5 Oct-2022 Data Science Wednesday!

Pandas:

→ It offerss data structures and opesations for manipulating numerical tables and time series

→ Pandas is used for analyzing the data and its manipulation

* Series is a specific row / specific column

→ dl does prediction on basis of statistics

→ 0.025% of dl-applied

→ shape returns a tuple

→ iloc: integer location.

→ iloc can be used to fetch columns.

* data.iloc[:, [4,5,16]] → all rows with 4, 5, 16 no column

Filtering dataframe on a condition

```
mask = data['city'] = "Hyderabad"
data[mask] → dataframe / mask : boolean series

check = data[mask1 & mask2]
check

data['season'].value_counts()

data['winner'].value_counts().head().plot(kind="bar")
```

win run is on horizontal (in histogram)
frequency on Y-axis

13/Oct-2022 Data Science Mahi's bday ♡

→ plt.style.use("ggplot") built in option available — seaborn and matplotlib folders
color/shape/background

→ Matplotlib and seaborn are used for data visualization in the pictorial view/form

→ import matplotlib.pyplot as plt
main library sublibrary alias

→ plt.tight_layout( ) → to take graph in center / a little bit alignment from edges

* for arrow we write 'v' in marker

→ plt.xkcd( ) like graph is sketched

→ plt.bar( ) } bar and plot (line graph) will be mixed
plt.plot( ) }

→ arange (18) means indexes from 0 to 17

→ Counter( ) built in function of pandas
for different languages

→ most_common( ) function
[name: index[0]] , index[1] → count

→ wedgeprops ( ) { for boundary specification
shrink

→ autopct = '%0.1f%%' → 1 decimal places, float

→ explode is for the pie chunk which is out of whole body

→ shadow = True for shadow of c

13/Oct Data Science

→ startangle = 90, cutting pie chart at 90° angle

→ Stackplot helps in visualizing pattern of every user after every minute

for first value, it shows the area of first, then add the value with second one and then add it with third one.

But sum of all values must be same in whole list

→ plt.fill_between (ages, py_salaries, dev_salaries, where = (py_sal > dev_salaries, interpolate = True, color = "red", alpha = 0.25, label = "Above Avg")

alpha → Transparent/opacity

→ Histogram tells you count of occurrence

→ We take log to make values in a range of small values.

→ log helps out to remove outliers from data

20/Oct-2022 Data Science Thursday

→ Scatter plot is for bivariate analysis (it needs two variables for plotting)

→ sizes(15, 200)
minimum radius, maximum radius

→ relplot stands for relational plot
→ Scatter plot is for numerical data, x and y should be numerical. But hue parameter is always categorical. style can or cannot be numeric.

→ catplot (x-axis: categorical, y-axis: numerical) categorical plot

→ First quartile means 25% values of your data is less than this value.
median is 50% of data.

→ Boxplot is used to detect ~~boxplot~~ outliers
→ Outlier disturbs your mean in a drastic way.

→ Median is changed according to confidence interval

→ pdf for discrete data
→ Descriptive Statistics Inferential Statistics
→ Central Tendency: ~~values of~~ data where values are more
→ std is the under root of variance, so std values are small than var

26/Oct/2022 Data Science Wednesday

→ Histogram is for single variable
→ Scatter plot is for 2 variables
→ Distplot :- to do univariate analysis on continuous/numerical data
→ Combines :- histogram, kernel density function, Rugplot
→ x-axis: bins, y-axis: count/frequency → histogram

→ By default rug is false in displot.
→ Rugplot tells you the information of count of every value in the bin

