Research Methodology: Analysing Quantitative Data

Yesoda Bhargava

January 8, 2021

Introduction

- What is quantitative research?
- Numbers are meaningless unless we can discern patterns that lie within them.
- **Statistics**: A group of computational procedires that enable us to find patterns and meaning in numerical data.
- *What do the data mean?* What message do they communicate?

Exploring and Organizing a data set

- Before making a single computation-look closely at your data.
- Consider potentially productive ways of organizing them.
- Why?
- Example: Below are the marks of few students in reading test.
- Adam : 76, Alice : 80, Bill: 72, Chuck: 68, Kathy: 84, Margaret: 88, Mary: 92, Ralph: 64, Robert: 60, Ruth: 96, Tom: 56.
- What do you observe?
- What if I organize it as per gender?
- **Girls** Alice : 80, Kathy: 84, Margaret: 88, Mary: 92, Ruth: 96
- **Males** Adam : 76, Bill: 72, Chuck: 68, Ralph: 64, Robert: 60, Tom: 56
- Observations?
- How can we graphically see the difference, if any?

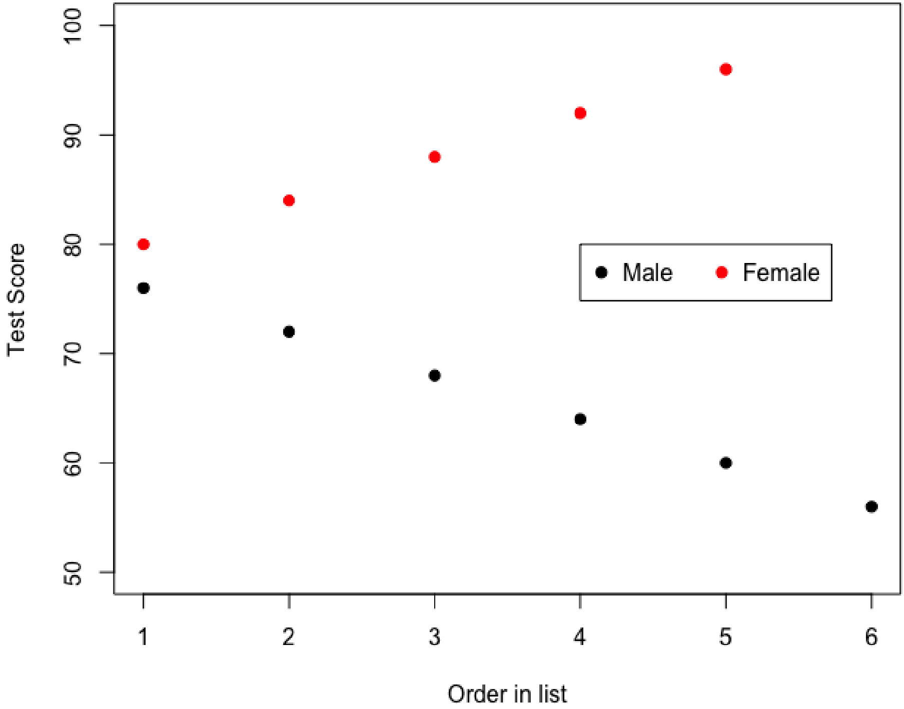

Figure 1: Visual representation of the Reading Achievement Test Scores.

Fundamental principle about data exploration

How the researcher prepares the data for inspection or interpretation will affect the meaning that those data reveal. Therefore, every researcher should be able to provide a clear, logical rationale for the procedure used to arrange and organize the data

Organizing data to make them easier to think about and interpret

- Graphing data is often quite useful for revealing patterns in the data set.
- It makes the nature of pattern more transparent and obvious.

Choosing appropriate statistics

- If we are to summarize the test scores, we can use *mean*, or, an *average*.
- Compute mean of the test scores in the previous example.
- How do you interpret it?
- Compute mean of girls and boys separately. Interpret it.

