Data analysis and interpretation

Lecture 1

Based on Sheldon M Ross

Statistics

- Statistics: art of learning from data
- Statistics: collection of data, description of data, analysis of data, and, drawing of conclusions from data
- Our course title: Data analysis and interpretation

Statistics

- Not so much explicit emphasis on data collection and description
- Implicitly assumed that the data collection is proper and you know how to describe it
- We will brush up on these aspects before proceeding to the course content: in a couple of lectures and tutorials

Data collection

- Data not available:
  - Methods of designed experiment
  - Sampling methods
    - Suppose I want to compare two different methods of teaching: how do I divide this class for collecting data on that?

Data collection

- Class division into groups: random
- Then, the data are unbiased
- The test scores of the two groups: will indicate the efficacy of the teaching method

Descriptive statistics

- Description and summarising data
- Will give lots of hands-on exercises
- The best way to learn: doing it yourself

Inferential statistics

- Drawing conclusions from data
- Suppose average scores of one group is a lot higher than the other: chance or genuine?
- Probability model: assumption about chances of obtaining different data values

Theory of probablility

- We live with chance…
- Probability is a measure of “Chance”
- Needed to better understand whether the events occurring are due to chance or by design.
- Very basic measure to understand events occurring by chance and inferring on such events’ future occurrence.

Definitions

- Event: is a happening or an outcome of an experiment
- Population: total collection of elements
- Sample: Subgroup chosen for examination
- Question: Is the sample representative?

  Can we consider air-travellers representative of the citizens who voted population?

Random Event

- Random Event: An event that occurs by chance.
  - do not worry about getting the proportions right (say 50% women) – leave it to -“chance”

Descriptive statistics

- How to describe and summarise data?
- Tables and graphs

An example: The Hindu, 4.8.2019
Tabulated data

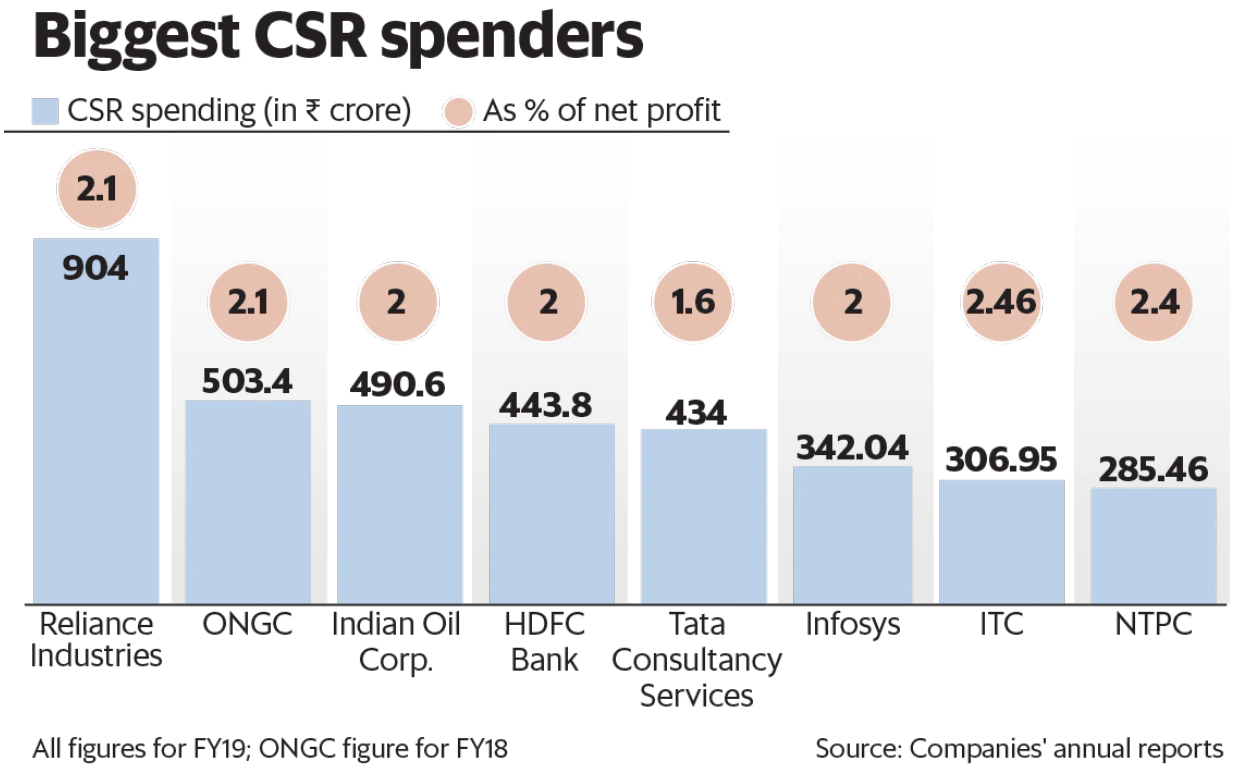
On the fast track

The Home Ministry cleared action against 400 terror suspects in 3 years. The maximum clearances were given in June 2019

Month/year and the number of prosecutions sanctioned:

June 2019: 44	August 2018: 16	April 2017: 27
May 2019: 12	April 2018: 7	March 2017: 9
April 2019: 19	March 2018: 23	December 2016: 10
March 2019: 21	February 2018: 9	November 2016: 7
February 2019: 31	December 2017: 2	September 2016: 2
January 2019: 19	November 2017: 6	August 2016: 6
November 2018: 17	August 2017: 22	July 2016: 38
September 2018: 23	June 2017: 7	June 2016: 28

Another example: Mint, 1.8.2019 Graph



Data representation

- Frequency table, line graph, bar graph, frequency polygon, relative frequency table and graph, pie chart, grouped data, histograms, Ogives, stem and leaf plots, sample percentiles and box plots
- We will have a tutorial some time this or next week

Measures of Central Tendency

- Average / Mean
  - **Arithmetic Mean**
  - Geometric Mean
  - Harmonic mean
- Median
- Mode

Arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median: centre of the data

- Order the data values
- If the number of data points is odd, sample median is the value in the position (n+1)/2
- If the number of data points is even, sample median is the average of values in positions n/2 and n/2+1

Mean and median

- Mean: affected by extreme values
- Median: not affected by extreme values
- Mean or median: depends on what you need

Mode

- Value that occurs with the highest frequency
- If no single value occurs most frequently, the modal values are all the values that occur at the highest frequency

Measures of Dispersions

- Variance / Standard Deviation
- Range

Sample variance

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n-1}$$

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Sample variance

- How to prove the algebraic identity in the previous slide?
- What is the effect of adding a constant to all data points on sample variance?
- What is the effect of multiplying all data points by a constant on sample variance?
- Hint: consider $y = a\, x_i + b$

Sample standard deviation

- Positive square root of the sample variance

Chebyeshev inequality

- Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set
- Assume $s>0$
- Chebyeshev inequality states that for any $k \geq 1$
- greater than $100\left(1-1/k^2\right)$ percent of the data lies within the interval $\bar{x}-ks$ to $\bar{x}+ks$

Chebyeshev inequality

- Note the word greater than
- This is the lower limit
- It can be sharpened based on the specification of the data set