→ Displot has probabilities on Y-axis

→ pdf makes from histogram / applying kernel density estimation

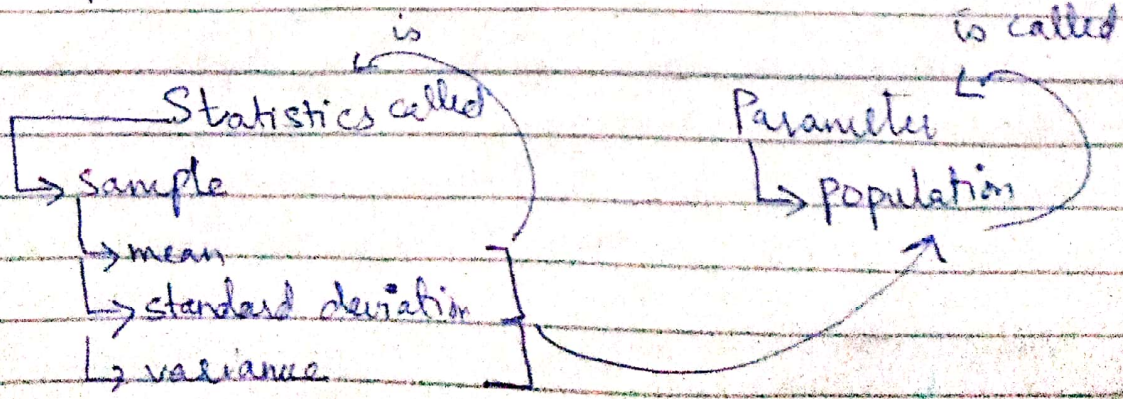
Statistics

→ collect data and draw conclusions from it
→ Descriptive and Inferential Statistics

Sampling from population

Draw conclusions from sample data

Sample must be representative of population



26/8/2022

DS:

→ Mean, std and var of population is called parameter
→ Mean, std and var of sample is called statistics

Descriptive Statistics

→ Mean
→ Median, Mode, std deviation
→ variance, MAD (Mean Absolute Deviation or Median Absolute Deviation)

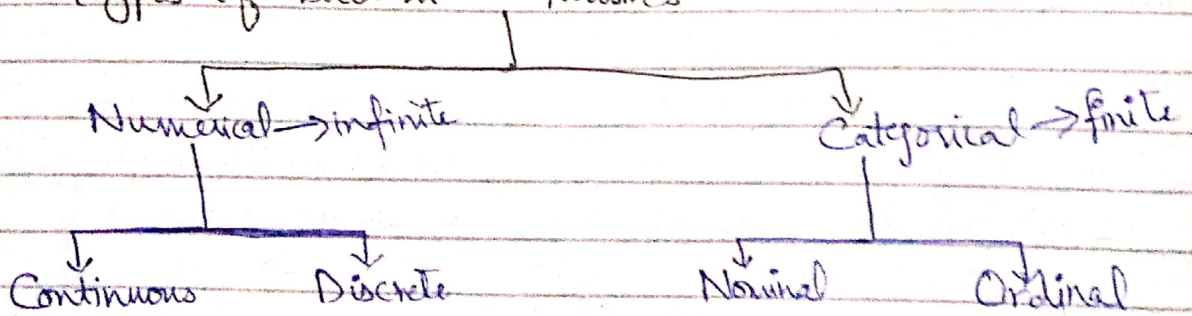
Inferential Statistics →

is something really different from other or occurred by chance

→ Confidence Interval
→ Hypothesis Testing

Descriptive Statistics

Types of Data in Statistics



→ Discrete: values are countable / distinct values
number of mobile apps in phone

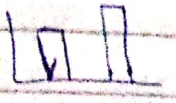
→ Nominal: values are independent of each other
→ Ordinal: values have relationship with each other / Order | 1 2 3 4 5 | 1 and 2, 3 and 4 like have relationship with each other for example, 2 is better 1 so its have relationship

→

26/oct-2022 Data Science Wednesday

→ Visualization Techniques
→ then Measure of Central Tendency
→ Measure of Dispersion

male / female



Visualization Techniques

→ Bar chart or pie chart when data is categorical and nominal
→ Graph do not tell us relationship
→ Bar and pie chart is used when data is categorical and ordinal

→ Histogram for Numerical data

→ Scatter plot for bivariate numerical data

→ for Numerical and Continuous data (univariate analysis) you can use boxplot

→ Bar chart when one variable is categorical and one is numerical

Bivariate Analysis

→ Categorical and Categorical data → Crosstab
↳ bar plot / heatmap / cluster map

Correlation

Measure of Central Tendency

→ where the most values occur in the data
→ Median does not change whether you add outlier
→ You should use 3 of them (mean, median, mode) to have fair idea of the centrality of data