Functions of Statistics

- Mainly tow major functions.
- **Descriptive Statistics** : Description of what the data look like, where their center or midpoint is, what is the spread of the data and how closely two or more variables within the data are associated with each other, etc.
- **Inferential Statistics**: Allow us to draw *inferences* about large populations by collecting data on relatively small samples.
- Inferential Statistics involve using one or more small samples and then *estimating* the characteristics of the population from which each sample has been drawn.
- Example: Estimation of the population mean from the sample mean.
- You sample a group of first-year post-graduate students in Pune and try to estimate their CGPA in the under-graduate. From this sample, you wish to say something about the population of first-year post-graduate students in Pune.
- Inferential statistics helps to make **reasonable guesses** about a large, unknown population by examining a sample that is known.

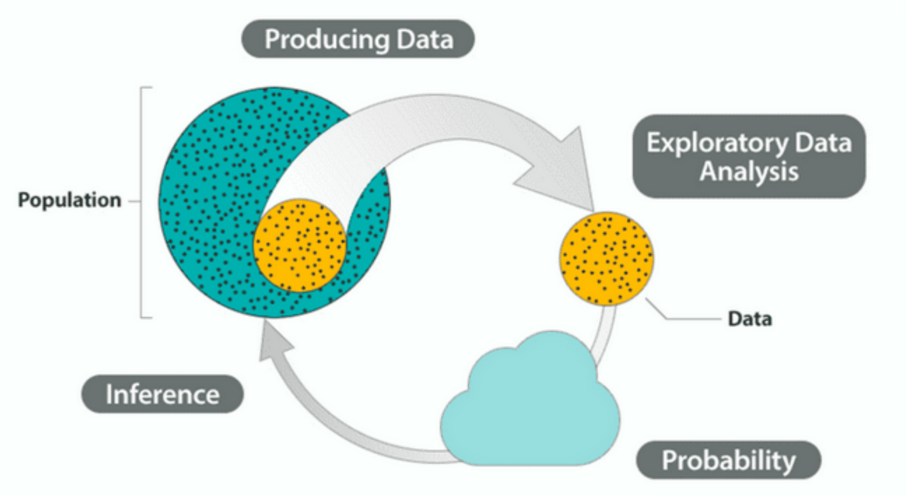

Figure 2: The Big picture of statistics.

Statistics as estimates of population parameters

- When statistics is used to draw *inferences* about a population from which a sample has been drawn, we are using them as *estimates of population parameters*.
- **Parameter**: characteristics/quality of a population, in-concept is a constant.
- Through statistics we want to draw a conclusion about this *parameter*.
- Sample is drawn to *estimate the parameter* and any calculation on sample (Sample mean, sample standard deviation etc.) is a **statistic**.
- Statisticians distinguish between **population parameter** and **sample statistic** by using different symbols.

Table 1: Examples of conventional statistical notations for population parameters ans sample statistics

The Characteristic in Question	Population parameter	Sample Statistic
The mean	μ	M or $\bar{X}$
The standard deviation	σ	s or SD
Proportion or probability	P	p
Number or total	N	n

Considering the Nature of the Data

- Consider whether your data is:
  - continuous or discrete.
  - Nominal, ordinal, interval, or ratio scale.
  - Reflect a normal or non-normal distribution.
- Continuous Versus Discrete Variables
  - A continuous variable reflects an infinite numbers of possible values failing along a particular continuum.
  - Example: chronological age.
  - Discrete variable has a finite and small number of possible values. Example: student's high school grade level, A student can be either in 9th, 10th, 11th or 12th.
  - A student cannot be in 9.25th grade.

Nominal, Ordinal, Interval, and Ratio Data

- **Nominal Data**: Numbers used only to identify different categories. Eg. Coding males and females or political party affiliation : Republicans, Democrats, Other affiliation.
- In neither case do the numbers indicate that participants have more or less of something.
- **Ordinal Data**: Those for which the assigned numbers reflect an order or sequence. They indicate the degree to which people, objects, or other entities have certain quality or characteristics (a variable) of interest.
- But they do not however, tell us anything about how great the differences are between the people, objects, or other entities.

