INTRODUCTION TO DATA ANALYTICS

Class #4

Descriptive Statistics

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

- Change your thoughts and you change your world.
    - NORMAN VINCENT PEALE, American - Clergyman

TODAY'S DISCUSSION…

- Introduction
- Data summarization
  - Measurement of location
    - Mean, median, mode, midrange, etc.
  - Measure of dispersion
    - Range, Variance, Standard Deviation, etc.

TRP: AN EXAMPLE

- Television rating point (TRP) is a tool provided to judge which programs are viewed the most.
  - This gives us an index of the choice of the people and also the popularity of a particular channel.
- For calculation purpose, a device is attached to the TV sets in few thousand viewers' houses in different geographic and demographic sectors.
  - The device is called as **People's Meter**. It reads the time and the programme that a viewer watches on a particular day for a certain period.
- An average is taken, for example, for a 30-days period.
- The above further can be augmented with a personal interview survey (PIS), which becomes the basis for many studies/decision making.
- Essentially, we are to analyze **data** for TRP estimation.

DEFINING DATA

Definition 3.1: **Data**

A set of data is a collection of observed values representing one or more characteristics of some objects or units.

Example: For TRP, data collection consist of the following attributes.

- **Age:** A viewer's age in years
- **Sex:** A viewer's gender coded 1 for male and 0 for female
- **Happy:** A viewer's general happiness
  - NH for not too happy
  - PH for pretty happy
  - VH for very happy
- **TVHours:** The average number of hours a respondent watched TV during a day

DEFINING DATA

Viewer#	Age	Sex	Happy	TVHours
…	…	…	…	…
…	…	…	…	…
55	34	F	VH	5
…	…	…	…	…

Note:

- A data set is composed of information from a set of units.
- Information from a unit is known as an observation.
- An observation consists of one or more pieces of information about a unit; these are called variables.

DEFINING POPULATION

Definition 3.2: **Population**

A population is a data set representing the entire entities of interest.

Example: All TV Viewers in the country/world.

Note:

1. All people in the country/world is not a population.
2. For different survey, the population set may be completely different.
3. For statistical learning, it is important to define the population that we intend to study very carefully.

DEFINING SAMPLE

Definition 3.3: **Sample**

A sample is a data set consisting of a population.

Example: All students studying in Class XII is a sample, whereas those students belong to a given school is population.

Note:

- Normally a sample is obtained in such a way as to be representative of the population.

DEFINING STATISTICS

Definition 3.4: **Statistics**

A statistics is a quantity calculated from data that describes a particular characteristics of a sample.

Example: The sample mean (denoted by $\bar{y}$) is the arithmetic mean of a variable of all the observations of a sample.

DEFINING STATISTICAL INFERENCE

Definition 3.5: **Statistical inference**

Statistical inference is the process of using sample statistics to make decisions about population.

Example: In the context of TRP

- Overall frequency of the various levels of happiness.
- Is there a relationship between the age of a viewers and his/her general happiness?
- Is there a relationship between the age of the viewer and the number of TV hours watched?

DATA SUMMARIZATION

- To identify the typical characteristics of data (i.e., to have an overall picture).
- To identify which data should be treated as noise or outliers.
- The data summarization techniques can be classified into two broad categories:
  - Measures of **location**
  - Measures of **dispersion**

MEASUREMENT OF LOCATION

- It is also alternatively called as measuring the central tendency.
  - A function of the sample values that summarizes the location information into a single number is known as a measure of location.

- The most popular measures of location are
  - Mean
  - Median
  - Mode
  - Midrange

- These can be measured in three ways
  - Distributive measure
  - Algebraic measure
  - Holistic measure

DISTRIBUTIVE MEASURE

- It is a measure (*i.e. function*) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (*i.e. entire*) data set.

Example

- ✔ sum(), count()

ALGEBRAIC MEASURE

- It is a measure that can be computed by applying an algebraic function to one or more distributive measures.

- Example

$$\text{average} = \frac{\text{sum}()}{\text{count}()}$$

HOLISTIC MEASURE

- It is a measure that must be computed on the entire data set as a whole.

