e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module No 10: CS/DA/10- Data Analysis Foundations – Univariate analysis
Quadrant 1 – e-text

1.1 Introduction

The aim is to give the understanding of the importance of data analysis. Here we discuss different data measurement and data measurement scales. Univariate data analysis and measures of univariate analysis are also discussed in detail with appropriate examples.

1.2 Learning Outcomes

- To Understand the basic statistical methods for exploratory data analysis of numeric attributes
- To learn about measures of central tendency, and dispersion for univariate data

1.3 Data Measurement

Measurement is a procedure for assigning symbols, letters, or numbers to empirical properties of variables according to rules. Measurement of the data is the first step in the process that ultimately guides the final analysis.

Measurement of the data is the first step in the process that ultimately guides the final analysis. Consideration of sampling, controls, errors (random and systematic) and the required precision all influence the final analysis.

Data validation is intended to provide certain well-defined guarantees for fitness, accuracy, and consistency for any of various kinds of user input into an application or automated system. Instruments and methods used to measure the data must be validated for accuracy. Precision and accuracy measures are used to determine the error.

1.4 Types of data

a) Univariate/Multivariate data

- **Univariate data:** Single-variable or univariate data refers to data where we're only observing one aspect of something at a time.

- **Multivariate data:** Multiple-variable or multivariate data refers to data where we're observing two or more aspect of something at a time

b) Cross-sectional data/Time-ordered data (business, social sciences)

- **Cross-sectional data:** Cross-sectional data, or a cross section of a study population, in statistics and econometrics is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time. Or mmeasurements taken at one time period.

- **Time ordered data:** Type of data collected at over period of time, with regarding to difference in timer or measurements taken at over time in chronological sequence.

1.5 Measurement Scales

In statistics, the term measurement is used more broadly and is more appropriately termed scales of measurement. Scales of measurement refer to ways in which variables/numbers are defined and categorized. Each scale of measurement has certain properties which in turn determines the appropriateness for use of certain statistical analyses. The four scales of measurement are nominal, ordinal, interval, and ratio.

i) **Nominal or Categrical:** Categorical data and numbers that are simply used as identifiers or names represent a nominal scale of measurement. Numbers on the back of a football jersey and your adhar number are examples of nominal data. Nominal scales are used for labeling variables, without any quantitative value. "Nominal" scales could simply be called "labels." All of these scales are mutually exclusive (no overlap) and none of them have any numerical significance.

- Eg: age ranges, colors, etc.

ii) **Ordinal:** An ordinal scale of measurement represents an ordered series of relationships or rank order. With ordinal scales, it is the order of the values is what's important and significant, but the differences between each one is not really known. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc."Ordinal" is easy to remember because is sounds like "order" and that's the key to remember with "ordinal scales"–it is the *order* that matters, but that's all you really get from these.

The best way to determine *central tendency* on a set of ordinal data is to use the mode or median; the mean cannot be defined from an ordinal set.

Eg: Classification of people, places, or things into a ranking such that the data is arranged into a meaningful order (e.g. poor, fair, good, excellent).

iii) **Interval:** A scale that represents quantity and has equal units but for which zero represents simply an additional point of measurement is an interval scale.

Eg: The Fahrenheit scale is a clear example of the interval scale of measurement. Thus, 60 degree Fahrenheit or -10 degrees Fahrenheit represent interval data. Measurement of Sea Level is another example of an interval scale. With each of these scales there are direct, measurable quantities with equality of units. In addition, zero does not represent the absolute lowest value. Rather, it is point on the scale with numbers both above and below it (for example, -10degrees Fahrenheit.

iv) **Ratio:** The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units. However, this scale also has an absolute zero (no numbers exist below zero).

Eg: Very often, physical measures will represent ratio data (for example, height and weight). If one is measuring the length of a piece of wood in centimeters, there is quantity, equal units, and that measure cannot go below zero centimeters. A negative length is not possible.

Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

1.6 Univariate analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression) and it's major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

The main purpose of univariate analysis is description.