Ratio Data

- **Interval data** : As is true for ordinal data, the numbers reflect differences in degree or amount.
- But in addition, differences between the numbers tell us *how much differences* exists in the characteristics being measured.
- Limitation of interval data is that a value of zero (0) does not necessarily reflect a complete lack of characteristic being measured.
- For example: IQ score of zero doesn't mean that a person has no intelligence whatsoever.
- **Ratio data**: are similar to interval data, they reflect equal intervals between values for the characteristics being measured.
- They have a true zero point: A value of zero tells us that there's a complete absence of the characteristics.
- Example: The difference between income level of 30,000 and 20,000 is the same as that between 40,000 and 30,000.
- People with income of 0 per year have no income.

Normal and Non-Normal Distributions

- Normal distributions or normal curve, also called as the *bell-curve*, have several distinguishing characteristics:
  - **It is horizontally symmetrical.**
  - **Its highest point is at its midpoint.**
  - **Predictable percentages of population lie within any given portion of the curve.**
- Example: Wheat production in the state of Delhi in India.
- If we would survey all farmers in the area and find out their per-acre wheat yield ina given year - we would find that few farmers have an unusually small wheat per acre whereas others may have bountiful.
- Reason: "That's the way it happened. "
- This is how nature behaves.

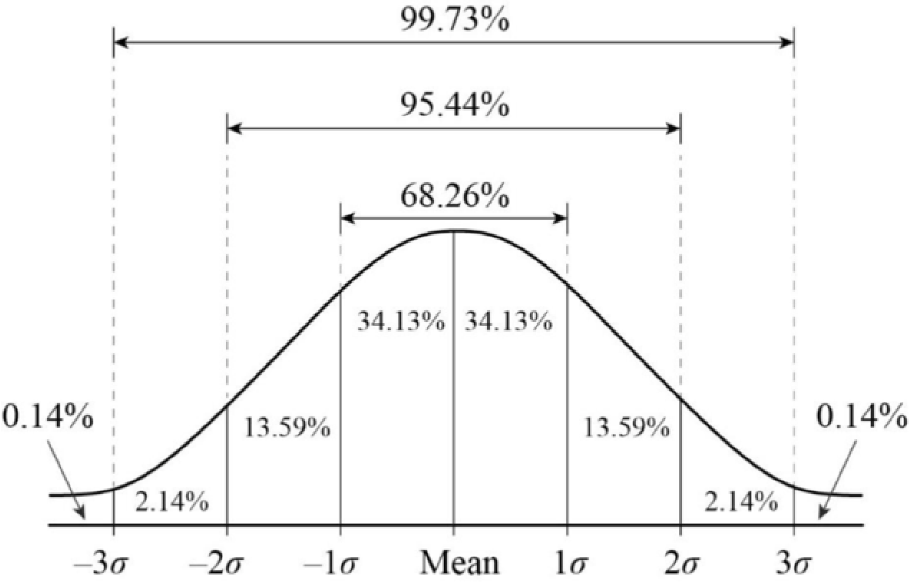

Figure 3: Percentages within each portion of the normal distribution.

Skewed Distribution

- Data don't reflect a normal distribution.
- "Skew" is the part of the distribution that stretches out a bit to one side.
- Positively skewed: If the peak lies to the left of the midpoint.
- Negatively skewed: If the peak lies to the right of the midpoint.
- Some data sets do not remotely resemble a normal distribution. Eg. Ordinal data, by virtue of how they are created, *never* fall into a normal distribution.
- Example: Giving each student a class rank according to the academic grade point average.
- There is only one student at each academic rank.