26/Oct-22 Data Science

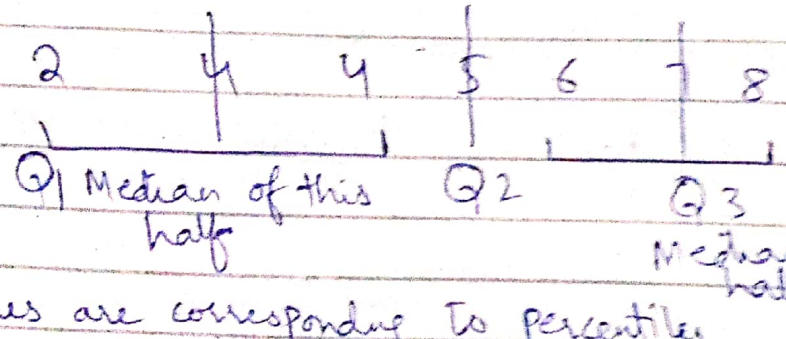
→ Measure of Spread :-

→ Range → max - min
→ Interquartile Range
→ Standard deviation
→ variance

Percentile: value/percentage of the percentage where below data is found

Quartiles : Divide the data into quarters

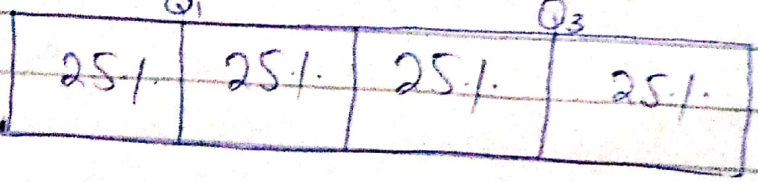
Q2 is the median



Quartiles are corresponding to percentiles

Interquartile Range

Interquartile range is the measure of statistical dispersion between upper (75th) and lower (25th) quartile.



$Q_3 - Q_1$ is Interquartile Range (IQR)

Boxplot's box difference is IQR

27/Oct 2022 Data Science

Mean Absolute Deviation:

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

$\bar{X}$ → mean of your data.

n → number of points in sample space

X → individual data point

Variance

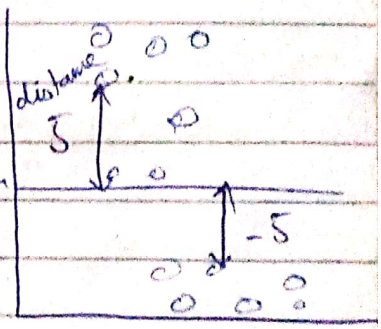
Means how far the datapoints are spread out from the mean.

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

μ → mean

It will cancel out each other's mean effect, thats why we use square

→ Outliers get more weitage

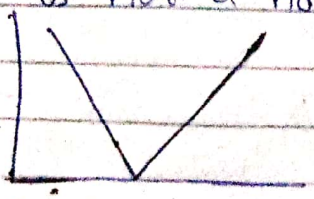


Problem

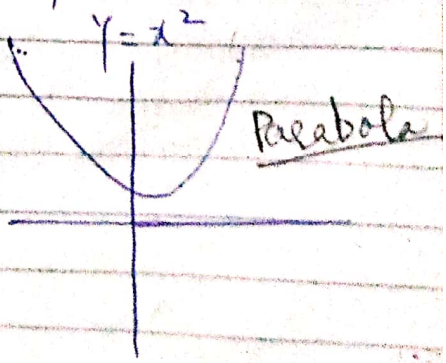
Data unit changes from cm to cm^2

Modulus is ~~not~~ a non-smooth function

$y = |x|$ graph

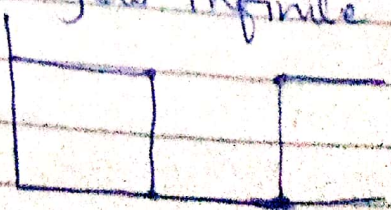


It just simply turns the straight line upward or downward



Discontinuity: y ki ath value k corresponding x ki 2 values ati han. Derivative gets infinite so graph gets non-smooth

If Derivate is 0 or ∞, then means discontinuity



27/Oct/22 Data Science

9. Median Absolute Deviation
↳ This function is in statsmodels library

→ Gives good results when data has outliers

Practical Statistics for Data Scientists
Book
Peter Bruce, Andrew Bruce and Peter Gedeck

Probability Density functions (PDF)
Kitna zaida, Kitna kam
→ pdf is made from histogram
→ kde is applied on histogram

pdf value = 1 occurence of 1 bin / Total occurrences of all bin values

→ Gives you smooth graph for univariate analysis
→ for multivariate, you get confused between values
But intersection point is taken to calculate high/low probability of species/class/values

Cumulative Density/Distributive function.
cdf. 1.5 pa 70% of values kam se kam
pdf: 1.5 aaya ki probability 0.28 or 28%
h
cumsum in numpy (cumulative sum)

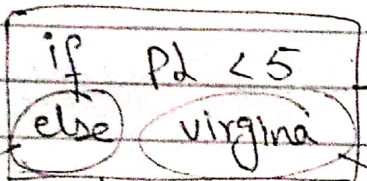
2/Nov/2022 Data Science Wednesday

→ KDE is smoothing function
→ PDF (probability ki density show karne ka function)

→ PDF tells you probability probabilities of occurrence
→ CDF tells you that below this point how much data/people exists
↳ Basically tells you the summary of data till that point

→ pdf is helpful when data is multivariate
↓ → occurrence of count of bin
1 point: 50/120 → total count

In what kind of scenarios cdf helps
→ When you have overlapping curves



Accordg to cdf. → is k ghlt hony ka chance : 10%.

