Statistics for Data Science.

Statistics

ABHISHEK SINGH SUNNY JAT

1) Descriptive Stats
   - Measure of central Tendency
   - Measure of Dispersion.

   Summarizing the data.
   Histograph, Pdf, Cdf, Probability Permutation, Mean, Median, Mode, Variance, Standard distribution

   1) Gaussian Distribution.
   2) Log normal Distribution.
   3) Binomial Distribution
   4) Bernoulli's Distribution
   5) Pareto Distribution (Power law)
   6) Standard Normal Distribution → Rymon
   7) Transformation and standardisation.
   8) qq plot.

2) Inferential Stats
   - Z test
   - t test
   - Anova test
   - CHISQUARE
   - Hypothesis testing.
   - Confidence Intervals

   Z table, t table, chi square table.

The two most important functions of descriptive stats are.

1) communicate information
2) Supporting reasoning about data.

Qus. What is Statistics ?

Ans Statistics is the Science of collecting, organising and analyzing data.
{ Better Decision making }

~~Qus~~

Data => Facts or pieces of information that can be measured.

Example => The IQ of a class
{98, 97, 60, 55, 75, 65} => Data.

Data is information such as facts and numbers used to analyze something or make decision

Ages of Students of a class
{30, 25, 24, 23, 27, 28} → DATA

Types of Statistics.

1) Descriptive Stats => It consists of organising and summarizing data.

2) Inferential Stats => It is a technique where in we used the data that we have measured to form conclusions.

Example:

1) Classroom of Maths students (20) students

data: Marks of 1st Sem

(84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97 ...)

what kind of question may come in descriptive Stats and what kind of questions may come in inferential stats.

Ex Descriptive Stats.

⇒ What is the average marks of the students in the class?

Ex⇒ Inferential Stats.

⇒ Are the marks of the students of this classroom similar to the other Maths classroom in the college?

Population And Samples.

Population ⇒ The entire group of subject about which we want information.

Parameter ⇒ the quantity about the population we are interested in

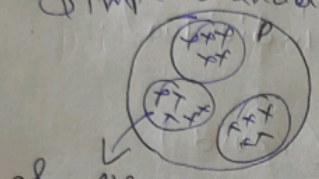
Sample ⇒ the part of the population from which we collect information

Statistic ⇒ the quantity we are interested in as measured in the sample.

Population (N) → Sample (n)

Sampling Techniques ⇒

1) Simple Random Sampling ⇒ when performing Simple random sampling every member of the population (N) has an equal chance of being selected for your Sample (n).



2) Stratified Sampling ⇒ where the population (N) is split into non-overlapping groups (Strata)

Ex ⇒ Gender [Male, Female] → There is no chance of overlapping.

Age Based.

(0-10) (10-20) (20-40) (40-100)

Ex Profession ⇒ .net PHP Python

But we can apply on Doctor, Egineer, Government job.

3) Systematic Sampling => Here from the population (N) we just pick up every nth individual from this population.

(N) ⟶ nth individual.

Eg => Mall ⟶ Survey (Covid)
↳ 8th person ⟶ Survey.

4) Convenience Sampling => I am doing a survey related to a specific topic in this particular example like data science
↳ Survey. ←
↳ DATA SCIENCE Only those people

Variables => A variable is a property that can take on any value.

Eg: Height = 172 cm
Weight = 65 kg

Two kinds of Variable; Ex => Age, Height, Weight

1) Quantitative Variables → Measured Numerically & Add, Subtract, multiple divide
2) Qualitative / categorical Variable
↳ Eg: Gender [M, F] (Based to some characteristics we can drive some categorical variables]

Eg: Blood Group

Quantitative Variables

Discrete Variable

Eg => Whole Number.

① No. of Bank Accounts.
You can say I have 1, 2 or 3 bank account.

② Total no. of children in a family.
=> 1, 2, 3, 5, 8 (Bargraph)

Continuous Variable

Eg => Height - 172.2, 162
Weight = 100kg, 78.2kg
Account of Rainfall = 1.25 inch, 1 inch

(Histogram)

* Variable Measurement Scales.

4 types of Measured Variables

1) Nominal {categorical data} -> Color, Gender.
2) Ordinal.
3) Interval.
4) Ratio

Ordinal => Order of the data matter, but value does not.

Eg

Students (Marks)	Rank
100	1
96	2
43	5
53	4
89	3

-> Ordinal Data.

Interval → Order matters, Value also matters, natural zero is not present.

Eg: Temperature (F)

70-80°F, 80-90°F, 90°-100°F

+

You have some range of value between them and the order also basically matters.

Ratio →

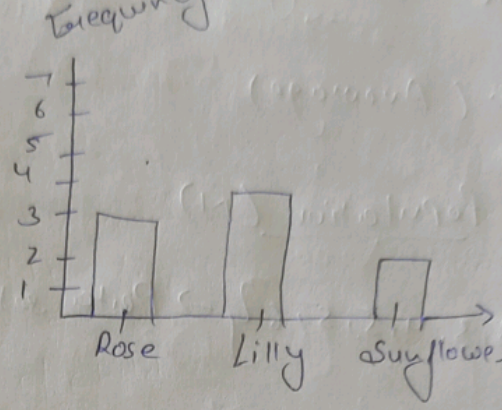
Frequency Distribution

Sample dataset ⇒ Rose, lilly Sunflower.
Rose, lelly, Sunflower,
Rose, lilly, lilly.

