Statistics

Statistics : Using data gathered on a group to reach conclusion about the same group.

Inferential statistics: gathering data from a sample and use the statistics to bring conclusion about the population

Population : a collection of persons, object or items EX:all automobiles, all employees of optum.

Census: When researchers gathers data from whole population of given measurement of interest they call it census. EX: US population census taken every 10 years.

Sample: a portion of whole or representation of the whole population. EX: 75% samples of dairymilks from testing quality.

Descriptive Statistics: Using data gathered on a group to reach conclusion about the same group.

Inferential statistics: Gathering data from a sample and use the statistics to bring conclusions about the population.

**Types of Variable:**

Categorical – nominal, ordinal

Numerical – interval, ratio

**Mean :**

**Media :**

**Mode :**

**Percentile:** 15, 28,34,68,82,85,96 🡪 (50\*7)/100 = 3.5 which is 4 and 68 is the 50% in the range

**Quantiles:** Q1=0-25 percentile, Q2=25-50 percentile, Q3=50-75 percentile, Q4=75-100 percentile

**Hypothesis testing:**

* Dairy milk of factor is 5gm (Ho Hypothesis)
* If motor has issues it would have issues
* Sample of 30 pks and average will measure
* 4.6 (Ha Null Hypothesis)
* 4.9 (Accepted null hypothesis)

**Variance:**

* Temp of Delhi: 20,19,39,37,46,42,38,32,28,19 - Far away from the mean, variance is more
* Temp of Mumbai: 19,18,20,25,18,17,20,25,23,21 - Near the Mean, Variance is less

1. Calculation: x = Temp of Delhi , x^ = avg of all values 36.5, variance = (x- x^ )2/N
2. Sigma = (20-36.5)2/10+(19-36.5)2/10+…….
3. Variance = Sigma square(σ2)
4. Standard deviation = Sigma or (root of variance)

Q . How should we decide 4.7 is to be accepted or rejected the null hypothesis?

Marks in a class 25,35,56,50,56,59,67,68,75,77,78,82,91,98

μ = population

x^ (x Bar) = samples

σ = Standard deviation

σ2 = Variance

Standard deviation: Square root of Variance

**Empirical Rule: Normal distributed**

* **μ+- 1 σ 68%**
* **μ+- 2 σ 95%**
* **μ+- 3 σ 99.7%**

μ = 5gm

σ=0.1gm

μ+- 2 σ = 4.8 or 5.2 covers 95% of the samples we can ignore.

* Credit card fraud detection and security.

**Chebyshev’s Theorem: Data is not normally distributed, any shape of plot we can use this.**

* Within K standard deviation of the mean μ+-k σ lie at least
* Proportion of the values assuming k>1

**μ+- kσ (1-1/k2)100 formula**

K=2 (1-1/4)100 = 75%

K=3 (1-1/9)100=88.8%

Q. How do we know data is normally distributed or not…

* plot the data.
* Skive and kuntoses to calculate.

Q. Computer profanations are young, Avg age of computer prof =28, standard deviation = 3 years. Not normal distribution, more of them are 20s and less of 40s. which max age to be targeted to attract employees.

**Μ=28, σ=3**

**0.8=(1-1/k2) assume 1/k2=0.2, k2=5, k=2.24**

28 +- 2.24\*3 = 21 to 35 is the age group max of the employees are been targeted.

**Z Score: Distribution of data is present**

**Formula : z = (x- μ)/ σ**

How many times the mean is away from standard deviation

μ=5gm σ=0.1 and sample is 4.8

4.7-5/0.1=3

So 3σ it is away from the mean which is good. We should accept the null Hypothesis.

**Q2 . If x^=5.15 and z score = 5.15-5/0.1=1.5 which is near to 2 where most 95% of hypothesis is correct.**

* **Confidences interval if 2 is not under our confidences interval we can reject the null hypothesis**
* **Confidences interval is 95%.**
* **Significance interval**
* **1 – Confidences level is significance interval.**

1. **If I want 95% of confidences level what should be my σ ?**

**0.95=1-1/k2**

**0.5=1/k2**

**k2=2**

**k=1.414**

* **if the value of k is less than 1.414 I should reject the null hypothesis.**
* **Z score table is available can check and decide the confidence level.**

**Coefficient of Variations:**

* **Cv= (σ /μ)100**

**How many times (standard deviation) σ with the (mean) μ is, cv helps us in saving in below share trap. Exact variance of the data is defined.**

**Actual spread in data**

* Share price of two share one has high variance and other has less which is risk. Which has high variance is risk.
* But it could be trap
* Share a =100

σ =5, μ=100

μ-2σ = 90 - Variance

loss is 10%

* CV= (σ /μ)100

(5/100)100=5%

* Share b = 1000

σ =10, μ=1000

μ-2σ = 980

loss is 2%

* CV= (σ /μ)100

(10/1000)100 =1%

Q. How is hypothesis testing is use full.

Bank data analysis. Ho is credit score, how that is impacting the customers leaving the bank or not.

