Statistics for Data Science with Baba Ammar

> **Libraries Pands, NumPy, SciPy**

---------- Statistics----------

Statistics is collection of methods for collecting , displaying, analyzing and drawing conclusions from data.

1. Statistics is everywhere
2. Will it rain? 55%-&0% chances (weather forcast)
3. Rate of USD prediction
4. Housing material prices increased
5. Un- employment rate fallen
6. Who gets paid how much ? "Superman ki salary.Baba ge example, har company ko superman chaye hota hy jo har kam kary, companies nechorti hain.
7. Average salary of Data Analst or Data Scientist?
8. Any comparsion in research
9. Yaar thesis ke stat laga do
10. ANOVA etc...

Difference between Data Analylst and Data Scientist

> A data analyst makes sense out of existing data, whereas a data scientist works on new ways of capturing and analyzing data to be used by the analysts.

You must talk in the language of statistics :-

Point out the stat used in following

> a. Average income in Pakistan = average

> b. Highest score in cricket match = maximum

> Fastest bowler = maximum

> Sab se kam run khaye = minimum

> 40 % teachers in Pakistan are female = percentage

> Kal barish kitni hoge, hoge b k nai hoge = variance

> Dollars kabi opar jata hy kabi neechy = prediction

> Hostels me larky ziada kharcha karty ha ya larkian = T-Test

> Faisalabad : Lahore : Karachi : Sargodha : ISlamabadad jugton (Jokes) ke ranking = ANOVA

Jon Hopkins Data

COVID CASES IN PAKISTAN PUBLICATION RESULTS isme statistics use ho rahe hy

Publication result me mean, mode, etc or different statistics use hoti hy

---- 50 Important Statistical Terms -------

List of 50 statistics terms

Here is a list of the most common statistical terms and their definitions:

> 1. Alternative hypothesis

An alternative hypothesis is also known as the null hypothesis. The null hypothesis is the opposite claim of the thesis. If the data you collect demonstrates that your original hypothesis was correct, then you can reject the alternative hypothesis.

Related: Alternative Hypothesis: Definition and When To Use It

> 2. Analysis of covariance

An analysis of covariance refers to a technique that evaluates if a dependent variable produces equal results across numerous independent variable situations. It specifically looks at the differences between the mean values of dependent variables that relate to each other.

Related: Covariance vs. Variance: What They Are and How They Differ

> 3. Analysis of variance

An analysis of variance is a method of evaluating the differences between variables in a study. It also includes a two-way study, which is a way to split variability into two parts, including systematic and random factors.

> 4. Average

The average refers to the mean of data. You can calculate the average by adding up the total of the data and dividing it by the number of data points.

Related: How To Calculate Average

> 5. Bell curve

The bell curve, which is also referred to as the normal distribution, displays the mean, median and mode of the data you collect. It usually follows the shape of a bell with a slope on each side.

> 6. Beta level

A beta level refers to the probability of accepting the null or alternative hypothesis. The beta level also means that the hypothesis is incorrect.

Related: What Is a Positive Correlation in Finance?

> 7. Binomial test

You can use a binomial test when you are studying a hypothesis with two potential outcomes and you believe one has a higher probability of being true. Its theory is based on probability calculations and the study of small samples.

8. Breakdown point

The breakdown point is a point in which an estimator is no longer useful. A lower breakdown point means the information may not be useful, whereas a higher number means there is less chance of resistance.

9. Causation

Causation is a direct relationship between two variables. Two variables have a direct relationship if a change in one value causes a change in the other variable.

10. Coefficient

A coefficient measures a variable by using a multiplier. When conducting research and computing equations, the coefficient is often a numerical value that multiplies by a variable, giving you a coefficient of the variable. If a variable does not have a number, the coefficient is always one.

11. Confidence intervals

A confidence interval measures the level of uncertainty of a collection of data. It is also the level of probability that data will fall into a set of pre-assigned values.

12. Correlation coefficient

The correlation coefficient describes the level of correlation or dependence between two variables. This value is a number between negative one and positive one and can suggest when two variables may have an identifiable relationship.

13. Cronbach's alpha coefficient

Cronbach's alpha coefficient is a measurement of internal consistency. It shows the nature of the relationship between multiple variables in a subset of data. Additionally, Cronbach's alpha coefficient remains consistent in the number of items it measures and can increase when the average correlation between items increases.

