Lecture-1

Statistics (Definition, why we need statistics) (R language was mainly designed for Statistics)
Predictions.

* Descriptive Satistics

- Gives numerical and graphic procedures to summarize a collection of data in a clear and understandable way.
- Data properties and behaviour বোঝার জন্য we করা হয়,

Outlier: a data point that differs significantly from other observations.

* Inferential Satistics

- Provides procedures to draw inferences about a population from a sample.
- কোনো decision নেওয়া.

Descriptive statistics we কার কোনো decision → আর inferential statistics. Predictive decision.

Ex: Gmail এ spam → inferential statistics we করা হয় to identify if it is spam mail or not. word pattern, বা QR check করে যেটা spam কিনা decide করে gmail. আবার সময় important ও spam → চলে যায়,

Lecture-2

Descriptive Measures

* Central Tendency measures : Mean বা Average.

They are computed to give a "center" around which the measurements in the data are distributed.

A : 100
B : ৯৫
C : ৫
D : 7
E : 10
F : 6
G : 9

$$\text{Average} = \frac{100+৯৫+৫+7+10+6+9}{7}$$

$$\simeq 33$$

Real measure পাওয়া যায় না.

Average never handles extreme value.

Salman F. Rahaman - $100000

Normal people - $100

মাঝামাঝি যায় = Average.

Nature এর সবকিছু center এ থেকে চায়,

* Variation or Variability measures: standard deviation

They describe "data spread" or how far away the measurements are from the center.

এরকম data থেকে আরেকটা data-র পার্থক্য.
data কী পরিমান deviate করছে।
standard deviation যত কম তত ভালো, center tendency তত correct. data গুলো highly correlated.

* Relative standing measures:

They describe the relative position of specific measurements in the data.

Measures of Central Tendency (or Location)

* Mean (Extreme value handle করতে পারে না)

- Arithmetic mean : যখন সবগুলো data-র weight একই

- Weighted mean :

A : ৫০ → 3 hrs
100 → 2 hrs
100 → 1 hr

arithmetic mean = $\frac{50+100+100}{3}$

$3\times50 = 150$
$2\times100 = 200$
$1\times100 = 100$

weighted mean = $\frac{3\times50+2\times100+1\times100}{3}$

B : 100 → 3 hrs class
৫০ → 2 hrs
৫০ → 1 hr

arithmetic mean = $\frac{100+50+50}{3}$

weighted mean = $\frac{3\times100+2\times50+1\times50}{3}$

- Geometric mean :

* Median (মাঝামাঝি value)

data তে কোনো extreme value না থাকলে বা কম থাকলে mean and median কাছাকাছি হবে,

mean and median এর পার্থক্য কম মানে data গুলো balanced data গুলো ascending বা descending order এ থাকতে হবে.

* Mode

data set এ সবচেয়ে বেশি সংখ্যক value কোন value-র আছে সেটা।

Median and mode can handle extreme values.

Population :

The entire group about which information is desired.

Sample :

Sample is a subset of population.

A proportion or part of the population usually the proportion from which information is gathered.

n=3

p=7

15, 25, 60 < 100 -

sampling

15

25

60

119

125

1040

4049

Randomly choose করতে হবে data behaviour বোঝার জন্য .

Random sampling

Lecture-3

Populations and samples

* Good sampling is a sign where all the population can be represented.

* Population Mean

$$\mu = \frac{\sum X}{N}$$

For ungrouped data, the population mean is the sum of all population values divided by the total number of population values.

- μ is the population mean
- N is the total number of observations
- X is a particular value.
- $\sum$ indicates the operation of adding.

A parameter is a measurable characteristic of a population.

A statistic is a measurable characteristic of a sample.

The Kiers family owns four cars. The following is the current mileage on each of the four cars:

56000, 23000, 42000, 73000

$$\mu = \frac{\Sigma X}{N} = \frac{56000+23000+42000+73000}{4}$$

$$= 48500$$

Mean vs Average

Sample Mean

For ungrouped data, the sample mean is the sum of all the sample values divided by the number of sample values.