Flower	Frequency	Cummulate Fr
Rose.	3.	3
lilly.	4.	7
Sunflower.	2.	9
	Total 9	

32 → This is frequency of above dataset.

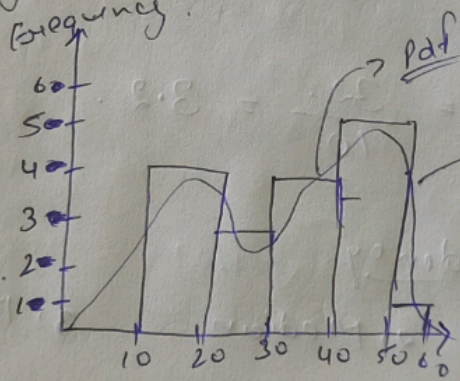
* Bar Graph (Distributive)



* Histogram (Continuous)

Age = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 42, 43, 50, 51 }.

↳ Bins (Grouping)
Default = 10



Probability Density function.

Your Value is Continuous.

↑
Pdf => Smoothning of histogram.

* Arithmetic Mean for Population & Sample

Mean (Average)

Population (N) | Sample (n)

$x = \{1,1,2,2,3,3,4,5,5,6\}$ → Data set

meu

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$\mu = \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$\mu = \frac{32}{10} = 3.2$$

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

$= 3.2$

* Central Tendency.

1) Mean 2) Median 3) Mode.

=> It refers to the measure used to determine the center of the distribution of data.

Outliers

Median: $\{1,1,2,2,3,3,4,5,5,6,100\}$

Mean $= \frac{32+100}{11} = \frac{132}{11} = 12$

$\mu = 3.2$

$+100$

$\mu = 12$

vivo V20 Pro 5G

Median = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100}

⇒ First Sort the number

⇒ If the total count of the number is odd number
= median = 3

Distribution:

⇒ If the total count of the number is even.

Avg = $\frac{3+4}{2} = 3.5$

{ Median works well with outliers }.

Outliers

Mode ⇒ {1, 2, 2, 3, 4, 5, 5, 6, 6, 6, 7, 8, 100, 200}

3

Mode = { most frequent Element }.

Mode = 6 ⇒ Measure of central Tending.

Types of flowers.
Rose.
lilly.
Sunflower.

Petel length
Petel width
10% missing data

{ Missing Value → most frequent number } Well work on Categorical Value.

Measure of Dispersion

1) Variable 2) Standard Deviation (Dispersion) = Spread

1) **Variance** ⇒ Variance is the expectation of the squared deviation of a random variable from the population mean or sample mean. **Population Variance** **Sample Variance**

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \frac{10.84}{6} = 1.81 \qquad s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Data Set

x	μ	$x - \mu$	$(x-\mu)^2$
1	2.83	−1.83	3.34
2	2.83	−0.83	0.68
2	2.83	−0.83	0.68
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

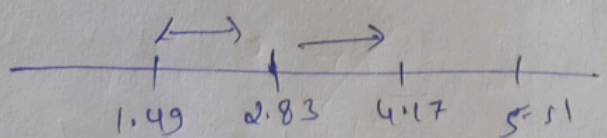
$$\sigma = \sqrt{\text{Variable}} = \sqrt{1.81} = 1.345$$

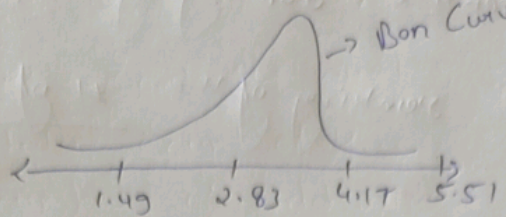
Our Side Standard division ⇒ Elements are basically present between two points.

2.83 + 1.34 = 4.17

4.17 − 1.34 = 5.51

4.17 − 1.34 = 5.51





* Percentiles And Quantiles {Find outliers}.

Percentage : 1, 2, 3, 4, 5

% of the number that are odd ?

% = No. of numbers that are odd / Total Numbers.

$= \frac{3}{5} = 0.6 = 60\%$.

Percentile ⇒ A percentile is a value below which a certain percentage of observation lie.

Example ⇒ Dataset ⇒ 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12.

⇒ What is the percentile ranking of 10 ?

Ans Percentile Rank of 10 = # of values below 10 / N. × 100

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

$= \frac{1600}{20} = 80\%$.

Meaning that ⇒ 80% of entire distribution is less 10.

=> What is Percentile ranking of 11?

Ans

Percentile ranking of 11 = $\frac{\#\text{ of values below 11}}{n} \times 100$

$= \frac{17}{20} \times 100$

$= \frac{170}{2} = 85\%$

Meaning = 85% of entire distribution is less than 11.

(Q) What value exists at percentile ranking of 25%?

Ans

Formula = Value = $\frac{\text{Percentage}}{100} \times (n+1)$

Value = $\frac{25}{100} \times 21$

$= \frac{21}{4} = 5.25$ → Index Position

= 5 is the value of 25%.

What value exists at percentile ranking of 75%?

value = $\frac{75}{100} \times 21$

$= \frac{1575}{100}$

$= 15.75$ is the value of 75%.