MID 01

1) • Central tendency
   • trimmed mean
   • mode
   • mean absoluted deviation
   • standard Deviation
   • variance

(Hossain mim 2021-1-60-071)

2) Box Plot — 2 টি — ছবি সহ

3) • STANDARD ERROR (Example) Explain
   • BOOTSTRAP METHOD
   • Example of Bootstrap
   • Algorithm of Bootstrap method.

4) Hypothesis testing
   ⇓
   Chi-Square test এর type

CSE 303 Dmim sir MID01 Dmim sir

Date : Page : Hossain Mim 2021-1-60-07

(P-P1) 60, 70, 80, 75, 65, 70, 80, 70, 65, 65

(Ans) sorted Order :— 60, 65, 65, 65, 70, 70, 70, 75, 80, 80

Here, $n = 10$

$$\text{mean}(\bar{x}) = \frac{\sum x_i}{n} = \frac{700}{10} = 70 \quad \text{(Ans)}$$

$n = 10$

$$\text{median} : \frac{\left(\frac{n}{2}\right)\text{th value} + \left(\frac{n}{2}+1\right)\text{th value}}{2} = \frac{5\text{th value} + 6\text{th value}}{2} = \frac{70+70}{2} = 70 \quad \text{(Ans)}$$

MID01 CSE303

(P-P2)

60, 70, 80, 75, 65, 70, 100, 70, 65, 65

$$\text{mean}(\bar{x}) : \frac{720}{10} = 72 \quad \text{(Ans)}$$

For Trimmed mean :—

sorted the dataset → : 60, 65, 65, 65, 70, 70, 70, 75, 80, 100

For P = 1 50, 60 and 100 will be discard.

$$\text{mean}(\bar{x}) = \frac{\sum_{i=P}^{n-P} x_{i+1}}{(n-2P)} = \frac{560}{10-2} = \frac{560}{8} = 70 \quad \text{(Ans)}$$

weighted mean :— $x_i = 40, 45, 80, 75$ and 10 have

$w_i = 1, 2, 3, 4$ and 5

$$(\bar{x}w) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i} = \frac{(40\times1)+(45\times2)+(80\times3)+(75\times4)+(10\times5)}{1+2+3+4+5} = \frac{720}{15} = 48 \quad \text{(Ans)}$$

MID01 CSE303

Hossain Mim 2021-1-60-07

mean Absolution deviation :—

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

Dataset 1 : 30, 40, 50, 60, 70

$\bar{x} = 50$

$$MAD :- \frac{|(30-50)| + |40-50| + |50-50| + |(60-50)| + |70-50|}{5} = \frac{20+10+0+10+20}{5} = \frac{60}{5} = 12 \quad \text{(Ans)}$$

Date :
Page :

Median Absolute deviation:—

$MAD = \text{median}(|x_1 - m|, |x_2 - m|, \ldots\ldots |x_N - m|)$

Dataset :(2) 0 ; 25 , 50, 75 , 100 $n = 5$ (odd number)

Ans: MAD: median (|0 − 50| , |25 − 50| , |50 − 50| , |75 − 50| , |100 − 50|) $\left(\frac{n+1}{2}\right)$ th

= median (50, 25 , 0 , 25 , 50) $= \left(\frac{5+1}{2}\right)$ $\left(\frac{6}{2}\right)$ 3th value

= median (0, 25, 25, 50, 50)

Middle one (3)

$n = 5$ (odd)

$= \left(\frac{n+1}{2}\right)$ th value.

$= \left(\frac{5+1}{2}\right)$ th

= 3th value

= 25 Ans

variance :— (σ^2) $\sigma^2 = \left(\frac{|x - \bar{x}|}{n}\right)$

Example:— The Highest : 600 mm, 470 mm, 170 mm, 430 mm, 300 mm

mean ($\bar{x}$) : $\frac{1970}{5} = 394$

variance (σ^2) : $\frac{(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-}{5}$

= 21704 Ans

Middle one (3)

standard deviation : (σ) = $\sqrt{21704}$

= 147.32 Ans

sample variance (s^2) :—

Dataset 1: 30, 40, 50, 60, 70

$n = 5$

$\bar{x} = \frac{250}{n} = \frac{250}{5} = 50$

$s^2 = \frac{(30-50)^2 + (40-50)^2 + (50-50)^2 + (60-50)^2 + (70-50)^2}{5-1}$

$= \frac{1000}{4}$

= 250 Ans

or standard deviation : $s = \sqrt{s^2}$

$= \sqrt{250}$

= 15.82 Ans

Dataset 2 : 0, 25, 50, 75, 101

$\bar{x} = \frac{250}{n}$

= 50

$s^2 = \frac{(0-50)^2 + (25-50)^2 + (50-50)^2 + (75-5 \quad (100-50)^2}{5-1}$

$= \frac{6250}{4}$

= 1562.5

Standard deviation :—

$s = \sqrt{1562.5}$

= 39.53

Ans

7.75 7.5

Date :
Page :