- Data matrix **D can** be thought of as an n×1 matrix, or simply a column vector, given as

  D= [x1, x2, x3……., xn]

  where *X is the numeric attribute of interest, with* $x_i \in R$.

X is assumed to be a random variable, with each point x_i ($1 \leq i \leq n$) itself treated as an identity random variable.

Some ways you can describe patterns found in univariate data include central tendency (mean, mode and median) and dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation.

You have several options for describing data with univariate data. Click on the link to find out more about each type of graph or chart:

- Frequency Distribution Tables
- Bar Charts
- Histograms
- Frequency Polygons
- Pie Charts

1.7 Measures of Univariate analysis

a) Range:

The "range" is just the difference between the largest and smallest values. The calculation of the range is very straightforward. All we need to do is find the difference between the largest data value in our set and the smallest data value. So we have the following formula:

Range = Maximum Value – Minimum Value.

Eg: *Find the range of the data set 4, 6, 10, 15, 18*

The data set has maximum of 18, minimum of 4 and range of 18 – 4 = 14.

So range is 14

Limitations of Range

The range is a very crude measurement of the spread of data because it is extremely sensitive to outliers. A single data value can greatly affect the value of the range. For example, consider the set of data 1, 2, 3, 4, 6, 7, 7, 8.
The maximum value is 8, the minimum is 1 and the range is 7. Now consider the same set of data, only with the value 100 included. The range now becomes 100 – 1 = 99. The addition of a single extra data point greatly affected the value of the range.

b) Mean(The average)

The mean is the average of the numbers i.e. a calculated central value of a set of numbers. To calculate the mean, just add up all the numbers, then divide by how many numbers there are. In other words mean is the arithmetic average of all numbers; the sum of all values divided by the total number of values. Mean uses both rank order and distance between ranks.

Eg: what is the mean of 2,7 and 9?

Add the numbers: 2+7+9=18

Divide by how many numbers(i.e,we added 3 numbers): 18/3=6

So the mean is 6

c) **Median**

Median is the middle value in a set of numbers arranged in order of magnitude. It is most appropriate for ordinal data - uses only the rank order, ignores distance. Is also sometimes good for interval and ratio level data that have some extreme values - for example, income figures could be misleading if the sample or population includes a few multi-millionaires.

Eg: Find median of 13, 18, 13, 14, 13, 16, 14, 21, 13.

The median is the middle value, so we have to rewrite the list in order:

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the (9 + 1) ÷ 2 = 10 ÷ 2 = 5th number:

13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.

Eg: Find median of 1,2,4,7

The median is the middle number. In this example, the numbers are already listed in numerical order, so I don't have to rewrite the list. But there is no "middle" number, because there are an even number of numbers. In this case, the median is the mean (the usual average) of the middle two values:

(2 + 4) ÷ 2 = 6 ÷ 2 = 3

So the median is 3.

d) Mode

The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

Eg: Find mode of 13, 18, 13, 14, 13, 16, 14, 21, 13

The mode is the number that is repeated more often than any other, so 13 is the mode.

Eg: Find mode of 1,2,4,7
The mode is the number that is repeated most often, but all the numbers in this list appear only once, so there is no mode.

