



Predictive Modeling with SAS Course

LIVE PROJECTS, REAL DATASETS, CASE
STUDIES AND PRACTICAL EXAMPLES

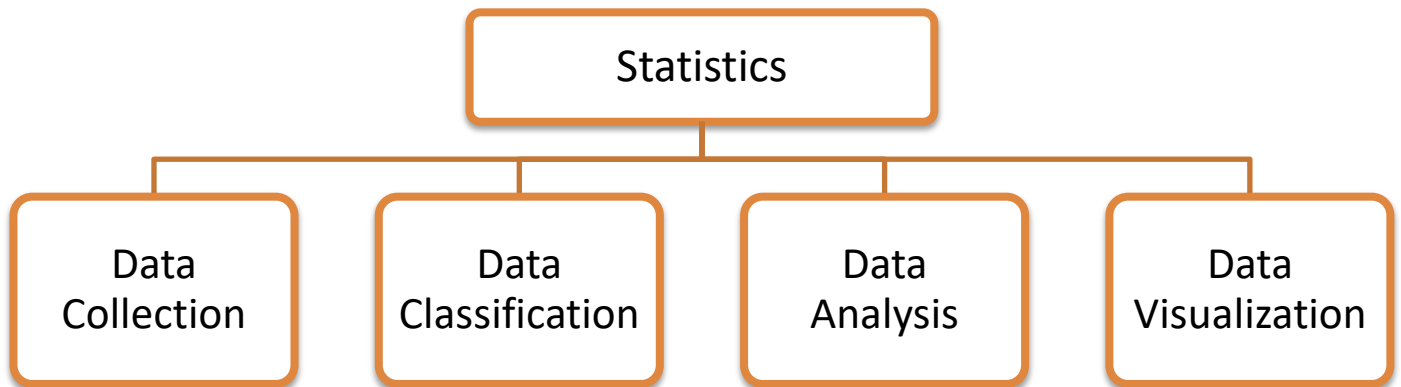


CHAPTER ONE

Introduction to Statistics & Predictive Modeling

What is Statistics?

 *Learning Statistics was never so Easy!*



Why to learn Statistics?

*It helps in **making more effective decisions**. Understanding of Statistical Analysis is required in almost every line of work.*

Examples

- 1. The average income in the United States is around \$72,000 (US Census Bureau 2014) **Seriously?***

Prior reading this fact, someone might think he was actually doing pretty well for yourself with \$60k salary, but now he wants to get a second job just to increase his net worth. *How to make money, damn it?*

Trick : The term "average" is way different from the mathematical term

What is Statistics? (Contd.)

The term “average” used in the fact is referring to “mean” which is affected by extreme positive or negative values.

A few rich people are skewing the average number. If you're earning less than the average income, it's not because your job is screwing you, it's because you live in the same country as Bill Gates, Mark Zuckerberg, Warren Buffet and many other billionaires.

The median income in the United States is around \$53,000 (US Census Bureau 2014)

2. One study shows that for every 100 Americans, there are 88 guns.

It means it's hard to find an American who does not have a gun.

Another study from the same year shows that only 43% of households have guns in them.

Trick :

In the first study, people who have a lot of guns skew the average upward. It showed incorrect picture of the situation.

What is Statistics? (Contd.)

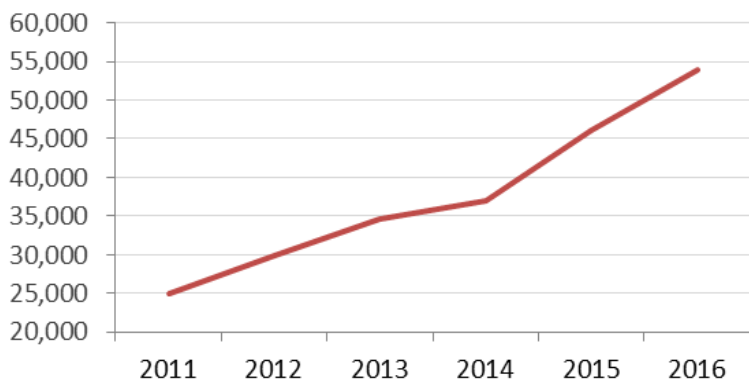
3. Biased sampling :

A survey finds that “**41% of retail bank customers would use mobile banking if it were available**” becomes meaningless when you find out the survey only polled people on their mobile devices.