Ha= null hypothesis credit score will affect the customer leaving the bank. (assume).

Q. Diary milk μ=5, machine has issues

Ho=5gm

Ha= test the system. Take sample of 50 diary milks find x^=4.6.

Confidence interval =95%

Critical value = 2 ------(x-+2 σ =95%)

**Chebeys Theorem:**

0.9=1-1/K2

1/k2=0.1

**K=3.2**

**Z score** = 4.6-5/0.1= -4

σ = 0.1, μ=5

x^ is -4 times σ away from the population mean. so we should reject the null hypothesis.

X+-2 σ = 95%

**IQR - Interquartile Range: Normal distributed.**

* Ascending order
* Min -1st one fourth Q1 - 1st two fourth-Q2 -1st three fourth-Q3 -Max
* IQR = Q3-Q2
* Removes extreme out liars
* Data spread in 50% of a dataset.
* Distance between Q1 and Q3.
* Smaller range with good data.
* If not in normal distributed, then when we need to log of the data to make it normal distribution.

**Covariance:**

* Mulita features.
* Relation Change in one feature vs change in another feature
* Covariance take correlation for calculation but no dependency
* Qualitative analysis.
* **Cov(x,y)=1/1-n   (i 1 to nΣ(x-x^)(y-y^))**

**Correlation:**

* How much extent of X & Y are related?
* Quantitative analysis
* Range from -1 to 1
* ----Graph go’s up
* Correlation 0 to 1 >>0
* Covariance is +ve
* ----Graph go’s down
* Correlation -1 to 0 <<0
* Covariance is -ve
* -----Random
* Covariance -ve or +ve –Good variance change in x change in y but not straight
* Correlation = zero
* ------Constant graph at y
* Covariance = 0
* Correlation = 0

**Correlation: Formula**

**i 1 to nΣ ((x-x^)(y-y^)/sqr(x-x^)2(y-y^)2)**

**Kurtosis, Skewness:**

**Skewness:**

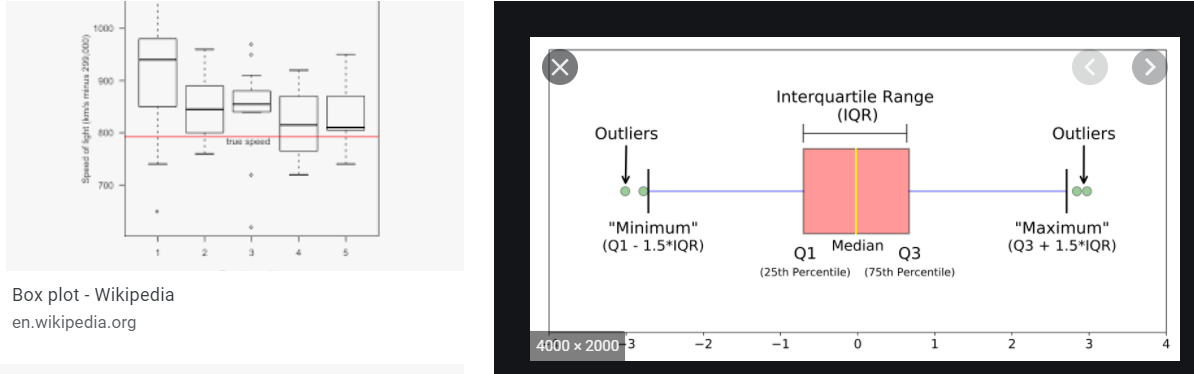
* Major on shape of data
* Box plots.
* Normal distributed symmetrical – skewness=0 (mean=median=mode)
* Negatively skewed – left skewed – Skewness <0 (good example: share when market is down)
* Positively skewed – right skewed – Skewness >0
* Formula
* Python pands .skew()

**Kurtosis:**

* Leptokurtic(+) Peak
* Mesokurtic (0) normal
* Platykurtic(-) widely spread data
* Python pandas .kurt()

**Box Plots:**

* Box and whisker plot
* Q1= min, Q3=max, Q2=median
* .boxplot



**Chi Square test** – Nominal features, in categorical features we can’t calculate the mean so in those cases we will use chi square test. Age, Gender etc.