* cdf helps in quantifying your decision

Normal Distribution

→ unknown task is done using known patterns (modding unknown in a way, so that it can be resolved with know techniques)

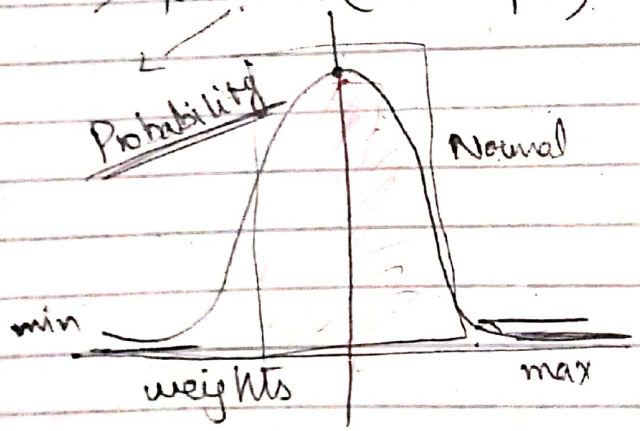
2/Nov.2022 Data Science Wednesday

→ fourier: sine and cosine wave combination can transform every signal

Area under the curve is calculated by (integration) its sum

Distribution:

1. Piosson
2. Bernoulli
3. Binomial
4. Normal Distribution
   ↳ Gaussian (Bell shaped)



→ There are less people in world, whose weights are less
→ There are less . . . , whose weights are maximum
→ There are more people, whose weights are normal

Central limit Theorem: says
* Every distribution can be converted into normal distribution

Normal Distribution Curve Equation

$$P(X=x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = mean
σ = standard deviation
σ^2 = variance

Normal distribution has same mean, median, mode

If $\mu = 0$, $\sigma = 1$

Equation:

$$P = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$\frac{1}{\sqrt{2\pi}}$ constant

Always for Normal curve, variance/std = 1 and $\mu = 0$

Standard Normal Variate : $\mu = 0, \sigma = 1$

loc = μ, scale = std in formula

↳ subtract mean and Divide by std to make the distribution Standard normal Variate

np.std(draw)

z-score : $\frac{\text{score} - \text{avg score}}{\text{std}}$ → mean

1.66 = $\frac{1.6 + 0.06}{\text{std}}$

z table 0.9515

xply by 100 = 95.15%

Student scored more than 95% of all students

3/Nov/2022 Data Science Thursday

statistics X

Data Analysis Process

Steps
→ Asking questions
→ Data Wrangling / Munging
→ Exploratory Data Analysis
→ Drawing conclusions
→ Communicating results

Speech therapy for Autistic patients

① Asking questions:
for better questions:—
→ You must have subject/domain expertise
→ Experience

② Data Wrangling:
Steps:
① Gathering data
② Assessing Data
③ Cleaning data

— Gathering data (API : Application program interface)
* csv files
* API (data in json format)
* web scraping (data in html format)
* databases

— Cleaning data
* Missing data (mean)
* Remove duplicate data (drop_duplicates)
* Incorrect data type (astype)

3/Jan/2021 — Data Science — Thursday

③ EDA:

① Explore → Visualization

② Augment → changing according to the patterns seen in visualization
- → Removing outliers
- → Merging dataframes
- → Adding new column

This EDA is also called Feature Engineering.

④ Drawing Conclusions:
- → Machine Learning
- → Inferential statistics

} Data Scientist

→ Descriptive Statistics } Data Analyst.

Till EDA, Data Scientist and Data Analyst have to cover all the steps.

Note:
All the steps are not linear though!!!

Dirty Data: Quality issues (Duplicate, missing, corrupt, inaccurate) — noisy

Messy Data: Structural issues (each variable forms a column, each observation form a row, each observational unit forms a table)

Name
- First Name
- SurName

Contact