14. Dependent variable

A dependent variable is a value that depends on another variable to exhibit change. When computing in statistical analysis, you can use dependent variables to make conclusions about causes of events, changes and other translations in statistical research.

15. Descriptive statistic

Descriptive statistics are the results that describe the data of your study. This may include the mean or median of the data as well as any other information that describes a trend among the population.

16. Effect size

The effect size is a way to quantify significant differences between two populations or sets of data. Effect size considers the size of the population rather than focusing solely on the data you collect.

17. F-test

An F-test is any test that uses F-distribution. F-distribution refers to the process of comparing different models of statistics to determine which model works best for the specific study.

18. Factor analysis

Factor analysis refers to a process of listing data as a smaller function to better organize and interpret results. A factor is also a set of variables with similar responses that you can apply to a larger collection of data.

19. Frequency distribution

Frequency distribution is the frequency in which a variable occurs. It provides you with data on how often something repeats.

20. Friedman's two-way analysis of variance

Friedman's two-way analysis of variance is a statistical test that looks for differences among numerous studies. This nonparametric test focuses on the differences in a research study when the measurement of the dependent variable is not as necessary.

21. Hypothesis tests

A hypothesis test is a method of testing results. Before conducting research, the researcher creates a hypothesis or a theory on what they believe the results will prove. A study then tests that theory.

22. Independent t-test

An independent t-test is a comparison of two independent variables. Its main goal is to determine if the data you collect is enough to interpret that two variables are significantly different from each other.

23. Independent variable

An independent variable is a value that does not depend on another factor to exhibit changes. Independent variables are necessary for conducting research that focuses on causal relationships. Additionally, independent variables are necessary in regression testing, where analysts can measure correlations between one or more independent variables and dependent variables.

24. Inferential statistics

Inferential statistics is a test that you use to compare a certain set of data within a population in a variety of ways. Inferential statistics include parametric and nonparametric tests. When conducting an inferential statistical test, you take data from a small population and make inferences as to whether it will provide similar results in a larger population.

25. Marginal likelihood

The marginal likelihood is a variable's likeliness to marginalize. It involves a method of assigning a likeliness value to each variable and multiplying the probability of it occurring.

26. Measures of variability

Measures of variability, sometimes referred to as measures of dispersion, refer to an explanation of significant variances among scores. It may include the interquartile range, range, standard deviation and variance of two or more variables with results that are far apart from one another.

> 27. Median

The median refers to the middle point of the data. Typically, if you have a data set with an odd number of items, the median appears directly in the middle of the numbers. When computing the median of a set of data with an even number of items, you can calculate the simple mean between the two middle-most values to achieve the median.

> 28. Median test

A median test is a nonparametric test that tests two independent groups that have the same median. It follows the null hypothesis that each of the two groups maintain the same median.

> 29. Mode

Mode refers to the frequency in which a number repeats in a collection of data. If two or more numbers are repeating, the mode then becomes the value that repeats the most.

> 30. Multiple correlations

Multiple correlations are an estimate of how well you can predict a variable using a linear function of other variables. It uses predictable variables to come to a conclusion.

> 31. Multivariate analysis of covariance

A multivariate analysis of covariance is a technique that assesses statistical differences between multiple dependent variables. The analysis controls for a third variable, the covariate, and you can use additional variables depending on the sample size.

> 32. Normal distribution

Normal distribution is a method of displaying random variables in a bell-shaped graph. Most data naturally forms a bell curve, which is a normal distribution.

> 33. Parameter

A parameter is a quantitative measurement that you use to measure the population. It is the unknown value of a population that you conduct research on to learn more.

> 34. Pearson correlation coefficient

The Pearson correlation coefficient measures the strength of linear correlation between two variables. It differs from coefficient correlation in that the coefficient correlation only measures a single correlation.

> 35. Population

Population refers to the group you are studying. This might include a certain demographic or a sample of the group, which is a subset of the population.

> 36. Post hoc test

A post hoc test analyzes the results of a study. It identifies certain differences between a minimum of three sample groups.

37. Probability density

Probability density is a statistical measurement that measures the likely outcome of a calculation. You can display this on a graph with the probability lying outside of the normal curve.

38. Quartile and quintile

Quartile refers to one of a total of four groups of data. Quintile refers to one of a total of five groups of data.

39. Random variable

A random variable is a variable in which the value is unknown. It can be discrete or continuous with specific values to measure. Conversely, it can also be a value of a range.