- n is the total number of values in the sample.

$$\bar{X} = \frac{\Sigma X}{n}$$

A sample of five executives ~~records~~ received the following bonus last year ($k):

14.0, 15.0, 17.0, 16.0, 15.0

$$\bar{X} = \frac{\Sigma X}{n} = \frac{14.0+15.0+17.0+16.0+15.0}{5} = 15.4$$

Arithmetic Mean Properties:

* Every set of interval-level and ratio-level data has a mean.
* All the values are included in computing the mean.
* A set of data has a unique mean.
* The mean is affected by unusually large or small data values.
* The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.
* Affected by extreme values (outliers).

Consider the set: 3, 8, 4

The mean is 5

$$\sum (X - \bar{X}) = [(3-5) + (8-5) + (4-5)] = 0$$

Weighted Mean

The weighted mean of a set of numbers $x_1, x_2, \ldots, x_n$, with corresponding weights $w_1, w_2, \ldots, w_n$ is computed from the following formula:

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

Course	Grade	Points (x)	Credits (f)
Math	A	4	5
History	B	3	3
Health	A	4	2
Art	C	2	2

$$\bar{x}_w = \frac{5 \times 4 + 3 \times 3 + 2 \times 4 + 2 \times 2}{5 + 3 + 2 + 2}$$

$$= 3.42$$

Geometric Mean

The geometric mean (GM) of a set of n numbers is defined as the nth root of the product of the n numbers.

The geometric mean is used to average percents, indexes, and relatives.

$$GM = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots \cdots X_n}$$

* Properties of root

The interest rate on three bonds were 5, 21, 4 percent.

$$\text{Geometric Mean} = \sqrt[3]{5 \times 21 \times 4} = 7.49$$

$$\text{Arithmetic Mean} = \frac{5+21+4}{3} = 10.0$$

The GM gives a more conservative profit figure because it is not heavily weighted by 21%.

center tendency ভালোভাবে represent করে,

Outlier কে overcome করতে পারে,

(যে dataset এ সব mean এর value same সেটা সবচেয়ে ভালো

The Median

Mean is highly sensitive to outliers or extreme values.

The median is the mid point of the values after they have been ordered ~~the~~ from the smallest to the largest.

- There are as many values above the median as below it in the data array.
- For an even set of values, the median will be the arithmetic average of the two middle numbers.

21, 25, 19, 20, 22

19, 20, 21, 22, 25

Median = 21

76, 73, 80, 75

73, 75, 76, 80

Median = 75.5

Properties of the Median

* There is a unique median for each data set.

* It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.

* It can be computed for ratio-level, interval-level, and ordinal-level data.

* It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.

** If the dataset is uniformly distributed then

mean = median = mode.

Standard deviation = 0

Lecture-4

Descriptive Statistics - data analysis করে decision নেয়া

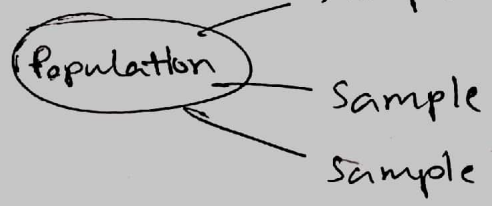
Inferential Statistics - data analysis করে mathematical model তৈরি করে future decision predict করা।

data analysis, prediction > statistics.

Statistic - value that describes samples

Parameter - value that describes a population.

Large pool of data source



Statistical analysis এর জন্য যে domain টা set করা হয় সেটাই population

Sampling করা হয় selectively and randomly.

Sample mean - Statistic

Population mean - Parameter.

