15/7/24

Statistics for DS :
Complete Notes & explanation.

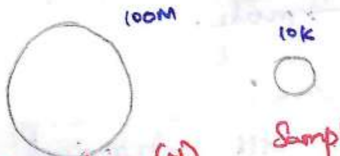
→ Statistics : Science of collecting, organizing & analyzing data.
purpose : better decision making.

data ↓ Facts (or) pieces of info that can be measured.

↳ Descriptive Stats : It consists of organizing & summarizing data.
(eg) Average, mean, mode, median.

↳ Inferential Stats : Technique where we use the data that we have measured to form conclusions.

(eg) Taking exit poll for election



Infrantial Stats Population (N) Sample (n) Descriptive Stats

↳ Sampling Techniques :

* Simple Random Sampling : just picking samples randomly.
* Stratified Sampling : Population (N) is split into non-overlapping groups. (strata).
  (eg) dividing the total population with male & female.
  (eg) dividing the N with Age-groups. (Samples).
* Systematic sampling : (N) → elements are selected at a regular interval after a random starting point.
  (eg) To survey 1000 employees for job satisfaction, we can select 100 employees. (samples)
* Convinient sampling : only interested people / with domain knowledge are selected for the survey.
  (eg) Only Data Science people are selected for DS survey. (Samples)

↳ Variables : A property that can take on any value.

* Quantitative : Age, weight, height, etc.
* Qualitative / Categorical : Gender, Blood groups etc.

IQ

0-10	10-50	50-100
Less	Med	Good

→ Discrete : No. of Bank Accounts, No. of people in family.

→ Continous : Height, weight etc.

↳ Variable Measurement scales :

* Nominal data : Categorical data (eg) Colors, Animals.
* Ordinal data : we focus on the order rather than the data. Categorical data with order.

(eg) Education level

Marks	Rank
100	1
90	3
97	2

} ordinal data.

* Interval data : Numeric data with intervals. (No true zero). (eg) Temperature
* Ratio scale : Numeric data with True Zero.
  (eg) weight - 0 kg.

↳ Measure of Central tendency: Refers to the measure used to determine the centre of the distribution of data.

* Mean: Average calculation. $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$

* Median: To leverage the problem of outliers, we go to median.

Steps: ① Sort the data in asc.
② choose the middle value as median.
③ In case of even no. of elements, we take the middle 2 elements & take average of the 2 elements.

* Mode: The value that occurs most frequently.

(eg) car 14 → mode.
Cycle 6
walk 5

Mode really works with categorical data (i.e.) if our dataset has cv, then we use mode to replace the missing data.

↳ Measure of dispersion: Dispersion refers to how well spread our data is.

* Standard deviation: $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Quadratic mean of the distance from the mean.

Indicates the avg. distance between each data point & mean. Each datapoint has some deviation from the mean. The average calculation of these deviations is called as std. deviation.

* Variance: Squared standard deviation.

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

→ Population variance: Measure of the dispersion of all data points in an entire population.

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

→ Sample variance: Measure of the dispersion of data points in a sample.

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

* Range: The diff. btwn the maximum & minimum value.

* IQR: Inter Quartile Range: Represents the middle 50% of the data. IQR = $Q_3 - Q_1$.

* Percentile: Is a value below which a certain percentage of observations lie.

(eg): Dataset: 2, 2, 3, 4, 5, 5, 6, 7, 8, 9, 9, 9, 10, 9

what is the percentile ranking of 10?.

$$x\% = \frac{\text{No. of values below } x}{x} \times 100 \quad (\text{No. of values below } 10).$$

$$= \frac{\overset{6}{\cancel{12}}}{\underset{7}{\cancel{14}}} \times 10\cancel{0} = \frac{600}{7} \simeq 85\%.$$

Thus, 85% of the values in the distribution are below 10.

→ Removing the outliers: Consider the following dataset:

[1, 2, 2, 3, 4, 4, 5, 6, 7, ~~8~~ 9, 9, 9, 10, 27]. → outlier.

* we can remove the outliers by using,

outliers > [lower fence ⟷ Higher fence] > Outliers

* IQR = $Q_3 - Q_1$ [Q_3 = 75% ; Q_1 = 25%].

$$Q_3 = \frac{\cancel{75}^{15}}{\cancel{100}\ \cancel{20}\ 4} \times \cancel{15}\,3 = 11 \rightarrow \text{Index.}$$

the Q_3 value is 9.

$$Q_1 = \frac{\cancel{25}^{5}}{\cancel{100}\ \cancel{20}\ 4} \times \cancel{15}^{3} = 3.6 \rightarrow \text{Index}$$

the Q_1 value is 3.