40. Range

The range is the difference between the lowest and highest values in a collection of data. It provides a guide for where you can assign numbers on a bell curve.

41. Regression analysis

Regression analysis refers to the statistical process of estimating a relationship between two different random variables. You can use regression analysis to predict the correlation between a dependent variable and a series of independent variables. Regression analysis can also be linear regression or multiple regression analysis, in which the number of variables to evaluate can change.

42. Standard deviation

Standard deviation refers to the distance a result is from the average. It informs you how far a single or group result deviates from the average.

43. Standard error of the mean

A standard error of mean refers to the standard error level in a study. You can find the standard error of the mean if you divide the standard deviation by the square root of the sample size.

44. Statistical inference

Statistical inference occurs when you use data to come to an inference or conclusion. Statistical inference can include regression, confidence intervals or hypothesis tests.

45. Statistical power

Statistical power is the probability of finding an effect in a test. The statistical power analysis estimates the minimum size of the population you need to conduct an effective study.

46. Student t-test

A student t-test is a hypothetical hypothesis that tests a small sample when you do not know the standard deviation. This can include correlated means, correlation, independent proportions or independent means.

> 47. T-distribution

T-distribution refers to a method of distributing probability measurements that relate to normal distribution patterns. It gives you a baseline for the standard distance between a sample and a real population.

> 48. T-score

A T-score refers to the number of standard deviations a sample is away from average. You can use it in T and regression tests.

> 49. Z-score

A Z-score, also known as a standard score, is a measurement of the distance between the mean and data point of a variable. You can measure it in standard deviation units.

> 50. Z-test

A Z-test is a test that determines if two populations' means are the same. To use a Z-test, you need to know the differences in variances and have a large sample size

---------TYPES OF DATA---------

> ## ----------First Type----------
>
> 1. a :- CROSS SECTIONAL DATA

13 Janauary ko kitny log video dek rahy hain it is cross sectional data.

> 1.b :- . Data with TIME SERIES

1 January se aj tak kitny bachy lecture dekthy hain this is data with TIME FRAME DATA COLLECTED AT ONE POINT

> ## ----------Data Type - 2----------

.

> 2. a :- UNIVARIATE

Data contains a single variable to measure entity
e.g Plant Height kitni barhati hy siraf pani se

Weight kitna barhata hy Naan khany se lets say one variable use kia

> 2.b :- MULTI VARIATE

Data contains > 2 variables to measure something

e.g yaha par 3 factors hain

1.Plant Height
2.Fertilizer amount

3.Irrigation time

agar ap yaha kahain k main sirf NAAN k sath kabab or coke pee raha ho to doube variable hogaya

----------Variable Types-1---------

--------- Categorical Variable---(Nominal)

Bionomial :-

Nominal me ranking karna ya hona lazmi nai

when we have only two options
True or False , Male- Female, Yes-No ye no quantitative data hota hy or relationship is given,/ Kashti wala data set me sex , male , female binomila tha for example Nashta kia ? han kia to apka relation ban gaya nashty k sath

Multinomial :-

For example travel choices. On how many devices you are using to watch Baba G Lecture, or kashti wala dataset me, who, jisme male, female, child, multinomial tha

----------Variable Type - 2 ---------

---------- Categorical (Ordinal)----------

Ordinal Variable :-

Categories can be compared
No Fixed unit of measurement for statistics
Data Ranked or ordered:
Mery pas kitnny Cell phones hy ?

Mery pas kitny phone or baki logo k pas to number of phone k hisab se rank hojaiga, ranking ban jati hy,

--------Variable Data Type-3---------- (Ratio Data)----------

Data have a natural zero.
Aj ziada mal bika hy kal ke nisbat

Measurement in units and ratios are Continuos Variable

Continuous and catergorical are two different variable

----------Variable Type-3----------(Interval Variable/ DATA)-----

Ordered Characterized YE data ordered hota or characterized hota hy

es June me garmi ziada 2020 ke june ke nisbat.
Garmi agar ek parameter hy, kiu k garmi ek feeling hy, 25 degree or 50 double feel wise boht faraq hota hy.

> Ratios are meaning less (50 degree is not double hot feel of 25 degree) , Difference meaningful hy boht, but ratio etna meaningful nahe

Measure of Central Tendency :-

Mean, Median and Mode :-

> Population vs Samples:-

Population research has more power other hand sample are used to reduce cost of data collection.

