Intermediate Stats

Agenda

1. Measure of central tendancy.
2. Measure of dispersion
3. Percentiles and Quantiles
4. 5 Number Summary (Box plots)

① Measure of Central Tendancy

- Mean
- Median
- Mode

$\{23, 24, 28, 29, 31, 32, 33\}$

⇓

This data can be represented with a central value.

:) Measure of CT is a single value that attempts to describe a set of data by identifying the central position.

1.1 Mean (Average)

:) We calculate average

Population (N) $x_i = \{1, 2, 3, 4, 5\}$ Sample (n)

⇓ ⇓

$$\text{Mean}(\mu) = \sum_{i=1}^{N} \frac{x_i}{N} \qquad \text{Mean}(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

{When to use mean?} ⇒
:) Use when data is normally distributed.
:) mean is sensitive to outliers.

1.2 Median { Used when data is skewed or contains outliers }

$x_i = \{1, 2, 3, 4, 5\}$

$\mu = 3$

range = 5 − 1 = 4

$x_j = \{1, 2, 3, 4, 5, 100\}$

$M = 19.16$

range = 100 − 1 = (99)

when range is large there is a possibility of outlier

Finding Median

1. Sort the data.
2. Find the central number
   - → If no. of elements are odd we take avg. of central number.
   - → If no. of elements are even we take central number.

Ex 1: $\{1, 2, 3, 4\}$

median = $\frac{2+3}{2} = 2.5$

Ex 2: $\{1, 2, 3, 4, 5, 6, 7\}$

median = 4

1.3 Mode ⇒ Most frequently occuring data.

$x_i = \{22, (24), 26, 27, 100, (24), 199, (24), 27, 32, (24), 33, 54, (24)\}$

⇒ Most repeated value in x_i = 24.

⇒ Which is the Mode in this Case.

{ When to use mode?

It is appropriate to use when data has two peaks (bimodal) }

② Measure of Dispersion

- → Variance (σ^2)
- → Standard Deviation (σ)

Consider samples,

$x = \{1, 1, 2, 2, 4\}$ $\qquad$ $y = \{0, 2, 2, 3, 3\}$

$$\mu_{x_i} = \frac{1+1+2+2+4}{5} = 2 \iff \mu_{y_i} = \frac{0+2+2+3+3}{5} = 2$$

Here, mean is same but we cannot say it follows the same distribution.

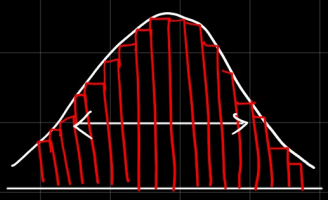
2.1 Variance (σ^2)

$$\text{population Variance } (\sigma^2) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

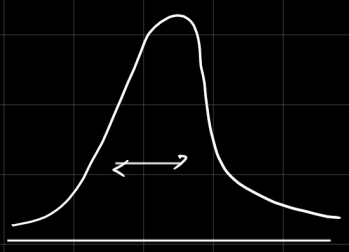
↳ tells us { spread of the distribution }

$$\text{Sample Variance } (s^2) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

→ Bessel's correction
Degree of freedom



) Range is higher
) Variance is higher



) Range is lower.
) Variance is lower.

2.2 Population Standard Deviation

Population

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}$$

Sample

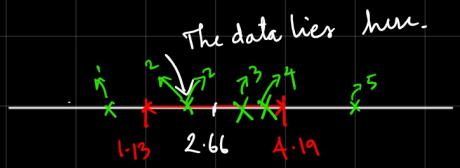
$$s = s^2 = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	2.66	-1.66	2.75
2	2.66	-0.66	0.43
2	2.66	-0.66	0.43
3	2.66	0.44	0.19
4	2.66	1.44	2.07
5	2.66	2.44	5.95
2.66			11.82

$$s^2 = \frac{11.82}{6-1} = 2.36$$

$$s = \sqrt{2.36} = 1.53$$

2.66
1.53
1.13



one standard deviation to right $= (2.66 + 1.53)$

one s.d to left $(2.66 - 1.53)$

(mean - variance)

③ Percentiles and Quartiles

Percentage = {1, 2, 3, 4, 5, 6, 7, 8}

Ex:- Percentage of even numbers = $\frac{\text{no. of even}}{\text{Total}} = \frac{4}{8} = \frac{1}{2} = 0.5$

Percentile :- A percentile is a value below which a certain percentage of observations lie.

Ex:- A persons GATE score is 90 Percentile

then it means, the person has scored better than 90 % of Students.

problem 1

Data :- {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

What is the rank of 10?

{Data must be sorted before solving}

Percentile of x = $\frac{\text{no. of values below } x}{\text{Sample}} * 100$

$= \frac{16}{20} \times 100$

$= 80$ percentile

which means 10 is greater than entire distribution.

Problem 2 →

What is the value that exists at the 25th percentile?

$$\text{Value} = \frac{\text{Percentile}}{100}(n+1)$$

$$= \frac{25}{100}(21) = \frac{21}{4} \approx 5.25 \Rightarrow [5]$$

$\frac{20}{4}$

we have to choose 5th index value in data

{When rank value is in decimal choose the corresponding index value and index +1 value}

Ex: $5.25 \Rightarrow \frac{5^{th}\text{ index} + 6^{th}\text{ index}}{2} = \frac{5+5}{2} = 5$

④ 5 Numbers Summary

*) Minimum
*) First Quartile (Q_1) 25%
*) Median
*) Third Quartile (Q_3) 75%
*) Maximum

→ helpful in understanding and removal of outliers.

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27} → Outliers.

We can use 5# summary to prove / conclude that '27' is an outlier.

[Lower fence ⟷ higher fence]

Lower fence ⇒ $Q_1 - 1.5(IQR)$

where $IQR = (Q_3 - Q_1)$

↓ 75th ↓ 25th percentile

IQR ⇒ Inter Quartile Range.

higher fence ⇒ $Q_3 + 1.5(IQR)$

$Q_1 = (25^{th} \text{ percentile}) = \frac{25}{100} \times (n+1)$ {n → length of data}

$= \frac{5}{\cancel{25}/\cancel{100}} \times (20) = (5^{th} \text{ index}) \Rightarrow \boxed{3}$

$Q_3 = (75^{th}\ \text{Percentile}) = \frac{75}{100} \times (20) = (15^{th}\ \text{index}) \Rightarrow \boxed{7}$

$$\boxed{IQR = Q_3 - Q_1 = 7 - 3 = 4}$$

Lower fence $= 3 - 1.5(4) = -3$

Higher fence $= 7 + 1.5(4) = 13$

It tells us that any value below -3 and above 13 are outliers.

{ i.e, in our case '27' is proved as outlier }

Data after removing outliers,

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9}

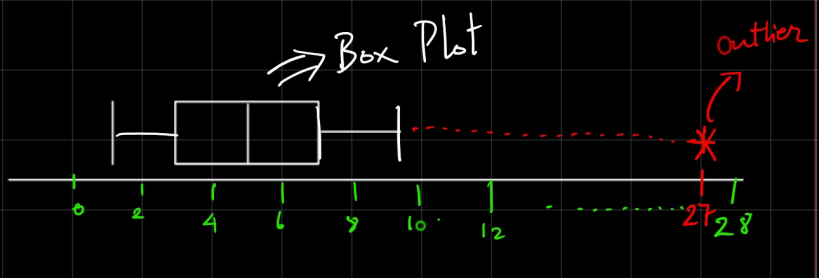
*) Minimum = 1

*) Q_1 = 3

*) Median = 5

*) Q_3 = 7

*) Maximum = 9



01/04/2023