Practice Example:—

Dataset:— 8, 9, 10, 10, (10), 11, 11, 11, 12, 13

mean ($\bar{x}$) : $\frac{105}{10}$

$= 10.5$ (Ans)

n = 10 (Even number)

median = $\frac{(\frac{n}{2})^{th} \text{ value} + (\frac{n}{2}+1)^{th} \text{ value}}{2}$

$= \frac{10+11}{2}$

$= 10.5$ (Ans)

mode : (10, 11) Bi-modal dataset (Ans)

Range: (max − min)

$= (13 - 8)$

$= 5$ (Ans)

Hossain Mim
2021-1-60-071

variance

$\sigma^2 = \frac{(8-10.5)^2 + (9-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (12-10.5)^2 + (13-10.5)^2}{10}$

≈ 1.85

$\sigma = \sqrt{1.85}$

$= 1.36$ (Ans)

MID of 2nd CT

sample variance

$s^2 = \frac{(8-10.5)^2 + (9-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (12-10.5)^2 + (13-10.5)^2}{(10-1)}$

$= 2.055$

$s = \sqrt{2.055}$

$= 1.43$ (Ans)

Mean Absolute Deviation:— $MAD: \frac{|8-10.5| + |9-10.5| + |10-10.5| + |10-10.5| + |10-10.5| + |11-10.5| + |11-10.5| + |11-10.5| + |12-10.5| + |13-10.5|}{10}$

$= \frac{5+3+3}{10} = \frac{11}{10} = 1.1$ (Ans)

median Absolute Deviation:—

MAD: median $(|8-10.5|, |9-10.5|, |10-10.5|, |10-10.5|, |10-10.5|, |11-10.5|, |11-10.5|, |11-10.5|, |12-10.5|, |13-10.5|)$

n = 10

$(\frac{n}{2})^{th}$ value $+ (\frac{n}{2}+1)^{th}$ value

$\frac{0.5+0.5}{2}$

or (Ans)

$= \text{median}\{2.5, 1.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 2.5\}$

$= \text{median}(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5, 2.5)$

Data: 0, 5.6, 8.7, 14.1, 14.1, 15, 17.2, 19.2, 19.3, 24.1, 24.7

max: 24.7
min: 0

IQR: $Q_3 - Q_1 \Rightarrow (19.225 - 7.925) \Rightarrow 11.3$

Date :
Page :

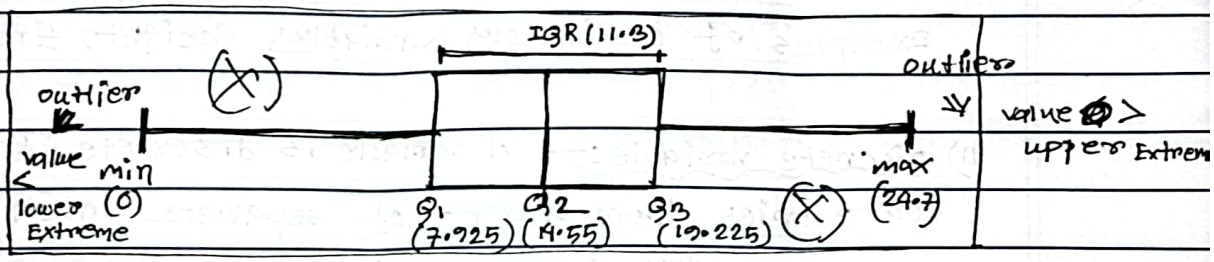
$Q_1 = 25\% \times 11$
$= 2.75$ th value
$= 5.6 + (8.7 - 5.6) \times 0.75$
$= 7.925$

$Q_3 = 75\% \times 11$
$= 8.25$ th value
$= 19.2 + (19.3 - 19.2) \times 0.25$
$= 19.225$

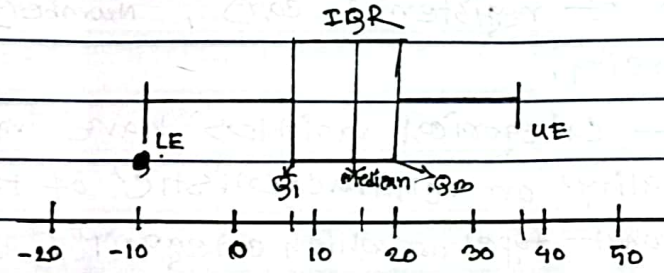
$Q_2 = 50\% \times 11$
$= 5.5$ th value
$= 14.1 + (15 - 14.1) \times 0.5$
$= 14.55$