O = observed value

E= expected value

**Formula: X2= Σ (O-E)2/E**

Slope would be graph**.**

**Bank customer from locations : contingency matrix**

**Status|France|spain|Germony| total**

**Leaving|10|18|5|33**

**Notleaving|20|30|10|60**

**Total|30|48|15|93**

**Ho=Geography dose not impact the exiting customers**

**Ha=Geography impact the exiting customers**

**Over all leaving = 33/93=0.35**

**Over all not leaving = 60/93 =0.64**

**Then leaving is impacted = 0.35x30=11 customer should leave**

**0.350X48=17**

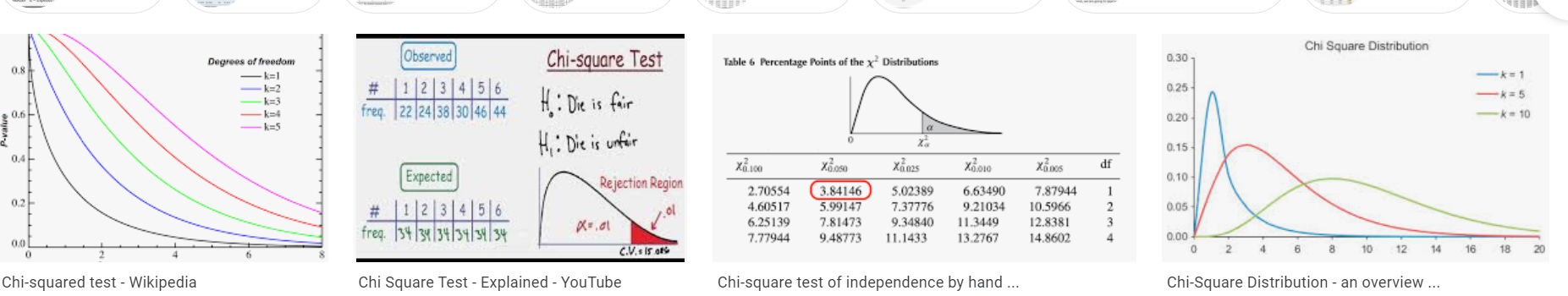
**0.35x15=5**

**For Not leaving Eo=19|31|10 based on leaving proportionate.**

**Chi square test = X2= Σ (O-E)2/E**

**Chi squre x2= (20-19)2/19+(30-31)2/31+(10-10)2/10+(10-11)2/11+(18-17)2/17+0=0.235**

* **Degree of freedom = (r-1)(c-1)=(2 rows-1)(3 columns ie Spain, Germany, France ) ie (2-1)(3-1)=2**
* **Significant level = 5%**
* **Confidence level = 95%**
* **Chi square table = degree of freedom we will have X2 values for 2 it is 5.99**
* **In our values x2=0.235 and c = 5.99**



* **We have module of python has chi from squre scipy imports stats**
* **Stats.**
* **O = [[20,30,10],[10,18,5]]**
* **E= [[19,31,10],[11,17,5]]**
* **2 = degree**
* **NOT CORRECT - Stats.chisquare(o,e,2,axis=1)**

**One sample T test** – Used for ratios

**Two sample T test** – Used for ratios

**Probabilistic Theory:**

Probability is a measure of how likely an event is .so it is 60% chance that it will rain tomorrow.

* General law of addition: P(X U Y)=P(X)+P(Y)-P(XnY)
* Special law of addition: P(XUY) = P(X)+P(Y)
* General law of multiplication: P(XnY)=P(X)\*P(Y|X)=P(Y)\*P(X|Y)
* Special law of multiplication: P(XnY)=P(X)\*P(Y)
* Marginal – P(X)
* Union – P(XUY)
* Join – P(XnY)
* Conditional P(X|Y)
* Conditional probability : P(A/B)= P(A n B)/P(B)
* Independent Event: P(X|Y)=P(X) and P(Y|X)=P(Y)
* Bayes Theorem: P(A/B)= P(B/A) \*P(A)/P(B)

**Linear regression:**

Sample example: flat in Hyderabad, price in lakh vs area in 1000sq feet.