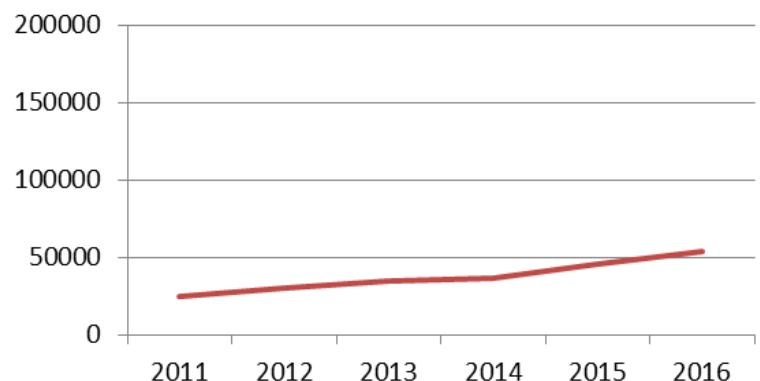
4.

Years	Number of Users
2011	25,000
2012	29,902
2013	34,545
2014	37,000
2015	46,000
2016	54,000

Number of Users



Number of Users



Both of the charts above contain exactly the same information, but which one more accurately shows the increase in number of users between the years 2011-2016? ***The only difference between the graphs is the scale.***

Test your knowledge

1. You need to calculate median salary of Group A and B
2. You need to calculate overall median salary
3. Check whether $\text{Median}(\text{GroupA}) + \text{Median}(\text{GroupB}) = \text{Overall Median}$

Refer "Exercise" tab of excel file named "Dataset"

What is Predictive Modeling?

Predictive

+

Modeling

To predict the future
based on historical data

By using statistical algorithms and
machine learning techniques

Why it is trending these days?

- ✓ 90% of all the data in the world today has been generated in the last two years alone. [Source : IBM Marketing Cloud]
- ✓ Faster, cheaper computers
- ✓ Data Mining softwares make it easy to predict future outcomes
- ✓ Differentiation to create a competitive advantage



According to a **Forrester study**, firms that use predictive analytics are **2.9 times more** likely to report revenue growth at a rate that is higher than the industry average. Top-performing companies are also much more likely to be using predictive analytics than others.

Use Cases of Predictive Analytics

Banking & Insurance

Customer Attrition

Likelihood a customer will leave a bank

Savings

Current A/C

Credit Card

Loan

Propensity to Buy

Likelihood a customer will buy a product

Probability of Default

Likelihood a customer will default a loan

Fixed Deposit Renew

Identify customers who are likely to mature FD

Customer Segmentation

Segmenting customers based on their spending pattern

Detecting Fraud

Predict Fraud in Insurance Claims

Sales Forecasting

Forecast Sales in the next 3 years

Use Cases of Predictive Analytics

Retail

✓ **Market Basket Analysis**

Identify items that are likely to be purchased together

✓ **Optimal Pricing**

Determine change in price which can lead to increase in profits by 10% considering the fact increase in price can lead to customer attrition