Upper Extreme: $Q_3 + IQR \times 1.5$
$= 19.225 + 11.3 \times 1.5$
$= 36.175$

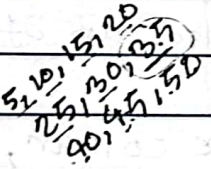
Lower Extreme: $Q_1 - IQR \times 1.5$
$= 7.925 - 11.3 \times 1.5$
$= -9.025$

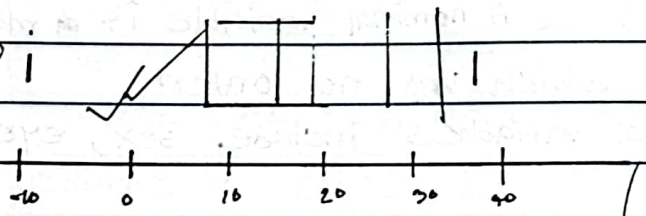


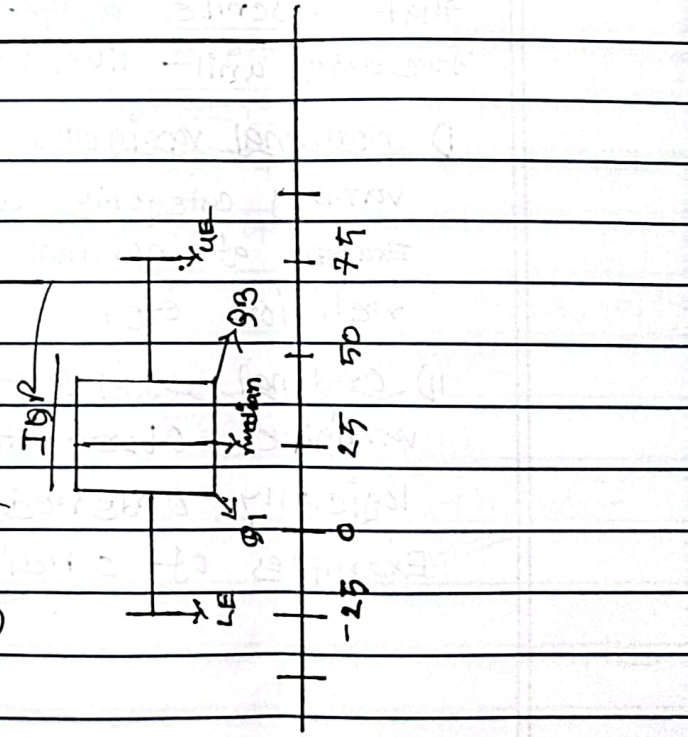
Box Plot



Hossain Mim
2021-1-60-071







$Q_1 = 25\% \times 10$
$= 2.5$ th value
$= 10 + (15 - 10) \times 0.5$
$= 12.5$

$Q_2 = 50\% \times 10$
$= 5$ th value
$= 25$

$Q_3 = 75\% \times 10$
$= 7.5$ th value
$= 35 + (40 - 35) \times 0.5$

$UE = (Q_3 + 1.5 \times IQR)$
$= 75$

$LE = (Q_1 - 1.5 \times IQR)$
$= -25$

$IQR = (37.5 - 12.5)$
$= 25$

Date :
Page :

variable: A variable is any characteristic, number or quantity that can be measured. A variable may also be called a data item. Age, sex, eye colour, country of birth etc are examples of variable.

Important

Numerical variables:— Numerical variable have values that describe a measurable quantity as a number, like 'how many' or 'How much'.

1) **Continuous variable:—** A continuous variable is a numeric variable. Observations can take any value between a ~~cert~~ certain set of real numbers.

Examples of continuous variables height, time, age etc.

11) **Discrete variable:—** A variable is discrete if it possible categories from a set of separate numbers

Examples of discrete variables, numbers of children in a family, Number of registered cars, Numbers of boy in the ~~schools~~ university.

Categorical variables:— Categorical variables have values that describe a 'quality' or 'characteristic' of the the data unit like 'what type' or which category'.

1) **Nominal variable:—** A nominal variable is made up of various categories which has no order

Example of nominal variables include sex, eye colour, religion etc.

11) **Ordinal variable:—** An ordinal variable is a categorical variable. Observation can take a value that can be logically ordered or rank.