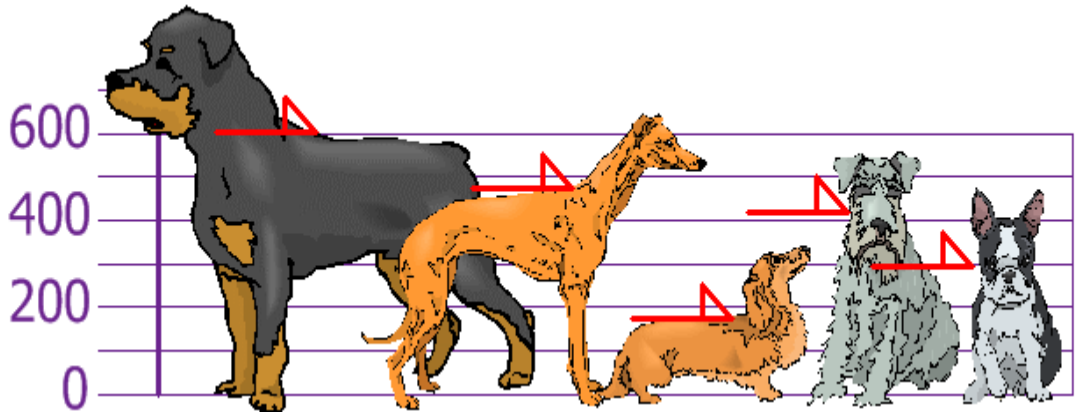
e) Variance

The variance is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the variance is big; and vice versa.
It is important to distinguish between the variance of a population and the variance of a sample. They have different notation, and they are computed differently. The variance of a population is denoted by σ^2; and the variance of a sample, by s^2.

The variance of a population is defined by the following formula:
$\sigma^2 = \Sigma (X_i - X)^2 / N$
where σ^2 is the population variance, X is the population mean, X_i is the *i*th element from the population, and N is the number of elements in the population.

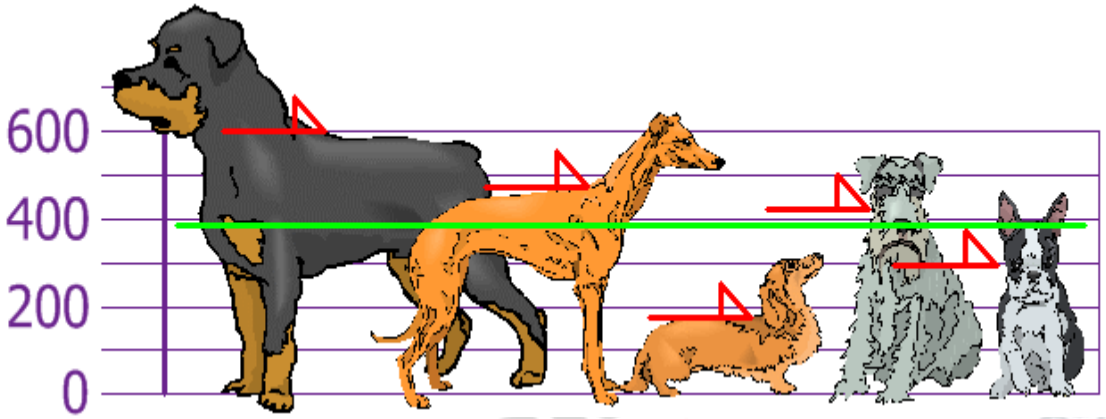
The variance of a sample is defined by slightly different formula:
$s^2 = \Sigma (x_i - x)^2 / (n - 1)$
where s^2 is the sample variance, x is the sample mean, x_i is the *i*th element from the sample, and n is the number of elements in the sample. Using this formula, the variance of the sample is an unbiased estimate of the variance of the population.
And finally, the variance is equal to the square of the standard deviation.

Eg: You and your friends have just measured the heights of your dogs (in millimeters):

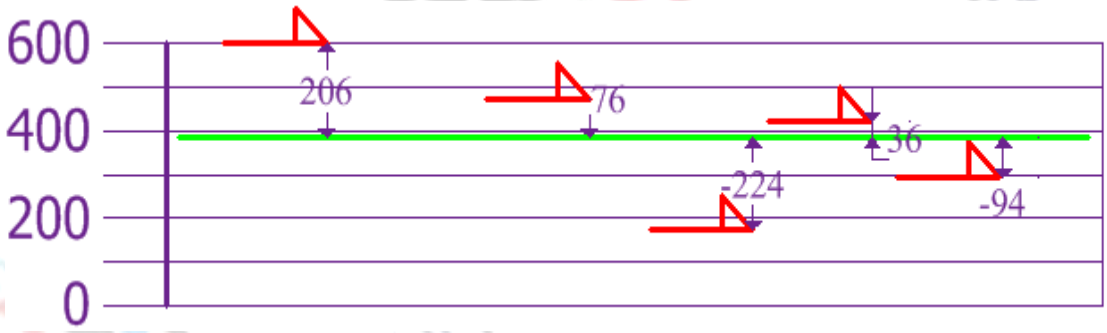


The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm. Find out the Mean, the Variance, and the Standard Deviation.
Your first step is to find the Mean:

