INTRODUCTION TO STATISTICS

STATISTICS – This is the science of collecting, organizing and analysing the data.

DATA – Collection of facts / pieces of information

Types of Statistics –

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics – Consists of organising and summarising the data using visualisation plots. Extensively used in EDA + FE. By using descriptive statistics, we can understand the data.

Example: Histogram, Bar chart, Pie chart, Distribution. What is the average age of students in the classroom? Relation between age and weight?

Inferential Statistics – It consists of collecting sample (n) data and making conclusions about population(N) data using some experiments. Conclusion can be made using hypothesis testing consisting of techniques like confidence interval, P-value, Z-test, t test, chi square test, Anova (F-test)

Example: University – 500 students (population); Class room A – 60 people (sample); Making conclusion of the average age of the entire university. Are the average age of the students in the classroom less than or greater than the average age of students in university?

Life cycle of a DS project –

1. Requirement Gathering – Identify the problem statement
2. Finalising the database / data source
3. Exploratory Data Analysis
4. Feature Engineering
5. Feature Selection
6. Model Training
7. Hyperparameter Tuning
8. Model deployment

Statistics will be used in steps 3,4,5,6,7,8

Sampling techniques –

1. **Simple Random Sampling**: Every member of the population (N) has an equal chance of being selected for the sample(n). Ex: Exit polls, lottery
2. **Stratified Sampling**: Strata -> Layers -> Clusters ->Groups. We focus on picking samples from a group Ex: Male/Female, educational degrees, blood groups

3. **Systematic Sampling**: Method which targets every nth individual out of population (N). Ex: Credit cards at airport – agent 1 will approach every 5th person and agent 2 will approach every 9th person to sell credit cards.
4. **Convenience Sampling**: Only those who are interested in the survey will be considered for participation. Ex: Students interested in DS program will be sent brochures and information regarding the course; Job application by candidates who are interested in the particular job

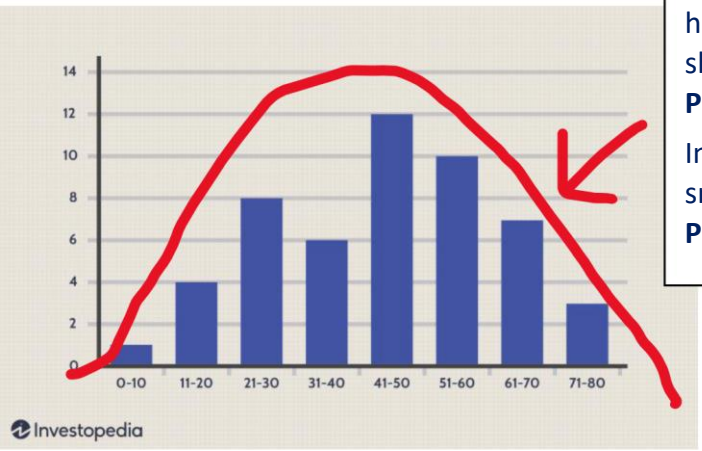
Variable – Is a property that can take any values. There are 2 different type of variables namely,

1. **Quantitative variable** – Measured numerically (mathematical operations can be performed) Ex: Age, weight, temperature, distance, rainfall etc.,
    a. **Discrete Variable** – Whole number (+/-) but no decimal values
    b. **Continuous Variable** – Numbers with decimal also
2. **Qualitative variable** – Categorical variables which are grouped together based on some common features. Ex: Gender, type of flowers, various movies etc.,

BASIC TERMINOLOGIES OF STATS

Histogram – A graphical representation that condenses data points into an easy interpretation of numerical data by grouping them into logical ranges / bins. Steps to be followed to create a histogram -

1. In a given variable, sort the numbers
2. Create no of groups -> bins
3. Define size of bins -> (max value of the given list / desired number of bins)
    Ex: Age = {1,2,3,4,5,6,7,8,9,10}
    Desired number of bins = 5
    Size of bins = 10/5 = 2



In case of continuous variables, once the histogram is smoothened, it takes a shape of a bell curve. This is called **PROBABILITY DENSITY FUNCTION (PDF)**.

In case of discrete variables, the smoothened curve is represented as **PROBABILITY MASS FUNCTION (PMF)**.

Measure of Central Tendency (Mean, Median, Mode) – This is a single value that attempts to describe a set of data identifying the central position of the dataset.