Examples of ordinal variables include

Assignmim
2021-1-60-071

Date :
Page :

There are mainly seven steps involved into it :—

i) Decide on the objectives :— These objectives may usually require significant data collection and analysis.

ii) what to measure and how to measure :— measurement generally refers to the assigning of the numbers to indicate different values of variables

iii) Data collection :— once you know what type of data you need for your statistical study then you can determine whether your data can be gathered.

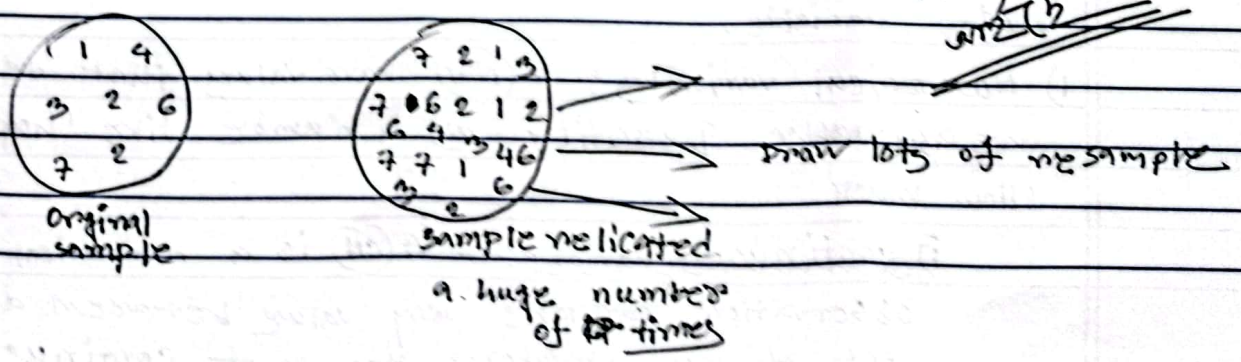
iv) Data cleaning :— This is another crucial step in data analysis pipeline is to improved data quality for your existing data.

v) summarising :— Exploratory data analysis helps to understand the data better

vi) Data modeling :—

vii) optimize and Repeat :— The data analysis is a repeatable process and sometime leads to continuous improvements, both to the business.

The Bootstrap :— one easy and effective way to estimate the sampling distribution of a statistic of a model parameters, is to draw additional samples, with replacement from the sample itself and recalculate the statistic or model each resample. This procedure is called bootstrap and it dose not necessarily involve any assumption about the data or sample statistic ~~being~~ being normally distributed.



Date :
Page :

The algorithm for a bootstrap resampling of the mean is as follows for a sample of size n :—

1) Draw a sample value, record, replace it
2) Repeat n times
3) Record the mean of the n resampled values
4) Repeat steps (1-3) R times
5) use the R results to :—
   i) calculate their standar deviation
   ii) produce a histogram or boxplot
   iii) Find a confidence interval.

Confidence Interval :— A (CI) is statistics is a rage of values thats used to estimate an unknown population parameter, such as the mean or proportion. It provides a range of plausible values for the parameter and is calculate from sample data.

For example 95% (CI) for mean weight of a population might be from 60kg to 70kg, indicating that we're 95% confident that the true population mean weight falls with in this rage.

The width of the interval is influenced by the sample size and the desired level of confidence.

variable :— A variable is any characteristic, number, quantity that can be mesured. A variable also be called a data item Age, sex, eye colour, country of birth etc are Example of variable.

1) Numerical variables :— (Nv) have values that describe a measurable quantity as a numbe like 'how many' or 'How much'

   i) continuous variable :— A (CN) is a numerical variable. observations can take any value between a certain set of real number. Example of continuous variable include weight, time, Age.

A confidence interval is sat[illegible] is [illegible] that used to estimate an unknown population parameter, such as the mean or proportion. Date

It provides a range of plausible values for the parameters and is calculated from sample data.

95% ⇒ It means that should we repeat an experiment or survey over and over again, 95% of the time our results will match the result s, we got from a ~~prop~~ population.

self-selection bais in ~~stat~~ statistic occurs when individuals or entities self-selection into a sample being studied, leading to a non-random or biased representation of the population.

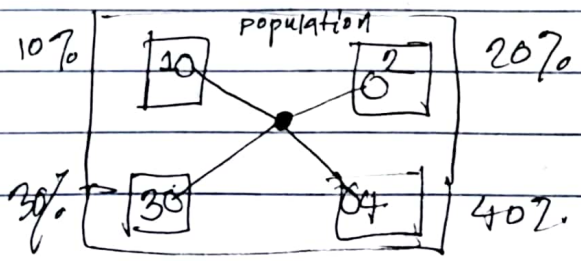
An Example self-selection bias in political polls.

consider a situation where a poll about a certain candidate online platform. people who actively seek out and participate in such poll might have stronger opinions or be more engaged it we topics. compared to general population.