✓ **Personalized Offers**

Target the right products and offers based on the individual customer DNA

HR

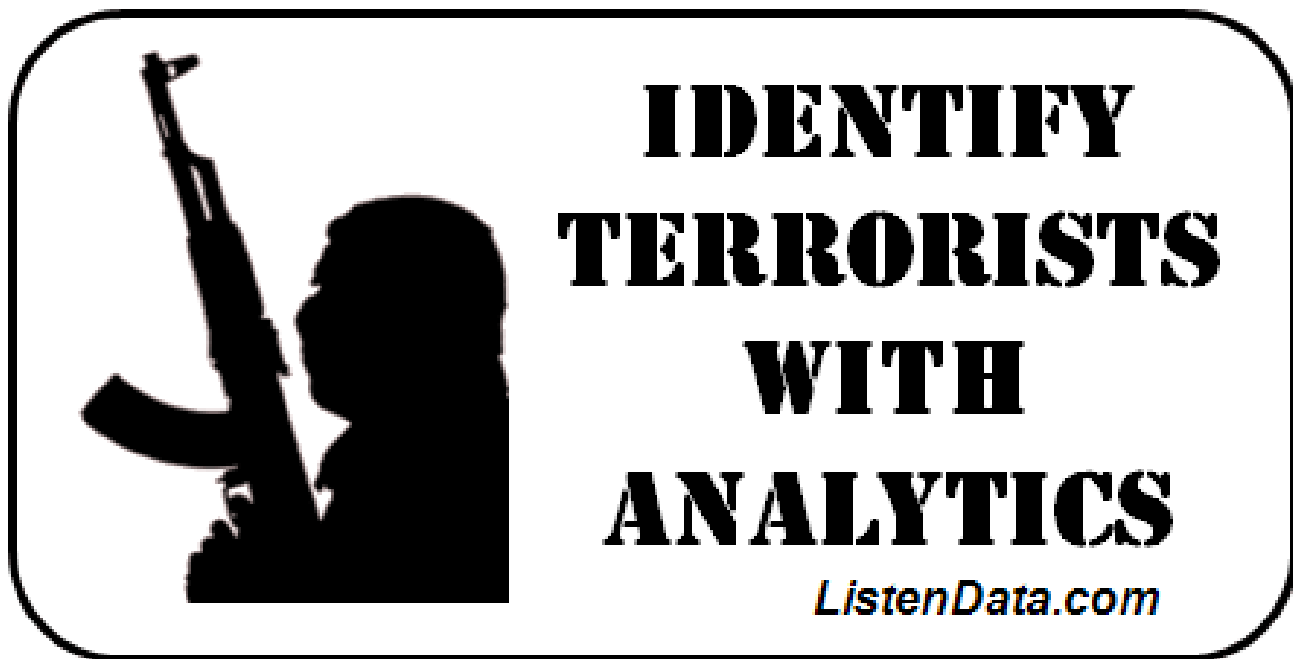
✓ **Employee Attrition**

Identify employees who can leave organization in next 3 months.

✓ **Employee Engagement**

Identify factors that can make employees stay or leave the organization based on employee engagement surveys.

Till what extent Predictive Modelling can be used?



Likelihood of being a Terrorist

US National Security Agency uses a machine learning algorithm to assess each person's likelihood of being a terrorist. They used Pakistan's mobile network metadata of 55 million people to develop a model to identify terrorists.