Mean = 600 + 470 + 170 + 430 + 300***5*** = 1970***5*** = *394*
so the mean (average) height is 394 mm. Let's plot this on the chart:



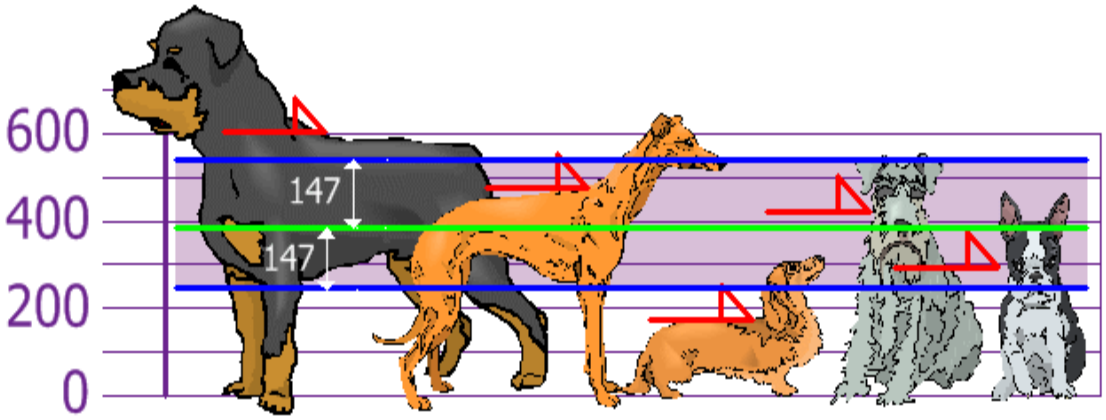
Now we calculate each dog's difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

$$\text{Variance: } \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$
$$= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5}$$
$$= \frac{108,520}{5} = 21,704$$

So the Variance is **21,704**



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

But, there is a small change with **Sample** Data. Our example was for a Population (the 5 dogs were the only dogs we were interested in). But if the data is a Sample (a selection taken from a bigger Population), then the calculation changes.

When you have "N" data values that are:

- **The Population**: divide by **N** when calculating Variance (like we did)
- **A Sample**: divide by **N-1** when calculating Variance

All other calculations stay the same, including how we calculated the mean.

Example: if our 5 dogs were just a sample of a bigger population of dogs, we would divide by 4 instead of 5 like this:

Sample Variance = 108,520 / 4 = 27,130

Think of it as a "correction" when your data is only a sample.

f) Standard deviation

The Standard Deviation is a measure of how spread out numbers are. Its symbol is **σ** (the greek letter sigma) The formula is easy: it is the **square root** of the **Variance.**

"**Population** Standard Deviation": $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

"**Sample** Standard Deviation**":** $$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Looks complicated, but the important change is to divide by **N-1** (instead of **N**) when calculating a Sample Variance. By referring the example in variance part, and the Standard Deviation is just the square root of Variance, so:

Standard Deviation, σ = √21,704

= 147.32...