X = {1,2,3,4,5}

Mean: (1+2+3+4+5) / 5 = 3
Population Mean (μ) = $\sum\_(i=1)\text{^}N\blacksquare\ [\![Xi/N]\!]$
Sample Mean (x̄) = $\sum\_(i=1)\text{^}n\blacksquare\ [\![Xi/n]\!]$

Median: If we have outliers, we should use median instead of mean.

1. Sort the numbers
2. Find the central number in the list
    a. If the number of elements is even, we find the average of central elements
    b. If the number of elements is odd, we just pick the central element

Mode: It is the most frequent element in the list OR most repeated element in the list. Practical implementation is when in a set of categorical variables, there are NaN values present. The most repeated value (Mode) can be used to replace NaN value.

Measure of Dispersion –

1. **Variance (σ_2)** – Talks about spread of data. Higher the variance, higher is the spread of data
2. **Standard Deviation (σ)** – Talks about how many standard deviation away a number falls from mean

Population Variance (σ_2) = $\sum\_(i=1)\text{^}N\blacksquare\ [\![(Xi-\ \mu)2\ /N]\!]$
Sample Variance (s_2) = $\sum\_(i=1)\text{^}n\blacksquare\ [\![(Xi-\ \bar{x})2/n-1]\!]$

It is natural to wonder why the sum of the squared deviations is divided by n−1 rather than n. The purpose in computing the sample standard deviation is to estimate the amount of spread in the population from which the sample was drawn.

Ideally, therefore, we would compute deviations from the mean of all the items in the population, rather than the deviations from the sample mean.

However, the population mean is in general unknown, so the sample mean is used in its place.

It is a mathematical fact that the deviations around the sample mean tend to be a bit smaller than the deviations around the population mean and that dividing by n−1 rather than n provides exactly the right correction.

Percentiles & Quartiles:

Percentile is a value below which a certain percentage of observations / data points lie.

Ex: 99th percentile – It means this person has got better marks than 99% of the entire students.

5 number summary: This can be used to remove outliers.

1. Minimum
2. First Quartile (25 percentile) Q1
3. Median
4. Third Quartile (75 percentile) Q3
5. Maximum

Example: {1,2,2,2,3,3,3,4,5,5,5,6,6,6,6,7, 8,8,9,27}

Lower Fence = Q1 – 1.5*IQR

Higher Fence = Q3 + 1.5*IQR

IQR (Iner Quartile Range) = Q3 – Q1

Q1 = 25/100 * (20+1) = 5.25 index (take average of 5^{th} and 6^{th} index – 3)

Q3 = 75/100 * (20+1) = 15.75 index (take average of 15^{th} and 16^{th} index – 7.5)

IQR = Q3 – Q1 = 7.5 – 3 = 4.5

Lower Fence = 3- (1.5 * 4.5) = -3.65

Higher Fence = 7.5 + (1.5*4.5) = 14.25

Conclusion: Since the lowest value in the dataset of 1, there is no outlier present in the lower fence.

However, 27 is greater than higher fence value 14.25. Hence this can be treated as an outlier and eliminated from the list.

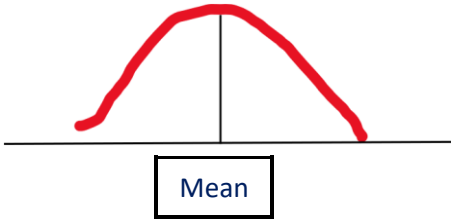
The five number summary would look something like this –

1. Minimum = 1
2. First Quartile (25 percentile) Q1 - 3
3. Median - 5
4. Third Quartile (75 percentile) Q3 – 7.5
5. Maximum – 9

INTRODUCTION TO DISTRIBUTION

1. Normal Distribution
2. Standard Normal Distribution
3. Z-Score
4. Standardization and Normalization

Gaussian / Normal Distribution –



This bell curve indicates distribution / spread of data.

1.Both sides around the mean are symmetrical / equal.
2.The area under the bell curve is 1 -> 100%

Emperical Rule of Normal / Gaussian distribution:

Assumptions of the empirical rule –

1.Within the 1st SD on either sides of mean, there are 68% of data present
2.Within the 2nd SD on either sides of mean, there are 95% of data present
3.Within the 3rd SD on either sides of mean, there are 99.7% of data present

{68-95-99.7%} – Emperical Rule

Use QQ plot, to determine whether a distribution is Gaussian or not.