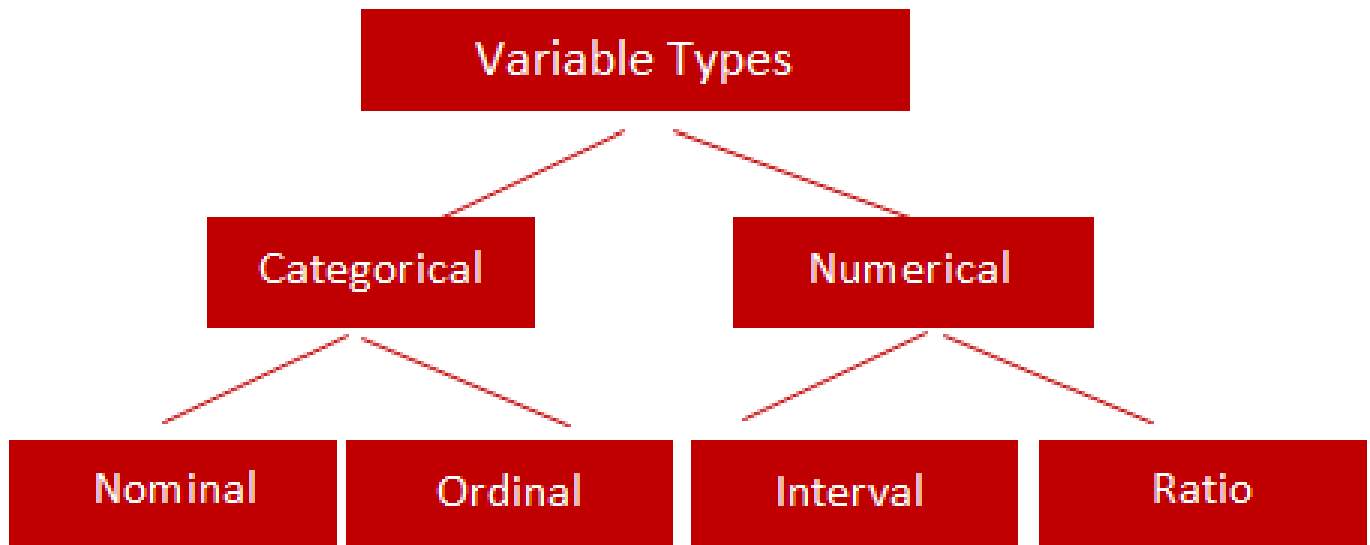
Around 4,000 people have been killed by drone strikes in Pakistan since 2004. **According to leaked documents on The Intercept**, these drone strikes happened based on results from the machine learning algorithm.

Model Result : A false positive rate of 0.18% across 55 million people would mean 99,000 innocents mislabelled as "terrorists"

Types of Variables

It is important to understand the different types of variables, because the type of variable can lead to different kinds of data and guide your analysis.

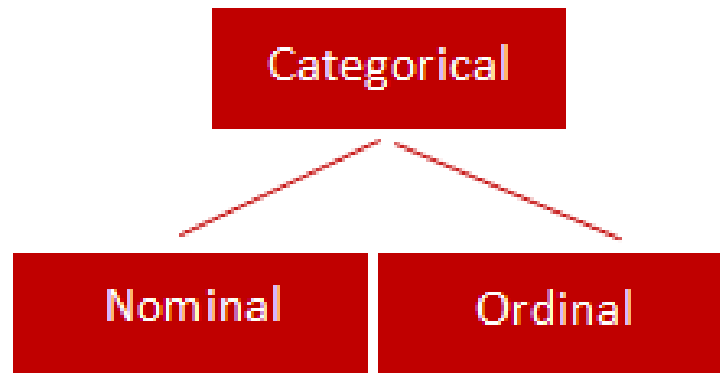
The following are types of variables:



Categorical

Qualitative data are often termed as **categorical data**. Data that can be added into **categories** according to their characteristics.

Categorical Variables



Nominal Variable (Unordered list)

A variable that has two or more categories, without any implied ordering.

Examples :

Gender - Male, Female
Marital Status - Unmarried, Married, Divorcee
State - New Delhi, Haryana, Illinois, Michigan

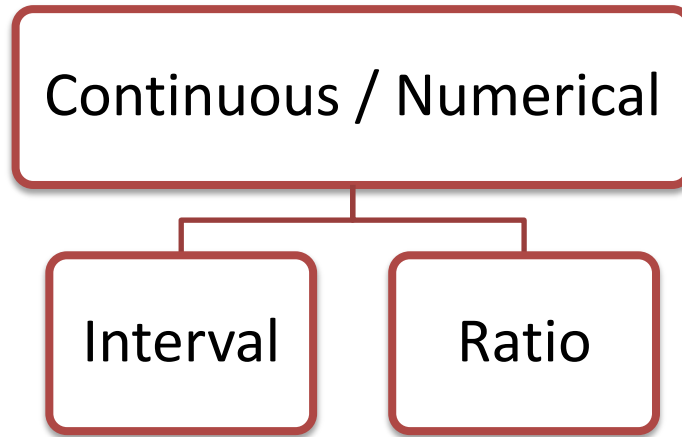
Ordinal Variable (Ordered list)

A variable that has two or more categories, with clear ordering.

