnecessary variables / Observations 8](#_Toc17287846)

[7 Variable Treatment 8](#_Toc17287847)

[7.1 Outlier Treatment 8](#_Toc17287848)

[7.2 Missing Value Treatment 9](#_Toc17287849)

[7.3 Derived Variables & Adding New Variables 9](#_Toc17287850)

DESCRIPTIVE STATISTICS

# Measure of Location

## Mean

### Arithmetic Mean

= AVERAGE ([Input Range])

### 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 , , ………,

Suppose you make a 2-year investment of $1,000, and it grows by 100% to $2,000 during the first year. During the second year, however, the investment suffers a 50% loss, from $2,000 back to $1,000. The rates of return for years 1 and 2 are R = 100% and R2 = −50%, respectively.

Value at the end of the investment period = 1,000(1 + Rg)2 = 1000(1 + 0)2 = 1000

= GEOMEAN ([Input Range])

### Weighted Mean

In some instances, the mean is computed by giving each observation a weight that reflects its importance. A mean computed in this manner is referred to as a weighted mean.

Where: = value of observation of *i &*  = weight of observation of *i*

When the data are from a population, μ replaces . As an example of the need for a weighted mean, consider the following sample of five purchases of a raw material over the past three months.



### Trimmed Mean

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

= TRIMMEAN(array,percent)

## Median

= MEDIAN ([Input Range])

## Mode

### Uni-modal

= MODE.SNGL ([Input Range])

### Bi-modal

### Multi-modal

= MODE.MULT ([Input Range])

## 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

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value. Percentiles divide the data in to 100 parts. The pth percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100 – p) percent of the observations are greater than or equal to this value.

= PERCENTILE(array,k)

If the score of 54 corresponds to the 70th percentile, we know that approximately 70% of the students scored lower than this individual and approximately 30% of the students scored higher than this individual.

## Quartiles

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. The division points are referred to as the quartiles and are defined as

* *Q1*= first quartile, or 25th percentile
* *Q2* = second quartile, or 50th percentile (also the median)
* *Q3* = third quartile, or 75th percentile002E

= QUARTILE.EXC(array,quart)

= QUARTILE.INC(array,quart)

= QUARTILE(array,quart)

# Measure of Variability

## Range

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

*Range = Largest value - Smallest value*

## Interquartile Range (IQR)

Measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR). This measure of variability is the difference between the third quartile, Q3 and the first quartile Q1. In other words, the interquartile range is the range for the middle 50% of the data.

## Variance

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

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

= VAR ([Input Range])

## 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 |
|  |  |

= STDV ([Input Range])

## 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 “X” % 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 (negative or positive), it cannot determine the strength of association between the variables because of the magnitude (units) of the variables.

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

= COVAR ([Input Range])

## Correlation of Coefficient [Pearson Product Moment]

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. The population parameter is denoted by the Greek letter rho. The advantage that the coefficient of correlation has over the covariance is that the former has a set lower and upper limit. The limits are −1 and +1, respectively—that is, −1 ≤ r ≤ +1 and −1 ≤ ρ ≤ +1.

When the coefficient of correlation equals −1, there is a negative linear relationship and the scatter diagram exhibits a straight line. When the coefficient of correlation equals +1, there is a perfect positive relationship. When the coefficient of correlation equals 0, there is no linear relationship. All other values of correlation are judged in relation to these three values. The drawback to the coefficient of correlation is that—except for the three values −1, 0, and +1—we cannot interpret the correlation

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

= CORREL ([Input Range])

## 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

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution

### 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). For data skewed to the left, the skewness is negative; for data skewed to the right, the skewness is positive. If the data are symmetric, the skewness is zero. For a symmetric distribution, the mean and the median are equal. The median provides the preferred measure of location when the data are highly skewed.

### Kurtoisis

Kurtosis is the height of distribution.

## 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. Measures of relative location help us determine how far a particular value is from the mean. The z-score is often called score is often called the ***standardized value*** and is interpreted as the number of standard observations (the value) is greater or lesser than the mean.

=STANDARDIZE(x,mean,standard\_dev)

## Chebyshev’s Theorem

Chebyshev’s theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean. A more general interpretation of the standard deviation is derived from Chebysheff’s Theorem, which applies to all shapes of histograms. 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.

*Mean & Standard Deviation 🡪 Z-Score 🡪 proportion of data values*

One of the advantages of Chebyshev’s theorem is that it applies to any data set regardless of the shape of the distribution of the data.

## 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.

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

A box plot is a graphical summary of data that is based on:

1. Smallest value
2. First quartile
3. Median
4. Third quartile
5. Largest value

# Graphical Descriptive

## Bar Graph

## Pie Chart

## Ogive

## Scatter Plot

## Kernel density plots

## Quantile - Quantile plots

## Line Graph

## Bubble Charts

## HeatMaps

# 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

=LN(number)

=LOG(number,base)

* Exponential - e^(number), where e ≈ 2.718

=EXP (number)

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