- Example

  Calculating median

  What about *mode?*

MEAN OF A SAMPLE

- The mean of a sample data is denoted as $\bar{x}$. Different mean measurements known are:
  - Simple mean
  - Weighted mean
  - Trimmed mean

- In the next few slides, we shall learn how to calculate the mean of a sample.

- We assume that given $x_1, x_2, x_3, \ldots, x_n$ are the sample values.

SIMPLE MEAN OF A SAMPLE

- **Simple mean**

  It is also called simply arithmetic mean or average and is abbreviated as (AM).

Definition 3.6: **Simple mean**

- ✓ If $x_1, x_2, x_3, \ldots, x_n$ are the sample values, the simple mean is defined as

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

WEIGHTED MEAN OF A SAMPLE

- **Weighted mean**

  It is also called weighted arithmetic mean or weighted average.

Definition 3.7: **Weighted mean**

When each sample value x_i is associated with a weight w_i, for $i = 1,2,\ldots,n$, then it is defined as

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

Note

When all weights are equal, the weighted mean reduces to simple mean.

TRIMMED MEAN OF A SAMPLE

- **Trimmed Mean**

  If there are extreme values (*also called outlier*) in a sample, then the mean is influenced greatly by those values. To offset the effect caused by those extreme values, we can use the concept of trimmed mean

Definition 3.8: **Trimmed mean**

Trimmed mean is defined as the mean obtained after chopping off values at the high and low extremes.

PROPERTIES OF MEAN

- **Lemma 4.1**

  If $\bar{x}_i$, $i = 1,2,\ldots,m$ are the means of m samples of sizes $n_1, n_2, \ldots, n_m$ respectively, then the mean of the combined sample is given by:-

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

(Distributive Measure)

- **Lemma 4.2**

  ✓ If a new observation x_k is added to a sample of size n with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\bar{x} + x_k}{n+1}$$

PROPERTIES OF MEAN

- **Lemma 4.3**

  If an existing observation x_k is removed from a sample of size n with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\,\bar{x} - x_k}{n-1}$$

- **Lemma 4.4**

  If m observations with mean $\bar{x}_m$, are added (*removed*) from a sample of size n with mean $\bar{x}_n$, then the new mean is given by

$$\bar{x} = \frac{n\,\bar{x}_n \pm m\,\bar{x}_m}{n \pm m}$$

PROPERTIES OF MEAN

- **Lemma 4.5**

  If a constant c is subtracted (*or added*) from each sample value, then the mean of the transformed variable is linearly displaced by c. That is,

$$\bar{x}' = \bar{x} \mp c$$

- **Lemma 4.6**

  If each observation is called by multiplying (*dividing*) by a non-zero constant, then the altered mean is given by

$$\bar{x}' = \bar{x} * c$$

  *Where, * is x (multiplication) or ÷ (division) operator.*

MEAN WITH GROUPED DATA

Sometimes data is given in the form of classes and frequency for each class.

Class □			…..		…..	
Frequency □			…..		…..	

There are three methods to calculate the mean of such a grouped data.

- Direct method
- Assumed mean method
- Step deviation method

DIRECT METHOD

Direct Method

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

Where, $\mathbf{x_i} = \frac{1}{2}$ **(lower limit + upper limit)** of the i^{th} class, i.e., $x_i = \frac{x_i + x_{i+1}}{2}$ (also called class size), and f_i is the frequency of the i^{th} class.

Note

$$\sum f_i (x_i - \bar{x}) = 0$$

ASSUMED MEAN METHOD

- **Assumed Mean Method**

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

where, A is the assumed mean (it is usually a value $x_i = \frac{x_i + x_{i+1}}{2}$ chosen in the middle of the groups $d_i = (A - x_i)$ for each i)