Examples :

Scale - Disagree, Neutral, Agree
Rating - Low, Medium, High
Education Level – High School, Undergraduate, Graduate, PhD

Continuous Variables



Interval Variable

It is a measurement where the difference between two values is meaningful. When the variable equals 0.0, it means something.

Examples -

Temperature in Celsius means something.

Temperature of 30°C is higher than 20°C, and temperature of 20°C is higher than 10°C. The size of these intervals is the same.

Ratio Variable

It's an interval data with meaningful zero starting point. When the variable equals 0.0, there is none of that variable. Practically all quantitative variables are ratio variables..

Examples :

1. Monthly Income of Analysts
2. Age
3. Height
4. Weight (Someone who has 80kg weight is twice as heavy as someone who has 40kg weight)

Descriptive Statistics

Suppose you have data for students placed in campus placement.

Descriptive Statistics answers the following questions :

1. What is the **average** annual salary during campus placement?
2. How much **variation** in the annual salary from its average value?
3. What is the **lowest** and **highest** annual pay offered?

It provides information on summary statistics that includes Mean, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count.

Measure of Central Tendency

It describes a whole set of data with a single value that represents the centre of its distribution.

There are three main measures of central tendency:

1. Mean
2. Median
3. Mode

Mean	Average value
Median	Middle value
Mode	Most frequent value

Descriptive Statistics

Mean

It is the sum of the observations divided by the number of observations

The mean of the values 5,6,6,8,9,9,9,9,10,10 is
 $(5+6+6+8+9+9+9+9+10+10)/10 = 8.1$

Limitation :

It is affected by extreme values. Very large or very small numbers can distort the answer

Median

It is the middle value. It splits the data in half. Half of the data are above the median; half of the data are below the median.

First step is to arrange all the numbers from smallest to greatest. **If there is an odd number of numbers, the middle one is picked.**

For example, consider the set of numbers

1, 3, 3, 6, 7, 8, 9

This set contains seven numbers. The median is the fourth of them, which is 6.

If there are an even number of observations, then there is no single middle value; **the median is then usually defined to be the mean of the two middle values.**

Descriptive Statistics

For example, in the data set
1, 2, 3, 4, 5, 6, 8, 9

The median is the mean of the middle two numbers: this is $(4 + 5) \div 2$, which is 4.5.

Advantage :

It is **NOT** affected by extreme values. Very large or very small numbers does not affect it

Mode

It is the value that occurs most frequently in a dataset

Advantage :

It can be used when the data is not numerical.

Disadvantage :

1. There may be no mode at all if none of the data is the same
2. There may be more than one mode

Note –

There may be no mode if no value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or four or more modes (multimodal).

Descriptive Statistics

When to use mean, median and mode?

Mean – When your data is not skewed i.e normally distributed. In other words, there are no extreme values present in the data set (Outliers).

Median – When your data is skewed or you are dealing with ordinal (ordered categories) data (e.g. likert scale 1. Strongly dislike 2. Dislike 3. Neutral 4. Like 5. Strongly like)

Mode - When dealing with nominal (unordered categories) data.

OK to compute	Nominal	Ordinal	Interval	Ratio
median and percentiles	No	Yes	Yes	Yes
add or subtract	No	No	Yes	Yes
mean, standard deviation	No	No	Yes	Yes
ratio	No	No	No	Yes

Test your knowledge

1. Suppose a shoe company is considering expanding into an area and is studying the size of shoes that competitors are offering.

Shoe size	5	6	7	8	9
Frequency	2	5	11	4	1

Which metrics – mean / median / mode, would they be more interested? And why?

2. Can median be calculated by sorting data in descending order rather than ascending?

3. If there are two modes, you should _____.

- i. calculate the average
- ii. report the higher of the two numbers
- iii. report both
- iv. report the one closer to the mean