= **147** *(to the nearest mm)*

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:

Sample Standard Deviation,s = √27,130 = 164 (to the nearest mm)

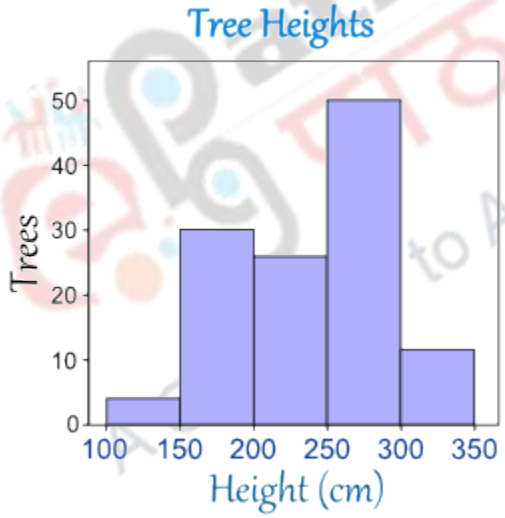
g) Histograms and distributions

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. It is a graphical display of data using bars of different heights. It is similar to Bar chart but a histogram groups numbers into ranges,and we have to decide what ranges to use.

Eg: Height if Orange Trees. You measure the height of every tree in the orchard in centimeters (cm). The heights vary from 100 cm to 340 cm. You decide to put the results into groups of 50 cm:

- The 100 to just below 150 cm range,
- The 150 to just below 200 cm range,
- etc...

So a tree that is 260 cm tall is added to the "250-300" range. And the result is:



The horizontal axis is continuous like a number line

A **probability distribution** is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

A variable is a symbol (A, B, x, y, etc.) that can take on any of a specified set of values.

When the value of a variable is the outcome of a statistical experiment, that variable is a random variable. Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,

X represents the random variable X.

P(X) represents the probability of X.

P(X = x) refers to the probability that the random variable X is equal to a particular value, denoted by x. As an example, P(X = 1) refers to the probability that the random variable X is equal to 1.

Probability Distributions

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. Consider the coin flip experiment described above. The table below, which associates each outcome with its probability, is an example of a probability distribution.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents the probability distribution of the random variable X.

Cumulative Probability Distributions

A cumulative probability refers to the probability that the value of a random variable falls within a specified range.

Let us return to the coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.

$P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$

Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation. In the table below, the cumulative probability refers to the probability than the random variable X is less than or equal to x.

Number of heads: X	Probability: P(X = x)	Cumulative Probability: P(X < x)
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

Uniform Probability Distribution

The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the uniform distribution.

Uniform Distribution. Suppose the random variable X can assume k different values. Suppose also that the $P(X = x_k)$ is constant. Then,

$P(X = x_k) = 1/k$

Example 1

Suppose a die is tossed. What is the probability that the die will land on 5?

Solution: When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the P(X = 5) = 1/6.

Example 2

Suppose we repeat the dice tossing experiment described in Example 1. This time, we ask what is the probability that the die will land on a number that is smaller than 5?

Solution: When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is equally likely to occur. Thus, we have a uniform distribution.

This problem involves a cumulative probability. The probability that the die will land on a number smaller than 5 is equal to:

$P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$

	**Case Studies**

A) http://www.learningcommons.uoguelph.ca/guides/university_learning/handouts/highlighting.pdf

Much of psychological research involves measuring observations of particular characteristics of either a population, or a sample taken from a population. These measurements yield a set of values or scores, and this set represents the findings of the research, or data. Often, it is impractical to completely measure the characteristics of a given population, known as parameters, directly. Thus, psychologists often focus on the characteristics of samples taken from a population. These characteristics are called statistics. The psychologist then uses these sample statistics to make inferences about population parameters.

B) http://www.investopedia.com/articles/financial-theory/10/gaussian-models-statistics.asp

Standard deviation measures volatility and asks what kind of performance returns can be expected. Smaller standard deviations may mean less risk for a stock, while higher volatility may mean a higher level of uncertainty. Traders can measure closing prices from the average as it is dispersed from the mean. Dispersion would then measure the difference from actual value to average value. A larger difference between the two means a higher standard deviation and volatility. Prices that deviate far away from the mean often revert back to the mean, so that traders can take advantage of these situations. Prices that trade in a small range are ready for a breakout.

Summary

- Importance of data measurement in analysis
- Four types of data – uni/multivariate and Cross sectional/Time-ordered
- Types of measurement scales – nominal, ordinal, interval and ratio-scaled