Thus, IQR = 9 − 3 = 6.

* lower fence : $Q_1 - 1.5[IQR]$

$= 3 - 1.5 \times 6$

$= 3 - 9 = -6.$

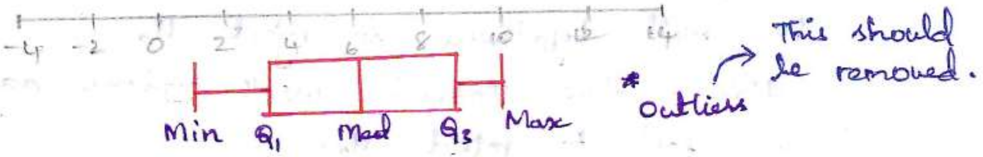
* Upper fence : $Q_3 + 1.5[IQR]$

$= 9 + 9$

$= 18.$

Thus, from this calculation, we can derive that and values below −6 and above 18 are considered as OUTLIERS.

* The Q_1, Q_3, IQR, Minimum Value, Max. value, Median are collectively called as Five Number summary. Using this, we can easily predict our outliers by drawing a box plot.



Inferential Statistics: Helps us make conclusions or inferences about a population based on a sample of data

↳ Hypothesis testing: A premise or claim that we want to test.

(eg) Investigation of a thing by collecting info from 200 people (sample).

* Null Hypothesis: Default (or) established. H_0 - currently accepted value for a parameter.

$H_0 = \mu = 1.2m$ (eg) Scientists who have already estimated the age of the earth is 1.2m years

* Alternative Hypothesis: H_a - Also called research hypothesis. Involves the claim to be tested.

$H_a = \mu \neq 1.2m$ (eg) New generation scientists are challenging the old scientists that the age of earth is not 1.2m years but 2.2m years

* Parameter: A numerical characteristic of a population (e.g) population mean, pop. std dev.

* Statistic: A numerical characteristic of a sample (eg).
sample mean,
sample std. dev.

* **Level of confidence:** A range of values that is likely to contain the population parameter with a certain level of confidence (95%). (i.e) 95% we can be sure that the true value is true.

* **Type-1 error:** Rejecting Null Hypothesis H_0 when it is actually true. It can be thought of as a "false positive".

  (eg) If testing a new drug,

  (α) H_0 - has no effect on patient

  H_A - does have an effect.

  After testing, we found the drug has an effect. However if the test results are due to randomness & the drug actually don't have any effect, the error type here is Type-1.

* **Type-2 error:** Failing to reject H_0 when it is actually false. "False negative".

  (eg) Same example can be taken but after testing we conclude that the drug has no effect but in reality it does have an effect on the patients.

  (β)

$$\alpha \propto \beta$$

* **P-value:** It is the propability for the "Null Hypothesis" to be true.

  → Low p-value : ($P \leq 0.05$): Indicates strong evidence against the null hypothesis, so reject H_0.

  → High p-value: Indicates weak evidence against H_0, so we fail to reject H_0.

  → Significance level (α): A threshold chosen before the test, commonly set to 0.05. If the p-value is less than α, the H_0 is rejected.

→ **Statistical tests:**

* **Z-test:** Used for comparing means with large sample sizes and known variances.

* **T-test:** used for comparing means with small sample sizes or unknown variances.

* **ANOVA:** used for comparing means across three (or) more groups.

* **Chi-squared test:** used for testing relationships btwn categorical variables.

(22) Using the choose columns & Remove other columns options will allow us to explicitly select the columns that we want to keep.

(23) Model view - Used to manage data model. Contains tables.
Report view - Manage roles, including creating them.
Data view - To transform & analyze data.

(24) When creating a quick measure, in PBI desktop, we apply calculations to fields.

(25) To create a measure, we can use : Data & report view.

(26) To reduce cardinality, reduce the no. of distinct values.

(27) Visuals that support conditional formatting: Matrix, table, cards, Bar & Column charts, Pie & Donut charts.

Descriptive :
* Describes & summarizes.

Infrentials :
* Draws conclusion on the population data.

Mode, frequency ←
Mode, freq, Median ←

M, M, M, Std-dev ←
M, M, M, SD, ratios ←