Standard Normal Distribution:

Assume a variable X belonging to Gaussian distribution with mean (**μ**) and standard deviation (**σ**). This variable can be converted into a different variable y belonging to standard normal distribution with mean (**μ=0**) and standard deviation (**σ=1**) using the formula Z-score (which can be interpreted by standard scaling / standardisation). **The main reason to perform this is to standardise all the different units into one comparable unit which can increase the speed of calculation.** 3

Ex: X = {1,2,3,4,5} -> **μ = 3; σ = 1.41**

Z-score = (Xi - μ) / (σ / sqrt (n))

Where -

(σ / sqrt (n)) -> standard error which is used in inferential statistics

Standardization – The process where a dataset or a distribution is transformed such that it is centred around zero with a standard deviation of 1. The main objective of standardization is to ensure that all the columns within a dataset can be compared on a similar scale.
Standardization can be helpful in cases where the data follows a Gaussian distribution.
Standardization does not get affected by outliers because there is no predefined range of transformed features. Ex: Standard Scaler.

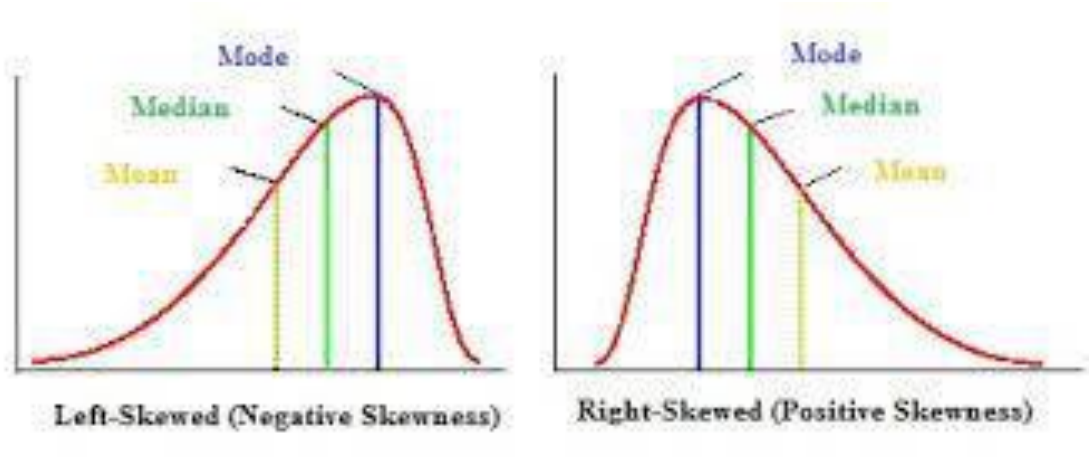
```
X_new = (X - mean)/Std
```

Normalization – The process of transforming a dataset where we define a specific range in which data needs to be transformed. Ex: MinMax Scaler. Normalization is useful when there

are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

```
X_new = (X - X_min)/(X_max - X_min)
```

Log normal Distribution - if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, X = exp(Y), has a log-normal distribution.



Area under the curve can be calculated by computing the Z-score and referring the Z value in the table, link given below.

Z table link - https://www.z-table.com/

MORE CONCEPTS OF STATS

1.Central Limit Theorem
2.Probability
3.Permutations and Combinations
4.Covariance, Pearson Correlation, Spearman Rank Correlation
5.Bernoulli's Distribution
6.Binomial Distribution
7.Power Law (Pareto Distribution)

Central Limit Theorem (CLT)

CLT says a population (N) which is either gaussian / log normally distributed, considering any number of samples with size of n >= 30, then the distribution of sample means follows a normal / gaussian distribution.

Probability – Measure of the likelihood of an event.

Ex – Tossing a fair coin. P(h) = 0.5

Additional Rule of Probability -

1. **Mutual exclusive events –** Two events are mutually exclusive if they cannot occur at the same time. Ex – Rolling a dice; Tossing a coin; Winter or Summer;
   P(A or B) = P(A) + P(B)
2. **Non-Mutual exclusive events** – Two events can occur at the same time. Ex – Picking random a card from a deck of cards, two events like a "heart" and "king of heart" can appear at the same time.
   P(A or B) = P(A) + P(B) – P(A*B)

Multiplication Rule of Probabilities -

1. **Dependent events –** Two events are dependent if they affect each other.
   Ex **–** From a bag of 4W and 3Y marbles, pick a marble. Probability of it being a white marble is 4/7. Later, pick a yellow marble, probability of it being a yellow marble is 3/6. Notice initially we had 7 marbles and next we have 6 marbles. Hence the first event has affected the outcome of the second event and hence the name dependent events.
2. **Independent events –** Two events are independent if they don't affect each other.
   Ex – Tossing a coin.