STEP DEVIATION METHOD

- **Step deviation method**

$$\bar{x} = A + \left\{ \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \; h \right\}$$

where,

A = assumed mean

h = class size (*i.e.*, $x_{i+1} - x_i$ *for the* i^{th} *class*)

$$u_i = \frac{x_i - A}{h}$$

MEAN FOR A GROUP OF DATA

- For the above methods, we can assume that…
  - All classes are equal sized

*Data with **in**clusive classes*

10 - 19	**20 - 29**	**30 - 39**	

*Data with **ex**clusive classes*

9.5 – 19.5	**19.5 – 29.5**	**29.5 – 39.5**	

OGIVE: GRAPHICAL METHOD TO FIND MEAN

- **Ogive** (pronounced as O-Jive) is a cumulative frequency polygon graph.
  - When cumulative frequencies are plotted against the upper (lower) class limit, the plot resembles one side of an Arabesque or **ogival** architecture, hence the name.
  - There are two types of Ogive plots
    - Less-than (upper class vs. cumulative frequency)
    - More than (lower class vs. cumulative frequency)

Example:

Suppose, there is a data relating the marks obtained by 200 students in an examination

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, …….

(Further, suppose it is observed that the minimum and maximum marks are 410, 479, respectively.)

OGIVE: CUMULATIVE FREQUENCY TABLE

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, …….

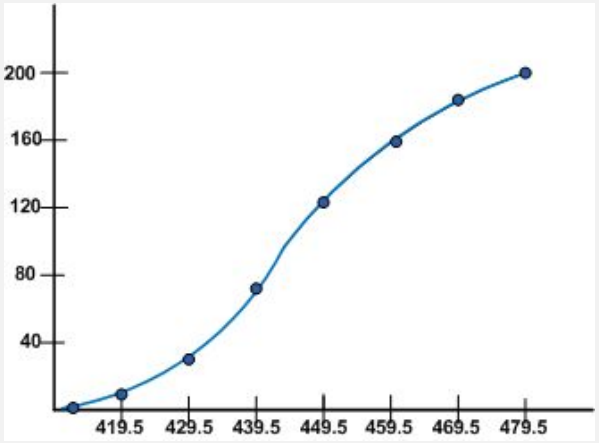
Step 1: Draw a cumulative frequency table

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

OGIVE: GRAPHICAL METHOD TO FIND MEAN

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Upper class	Cumulative Frequency
Less than 419.5	14
Less than 429.5	34
Less than 439.5	76
Less than 449.5	130
Less than 459.5	175
Less than 469.5	193
Less than 479.5	200

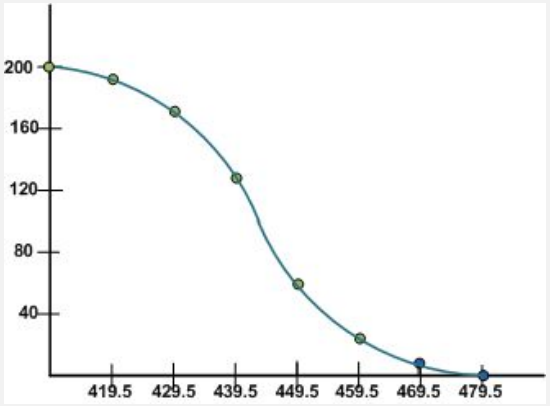


Step 2: Less-than Ogive graph

OGIVE: GRAPHICAL METHOD TO FIND MEAN

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

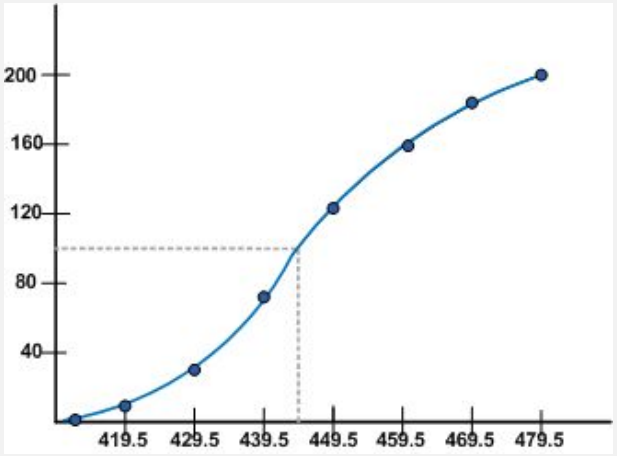
Upper class	Cumulative Frequency
More than 409.5	200
More than 419.5	186
More than 429.5	166
More than 439.5	124
More than 449.5	70
More than 459.5	25
More than 469.5	7