Population mean, $\bar{X} = \frac{\sum x_i}{n}$

Sample mean, $\mu = \frac{\sum x_i}{n-1}$

Sample mean এ bias থাকে, bias দূর করার জন্য n-1 দিয়ে ভাগ করা হয়।

280, 288, 300, 309

$$\bar{x} = \frac{280+288+300+309}{4}$$

$$\bar{x} = 294.25$$

$A = 280$ d_i = dispersion / বিচ্যুতি

$d_i = x_i - A$

$d_i \rightarrow 0, 8, 20, 29.$

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

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

$$= \frac{\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} A.}{n}$$

$$= \frac{\sum d_i + nA}{n}$$

$$\therefore \bar{x} = A + \frac{\sum d_i}{n}$$

$$\bar{x} = 280 + \frac{0+8+20+29}{4}$$

$$= 294.25$$

$A = 67$ (x_i is middle)

Heights (in)	No. of Students (f_i)	Class Marks (x_i)	$f_i x_i$	$d_i = x_i - A$	$d_i f_i$
60-62	5	61	305	-6	-30
63-65	18	64	1152	-3	-54
66-68	42	67	2814	0	0
69-71	27	70	1890	3	81
72-74	8	73	584	6	48
			6745		45

$n = 100$

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

$$= \frac{6745}{100}$$

$$= 67.45$$

$$\bar{x} = A + \frac{\sum d_i f_i}{n}$$

$$= 67 + \frac{45}{100}$$

$$= 67.45$$

Lecture. 15

Median for ungrouped data, Median is $\frac{N+1}{2}$-th observation; N is odd.

Initial data : 20, 10, 23, 25, 100, 37, 105.

Sorted data : 10, 20, 23, 25, 37, 100, 105.

↑ Median or Q_2

Initial data : 20, 10, 23, 25, 100, 37, 105, 200

Sorted data : 10, 20, 23, 25, 37, 100, 105, 200

Median for ungrouped data, Median is average of $\frac{N}{2}$-th and $(\frac{N}{2}+1)$-th observation.

$$\frac{25+37}{2} = 31 \rightarrow \text{Median}$$

* Grouped data

weight	frequency
30-40	18
40-50	37
50-60	45
60-70	27
70-80	15
80-90	8

Left inclusive convention

↓

30-39

40-49

⋮

Class size = 10.

weight	frequency	Cumulative frequency
30-40	18	18
40-50	37	55 → $(\Sigma f)_l$
Median → 50-60	45 → f_{median}	100
60-70	27	127
70-80	15	142
80-90	8	150

N = 150

Median for grouped data:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - (\Sigma f)_l}{f_{median}} \right) \times c$$

L = Lower class boundary of median class

N = Total number of items

$(\Sigma f)_l$ = Sum of frequencies of all class lower than the median class.

f_{median} = frequency of median class.

c = size of median interval.

$\frac{150}{2} = 75$, $\frac{N}{2}$ median value class থাকবে, সেটা median class.

Median class = 50-60 lower limit = 50

$f_{median} = 45$

$(\Sigma f)_l = 55$

$c = 10$

$L = 49.5$

True lower limit কে lower class boundary বলা হয়।

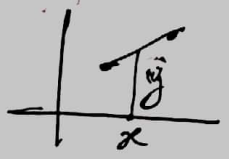
$L = \frac{50+49}{2} = 49.5$

$$\text{Median} = L + \left(\frac{\frac{N}{2} - (\Sigma f)_l}{f_{median}}\right) \times c$$

$$= 49.5 + \left(\frac{\frac{150}{2} - 55}{45}\right) \times 10$$

$$= 53.944.$$

[Linear interpolation মনে করা হয়।



Mode Calculation

৫ ২ ৪ 10 ৪ ৫ ৪

যে value র frequency সবচেয়ে বেশি সেটা mode.

৫ → 2

৪ → 3

10 → 1.

Mode = ৪. ; Uni-modal data

৫ ৪ 10 ৪ ৫ ৪ ৫

৫ → 3

৪ → 3

10 → 1

Mode = ৫, ৪ ; Bi-modal data.

দুইয়ের বেশি mode থাকলে multi modal data.

Lecture-6

1. Schaum series : statistic – Murray R. Spiegel
2. " " : Probability - Seymour

Qe.