Permutation –

nPr = n! / (n-r)! where,
n = total number of objects
r = number of selections

Combination –
nCr = n! / r!(n-r)!

Covariance – Covariance is the measure of changes between two random variables in statistics. In other words, we can say how will a variable y change when x changes.

Ex: Age and Weight. As age of a person increases, weight will tend to increase.

Covariance is one of the techniques used for feature selection. Covariance doesn't have any restriction on range or scale of +/- values that can be populated as the result. It may be +3000 or -6543 for example.

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

cov_{x,y} = covariance between variable x and y

x {i} = data value of x

y {i} = data value of y

\bar{x} = mean of x

\bar{y} = mean of y

N = number of data values

Three types of covariance –

1. Positive covariance – If x increases, then y also increases and vice versa.
2. Negative covariance - If x increases, then y decreases and vice versa.
3. Zero covariance – No relation between x and y

Pearson Correlation Coefficient (PCC) – The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. Covariance doesn't have any restriction on range or scale of +/- values that can be populated as a result. It may be +3000 or -6543 for example. To overcome this, pearson correlation coefficient has a range of values between -1 to +1. More the correlation value tending towards +1, stronger is the correlation (positive). More the correlation value tending towards -1, weaker is the correlation (negative). PCC is used only for linear data.

r = Cov(x,y) / (σx * σy)

where σx = standard deviation of x; σy = standard deviation of y

Spearman Rank Correlation – Whenever we encounter non-linear data, spearman rank correlation has to be used. This basically assigns a rank to every data point which is used to calculate the correlation. Rank is assigned in the ascending order for every data point. If two data points are same, then the assigned rank will be same.

r(s) = Cov[R(x), R(y)] / [σ(R(x)) * σ(R(y))]

where R -> Rank assigned to each data point

INTRODUCTION TO INFERENTIAL STATISTICS

Topics to be covered:

1. Hypothesis testing – Z test, t test, chi square test, Anova test (F test)
2. P-value
3. Confidence interval
4. Significance value

Inferential Statistics is a study where we make assumptions about the population (N) data with the help of sample (n) data and arrive at certain conclusions. In order to validate these assumptions / conclusions, we use HYPOTHESIS TESTING.

Hypothesis Testing – Steps to be performed to carry out hypothesis testing:

1. Null hypothesis – Ex: A coin is fair
2. Alternative hypothesis – Ex: A coin is unfair
3. Perform experimentation – Ex: Toss the coin 100 times -> record each result -> Let's say in one case, we got 50 times head and 50 times tail. We can conclude that this is a fair coin. Consider another scenario where we got 60 times head and 40 times tail.

   We can still agree that the coin is fair. But what if we got 75 times head and 25 times tail??? We might tend to disagree with the null hypothesis which states the coin is fair. In every situation, the range of acceptable outcomes will change and this range is called **CONFIDENCE INTERVAL.**
4. Conclusion – If the outcome is within the confidence interval, then **we fail to reject the null hypothesis**. Else we can say **null hypothesis is rejected.**

Point Estimate – The value of any statistics (sample mean $\bar{x}$) that estimates the value of a parameter (population mean μ) is called point estimate.

$\bar{x}$ +/- Margin of error = μ

Confidence interval ->

Lower CI: Point estimate – Margin of error

Higher CI: Point estimate + Margin of error

$$\mathbf{MOE}_{\gamma} = z_{\gamma} \times \sqrt{\frac{\sigma^2}{n}}$$

Problem Statement: On the quant test of CAT exam, a sample of 25 students has a mean of 520 with a population standard deviation of 100. Construct a 95% CI about the mean.

Solution:

n=25; $\bar{x}$ = 520; σ = 100; CI = 95%;

Significance Value (SV) = 1 – CI = 1 - 0.95 = 0.05

WKT, Lower CI = Point estimate – Margin of error

& Higher CI = Point estimate + Margin of error

Lower CI = 520 – Z(0.05/2) * (100/5)

Z(0.05/2) = Z(0.025) = Z(1-0.025) = Z(0.975) = 1.96

Lower CI = 520 – 1.96*(20) = 520 – 39.2 = 480.8

Similarly, Higher CI = 520 + 1.96*(20) = 520 + 39.2 = 559.2

Conclusion: Since the sample mean 520 is within the range of confidence interval ranging between 480.8 & 559.2, we fail to reject the null hypothesis.