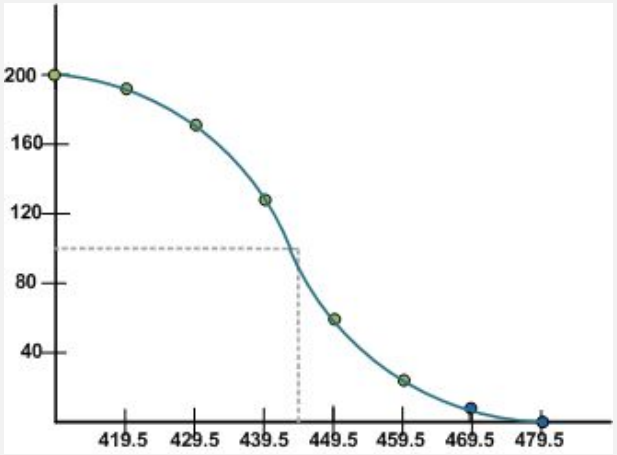
Step 3: More-than Ogive graph

INFORMATION FROM OGIVE

- Mean from Less-than Ogive

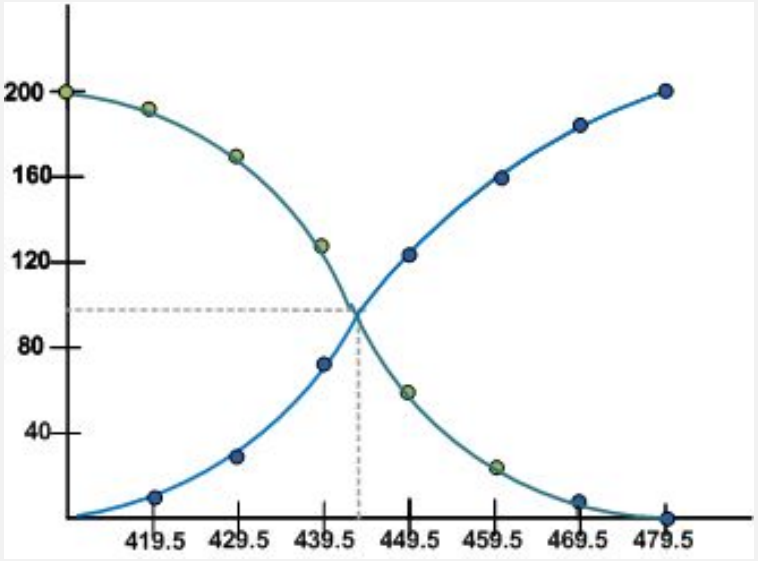


- Mean from More-than Ogive



INFORMATION FROM OGIVE

- Less-than and more-than Ogive approach



A cross point of two Ogive plots gives the mean of the sample

SOME OTHER MEASURES OF MEAN

- There are three mean measures of location:
  - Arithmetic Mean (AM)
  - Geometric mean (GM)
  - Harmonic mean (HM)

SOME OTHER MEASURES OF MEAN

- Arithmetic Mean (**AM**)
  - $S: \{x_1, x_2\}$
  - $\bar{x} = \frac{x_1 + x_2}{2}$
  - $\bar{x} - x_1 = x_2 - \bar{x}$
- Geometric mean (**GM**)
  - $S: \{x_1, x_2\}$
  - $\tilde{x} = \sqrt{x_1 . x_2}$
  - $\frac{x_1}{\tilde{x}} = \frac{\tilde{x}}{x_2}$
- Harmonic Mean (**HM**)
  - $S: \{x_1, x_2\}$
  - $\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$
  - $\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}$

REFERENCE

- The detail material related to this lecture can be found in

  Probability and Statistics for Enginneers and Scientists (8th Ed.) by Ronald E. Walpol, Sharon L. Myers, Keying Ye (Pearson), 2013 .

QUESTIONS OF THE DAY…

1. Which of the following central tendency measurements allows distributive, algebraic and holistic measure?
   - mean
   - median
   - Mode

   Which measure may be faster than other? Why?