CT Topic : Inferential, population, statistic
Descriptive sample parameter.

central tendency - arithmetic mean - grouped data.
geometric mean

Median for grouped and ungrouped data

Geometric Mean :

$$G = \sqrt[n]{x_1 . x_2 . x_3 . \ldots . x_n}$$

$$= (x_1 . x_2 . x_3 \ldots x_n)^{1/n}$$

$$\log G = \frac{1}{n} \log (x_1 . x_2 . x_3 \ldots x_n)$$

$$\log G = \frac{1}{2} [\log x_1 + \log x_2 + \ldots + \log x_n]$$

$$\log G = \frac{\sum f_i \log x_i}{n}$$

$$\Rightarrow G = 10^{\frac{\sum f_i \log x_i}{n}}$$

#

Weight	Frequency
30-40	18
40-50	37
→ 50-60	45
60-70	27
70-80	15
80-90	8

modal class

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

L = Lower class boundary of modal class

c = class size of modal class

Δ_1 = Difference between freq. of modal class and pre-modal class $= f_m - f_1$

Δ_2 = Difference between freq. of modal class and post-modal class $= f_m - f_2$

$L = 49.5$ $\quad f_m = 45$, $f_1 = 37$, $f_2 = 27$

$c = 10$

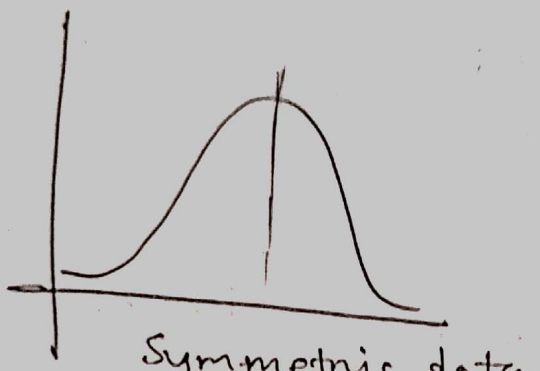
$\Delta_1 = 45 - 37 = 8$, $\quad \Delta_2 = 45 - 27 = 18$

$$\therefore \text{Mode} = 49.5 + \frac{8}{8+18} \times 10 = 52.58$$

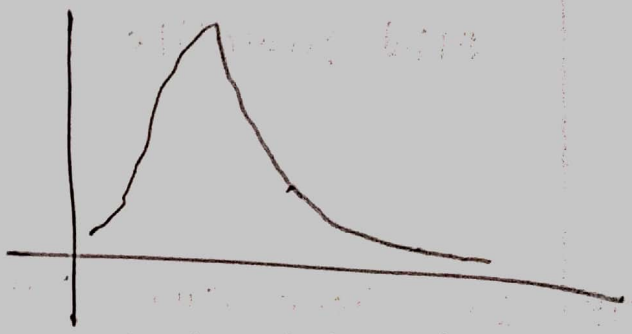
Empirical Rule:

For a moderately skewed data

Mode = 3 Median - 2Mean.



Symmetric data

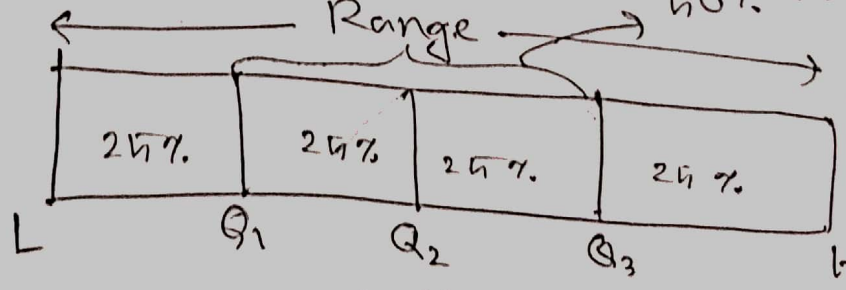


Moderately skewed data

not so symmetric
not so asymmetric

Quartiles: Quantiles are values that divide the total frequency into 4 parts.

Inter quantile range
50% central data
এটা important
কারণ outlier
detect করা যায়



Minimum value

Median

Maximum val

Range = H - L.

Q_1 = At most 25% data are smaller than Q_1 and at most 75% are larger.

Q_3 = At most 75% data are smaller than Q_3 and at most 25% are largen.

Inter quantile Range (IQR) = $Q_3 - Q_1$

Mid quartile = $\frac{Q_1 + Q_3}{2} = Q_2$ (not always)

If there are odd number of elements include median in two halves