Negative Skewness

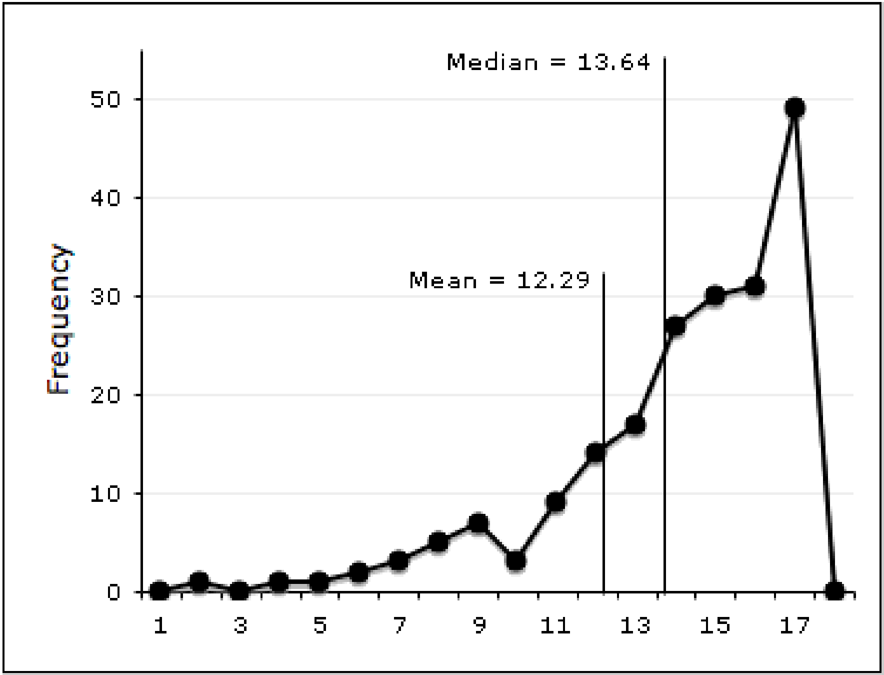

Figure 4: The distribution shown has a negative skew. The mean is smaller than the median.

Positive Skewness

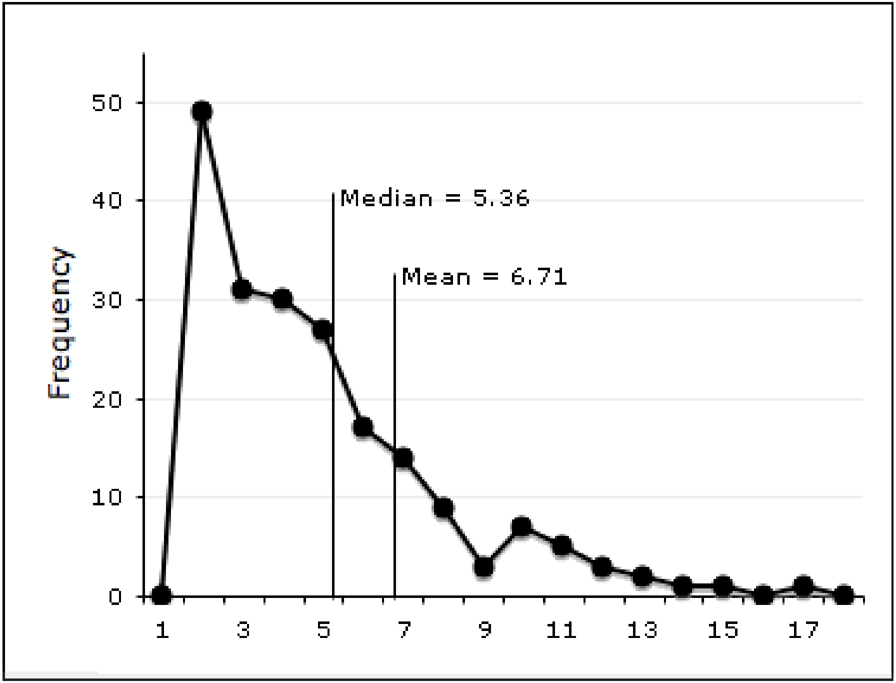

Figure 5: The distribution shown has a positive skew. The mean is bigger than the median.

Parametric and Non-parametric Statistics

- Choice of statistical procedures depends to some degree on the nature of the data set and the extent to which they reflect normal distribution.
- **Parametric Statistics** are based on certain assumptions about the nature of the population in the question:
  - The data reflect an interval or ratio scale.
  - The data fall in a normal distribution (e.g. the distribution has a central high point, and it is not seriously skewed).
- When either of these assumptions are violated, the results one obtains from parametric statistics can be flawed.
- **Nonparametric statistics** are not based on such assumptions. Applicable for data that is ordinal in nature rather than interval in nature.
- Most complex and powerful inferential statistics are based on parametric statistics.
- Non-parametric statistics are appropriate only for relatively simple analyses.