Problem Statement 2: n=25; $\bar{x}$ = 520; sample standard deviation (s) = 80; CI = 95%; WKT SV = 1-CI = 1-0.95 = 0.05

Since sample standard deviation is given, we need to use t-test and the formula changes to $\bar{x}$ +- t(a/2) * (s/sqrt(n))

Since we are using t-test, we need to calculate the degree of freedom which is nothing but n-1. In this case, degree of freedom will be 25-1 = 24.

Lower CI = 520 – t(0.05/2) * (80/sqrt(25)) = 520 – 2.064*16 = 486.976

Higher CI = 520 + t(0.05/2) * (80/sqrt(25)) = 520 + 2.064*16 = 553.024

ALWAYS USE Z-TEST WHEN WE HAVE POPULATION STANDARD DEVIATION OR WHEN SAMPLE SIZE IS GREATER THAN 30 AND USE T-TEST WHEN WE HAVE SAMPLE STANDARD DEVIATION AND SAMPLE SIZE LESS THAN 30

n>= 30 or population standard deviation ➋ Z-test

n>30 and sample standard deviation ➋ t-test

Conclusion: Since the sample mean 520 is within the range of confidence interval ranging between 486.976 & 553.024, we fail to reject the null hypothesis.

1-tail and 2-tail test:

Problem Statement: A factory has a machine that fills 80ml of medicine in a bottle. An employee believes that the average amount of medicine is not 80ml. Using 40 sample he

measures the average amount dispersed by the machine to be 78ml with a standard deviation of 2.5

1. State null and alternate hypothesis
2. At 95% CI, is there enough evidence to support if the machine is working properly or not?

Solution: Initial analysis of the problem statement reveals this is a 2 tail test since the employee believes that the average amount of medicine is not 80ml. It can be greater than or less than 80ml. Hence this is a 2 tail test.

Null hypothesis – Mean (μ) is 80

Alternate hypothesis – Mean (μ) is not equal to 80

WKT, $\mu = 80$; $n=40$; $\bar{x} = 78$; $s = 2.5$; CI = 0.95

SV = 1- CI = 1 – 0.95 = 0.05

In this case, since we have sample standard deviation, but sample size is greater than 30, we use Z-test

The SV is available on both sides of the distribution around the mean. Area under the curve can be found out by considering 1 – 0.025 = 0.975. From Z table, the value received against 0.975 is 1.96

Z-score = (Xi - μ) / (s / sqrt (n)) = (78-80) / (2.5/sqrt(40)) = -5.05

The value from Z table obtained for SV of 0.025 is +- 1.96. Since -5.05 does not fall in the range of +1.96 to -1.96, we reject the null hypothesis.

Chi Square test – This test explains about population proportions. It is a non-parametric test that is performed on categorical variables / datasets (both nominal and ordinal). Ex: Rank

Problem Statement: In 2000 USA census, age of individuals in a small town was found to be the following –

	< 18	18-35	>35
Expected	20%	30%	50%

In 2010, ages of n =500 individuals were sample and below are the results –

	< 18	18-35	>35
Observed	121	285	91

Using SV = 0.05 who would you conclude the population distribution of ages has changed in the last 10 years or not?

Solution:

H_0 – The data meets the expected distribution

H_1 – The data does not meet the expected distribution

Degree of freedom = No of categories – 1 = 3 – 1 = 2

Decision boundary: Refer to chi square table -> First column df -> under row 2 look out for SV of 0.05 -> The value obtained is 5.991

	< 18	18-35	>35
Observed	121	285	91
Expected	100	150	250

$X_2 = \sum (f_0 - f_e)_2 / f_e$ where –

X_2 – Notification of Chi Square parameter

f_0 – Observed value

f_e – Expected value

$X_2 = [(121-100)_2 / 100] + [(285-150)_2 / 150] + [(91-250)_2 / 250]$

$X_2 = 232.494$

Conclusion: Since X_2 is greater than 5.991, we reject the null hypothesis

Anova test (F-test) – Analysis of Variance (ANOVA)

ANOVA, or Analysis of Variance, is a test used to determine differences between research results from three or more unrelated samples or groups. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

The t-test is a method that determines whether two populations are statistically different from each other, whereas ANOVA determines whether three or more populations are statistically different from each other.