Important

stratified sampling is a sampling technique in statistic where the population is divided into distinct subgroups, called strata, based on certain characteristics that are relevant to the study.

The strata are homogenous whit in themselves but different from ~~each~~ each other in some key aspect



Sorted data:— 64, 65, 68, 69, 70, 71, (72), 72, 75, 75, 80, 81, 83, 85

25th percentile (Q1) = 25% × 14
= 3.5 th value
= 68 + (69 − 68) × 0.5
= 68.5

MID 01, 2025

50th percentile (median) = 50% × 14
= 7th value
= 72

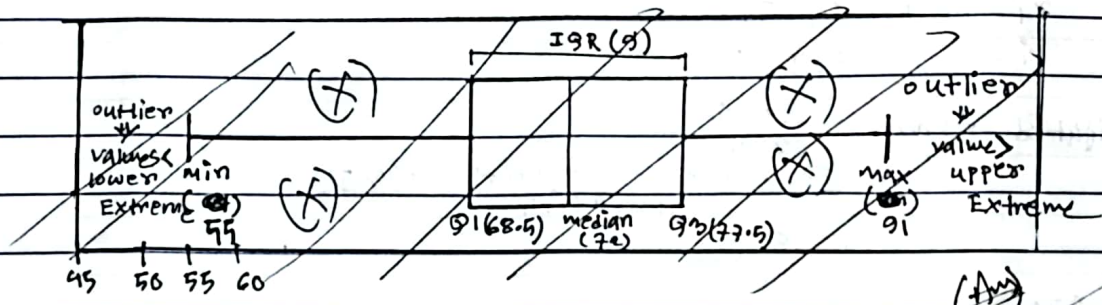
75th percentile (Q3) = 75% × 14
= 10.5 th value
= 75 + (80 − 75) × 0.5
= 77.5

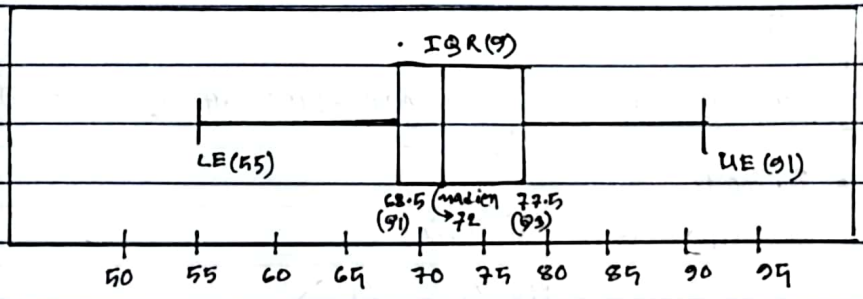
IQR = (Q3 − Q1)
= (77.5 − 68.5)
= 9

max = 85
min = 64

Upper Extreme = {Q3 + (IQR × 1.5)}
= {77.5 + (9 × 1.5)}
= 91

Lower Extreme: Q1 − IQR × 1.5
= {68.5 − (9 × 1.5)}
= 55





we can't see any outlier

Ans

H_0: There is no linked between gender and political party preferencemee

H_1: There is linked between gender and political party preference

Expected value for male republican $= \frac{240 \times 200}{420}$

$= 114.28$

Expected value for female republican $= \frac{240 \times 220}{420}$

$= 125.72$

Expected value for male democrat $= \frac{130 \times 200}{420}$

$= 61.91$

Expected value for female democrat $= \frac{130 \times 220}{420}$

$= 68.09$

Expected value for male Independant $= \frac{50 \times 200}{420}$

$= 23.81$

Expected value for female Indepent $= \frac{50 \times 220}{420}$

$= 26.19$

Now,

$x_c^2 = \frac{(O_i - E_i)^2}{E_i}$

* x_c^2 are all sum of values $= (1.79 + 1.63 + 1.074 + 0.97 + 1.61 + 1.46)$

$= 8.541$

we can take 4

degrees of freedom $(2-1) \times (3-1)$

$= (1 \times 2)$

$= 2$

we can see that alpha level of 0.05 and 2 degrees of freedom

- the critical statistic is 5.991
- the obtain statistic is 8.541

cr < ob so, H_0 rejected (Ans)

Standard Error:— The SE is a single metric that sum up the variability the sampling distribution for a statistic

$$SE = \frac{s}{\sqrt{n}} \rightarrow \text{(sample)}$$