Descriptive Statistics

- Descriptive statistics *describe* a body of data.
- Three things a researcher might want to know about a data set:
  - points of central tendency
  - amount of variability
  - extent to which two variables are associated with each other

Measures of Central Tendency

- A *point of central tendency* is a point around which data revolve.
- Three commonly used measures of central tendency are the mode, the median, and the mean.
- **Mode**: single number or score that occurs most frequently.
- Find mode in this data set:
- 3, 4, 6, 7, 7, 9, 9, 9, 9, 10, 11, 11, 13, 13, 13, 15, 15, 21, 26
- Mode is the *only* appropriate measure of central tendency for nominal data.
- **Median**: Numerical center of a set of scores/data points.
- It is the number in the very middle of the values, with exactly as many values above it as below it.
- Find median of above data set.
- **Mean** : The arithmetic average of values of a certain variable.

Measures of Variability: Dispersion and Deviation

- To understand the tendency of the data to be closely clustered around the mean or widely dispersed around the mean.
- Computed using standard deviation/variance.

$$\text{standard deviation}(s) = \sqrt{\frac{\sum(X-M)^2}{N}} \tag{1}$$

$$Variance(s^2) = \frac{\sum(X-M)^2}{N} \tag{2}$$

- Higher the value of these measures, higher is the spread/variability in the data.

Dispersion

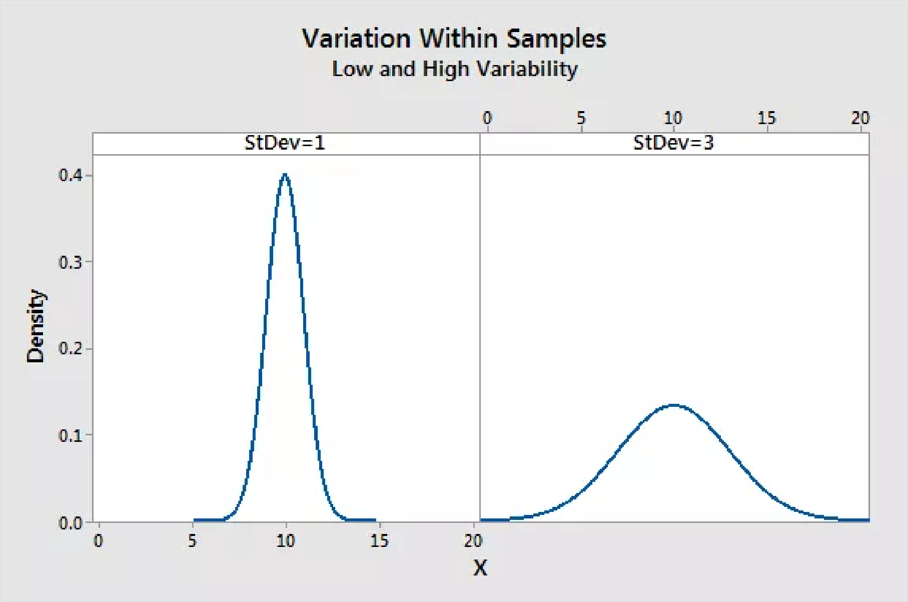

Figure 6: Variation with Samples. Low and high Variability.

Keeping measures of central tendency and variability in perspective

- Statistics related to central tendency and variability help us summarize the data.
- The ultimate goal in conducting research is not statistical manipulation but the *interpretation of the data*.
- Finding the medians, mean and variability do not mean research because we have not extracted any *meaning* from the data.
- This is where measures of association comes into picture: correlation.
- Correlation allows us to know the direction and strength of linear relationship between two variables.
- There is different type of correlation depending upon the nature of the data.
- Most common is the Pearson correlation coefficient between two continuous data.
- In-depth discussion on inferential statistics is out of the scope of this module.

Upcoming lecture

- In the upcoming lecture we shall see basic operations using two continuous variables, one continuous and one categorical variable and both categorical variables using R.
- Concepts covered would be: mean, variance, scatter plot, correlation, contingency table, box plots.
- Last three lectures will be on Reference Manager, Systematic Review, Poster Presentation.