Limited population ka sample lena asan or sasta hota hy.
Kuch Shehro ka data le kar ham baki sab pe apply karty hain.\

> Population is more accurate than sampling

> Sample is done when we have limited resources

Covid-19 ka data jo Pakistan me lia gaya wo sample set pe tha

Notions and Terms :-

N = Size of population

n = size of sample

Σ = Sum

μ = population mean

std(x) = standard deviation

σ2 = variance

σX = standard deviation

x = sample mean

s = sample standard deviation

s2 = sample variance

> Statistics :-

Statistics is a collection of methods for collecting, displaying, analyzing, and drawing conclusions from data.

A population is any specific collection of objects interest.

A sample is any subset or sub collection of the population, including the case that the sample consists of the whole population, in which case it is termed a census.

> Measurement :- Measurement is a number or attribute computed for each member of a population or of a sample. The measurements of sample elements are collectively called the sample data.

> A parameter is a number that summarizes some aspect of the population as a whole

> A statistic is a number computed from the sample data

> Descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data.

> Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population

> Qualitative data are measurements for which there is no natural numerical scale, but which consist of attributes , labels or other non-numerical characteristics (wo data jo categorical ho)

> Quantitative data are numerical measurements that arise from a natural numerical scale

> Mean is the sum divided by the number of observations (Average)

The mean is the average or the most common value in a collection of numbers.

> Mean and its properties :-

Mean is actually an average, its meaningful for intervals and ratio data.

Outliers change the means of a data, therefore median is useful.

The Median is the middle number in a sorted , ascending or descending , list of numbers.

> Population mean. μ = (Σ * X)/ N where :

```
Σ means “the sum of.”
X = all the individual items in the group.
N = the number of items in the group
```

> Sample mean :- $\bar{x}$ = (Σ xi) / n :

```
x̄ just stands for the “sample mean”
Σ is summation notation, which means “add up”
xi “all of the x-values”
n means “the number of items in the sample”
```

> Median is the middle number in a sorted, ascending or descending, list of numbers

Outliers change the mean of a data, therefore median is usefull for that data.

> Medain properties :-

Median is a middle number in a sorted, ascending or descending, list of numbers

- Mean is unuqie for each dataset
- Outliers dont have any effect on median
- Ratios , intervals and ordinals data has best use

> Mode is the value that occurs most frequently

- 18 years age is most common in a class.

Dispersion :-

How much data spread around mean is called dispersion. sides pe data kitna phel raha hy wo dispersion kehlata hy.

Dispersion is caused by between minimum and maximum of data

Range = minimum to maximum

Dispersion wo hy jo data phelta hy min to max.

For example samosa rate

Prices = (30,30,35,45,10,61,70,90,115)

Mean = 54

Median =45

Mode = 30

Std = 33.26

variance = 1106.5

std = square root of variance

min = 45

max = 115

std has same units as sample or population

Standard deviation (std)

> Role of standard deviation and mean

Bell curve me jo centre hoga wo mean hoga jaha par values ziada ja rahe ho, agar steep hoge curve to dispersion kam hoge or agar curve flat or near to flat or less steep hoge to dispersion ziada hoge.

agar standard deviation ziada hoge tu means dispersion ziada hy

Baba ge ke karahi gosht and biryan ke example yad rakhna.

Mean only gives us a small picture

Means are incomplete without dispersion (SD). Mean with SD is more useful than only mean by itself

Fundamentals of Visulization :-

> Types of visualization depends on the variable type :-

> Categorical variable :-

- Counts (plot type)
- Male vs Female
- 0 vs 1
- Yes vs No

> Continuous Variable

- Scatter plot
- Statistical proportions (mean and their comparison)

> Catergorical variable are :-

- Qualitative
- No numerical meaning
- Represented in texts
- for example character, factors in R

> Continuous variable are :-

- Quantitative
- Numerical
- Mostly represented in number (for example: Numerical variable in R)

Categorical :-

- True or False / Yes or No , it can me nominal or multi nominal

For example :- Do you like mangoes ? "Yes" or "No"

Continuous :-

- Amount
- Number
- Age
- Plant Height
- Number of bacterial colonies
- Cholorophyll content
- Fertilizer amount

Chart suggestion - A thought starter from the Extreme Presentation Method created by Dr Andrew Ambela