Y=mx+c where: y=actual price, Y^=perdition, x is the size of the flat.

Error = Y-Y^ (based on experience m & c values to be improved to get the accuracy)

* If error are negative need to move to correction could be left to right and vice versa for positive error.

**(Mean square Error**) E = 1/2n **Σ(y-y^)2 =J (Cost function)**

1200 sqft – 20L

1800sqft -42L

3200sqft – 44L

3800sqft – 25L

4200sqft -62L

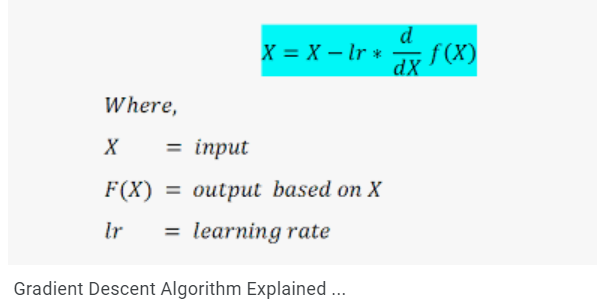
**Regression equation:** y=mx+c

* **Parameters: m & c**
* **J (Cost function) =** 1/2n **Σ(y-y^)2**
* **Goal = minimize the J by changing m & c.**

**Linear Regression:**

to decrease the cost function J we need to min m & c. For which we use different algorithms

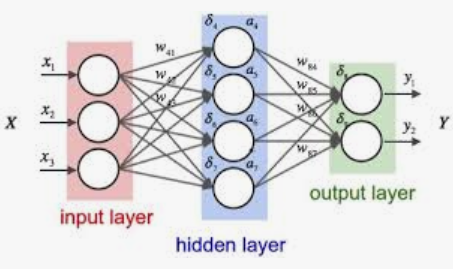
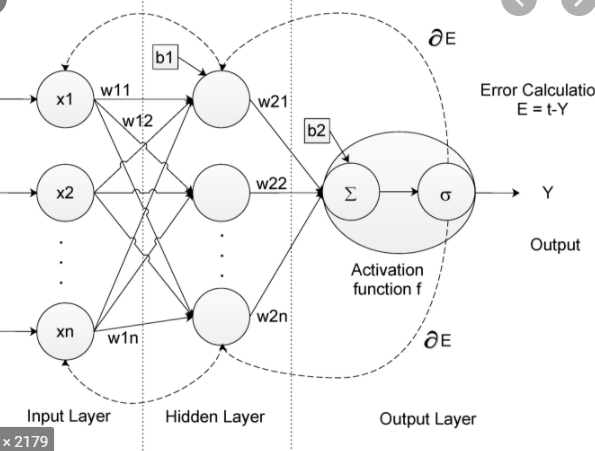
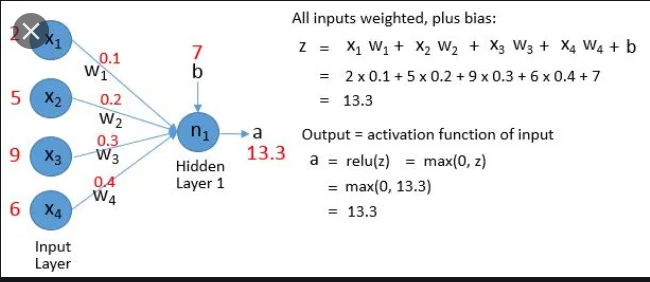
**Gradient Descent Algorithm:**

* Mn(new valume of m)=m(old) – lr(learning rate)dJ/dm
* Cn(new value of c)=c(old)-lr(learning rate)dJ/dc
* LR = Learning rate is how fast the learning rate is been done. Range 0 to 1
* 
* **Linear regression**

Neural Networks:

* Neurons connect with each other to pass the data.
* Same process we will have min 1 input and 1 output for each NN and N number of Neurons in hidden layer.
* Computation happens in hidden layers but not in input layers.
* Two process on every computational neuron

1. Weighted sum
2. Activation function

* 
* Z=x1w1+x2w2+x3w3+b (b = bias)
* Cost function y^ = f(z) = f(x1w1+x2w2+x3w3+b)
* Identity function f(x)=x
* Sigmoid function f(x)=1/(1+e^-x)
* Process of training neural network is the process of training linear regression.
* Artificial Neural networks – ANN
* ANN with weights as random values.
* We feed labeled data to random values weights ANN and calculate cost function J
* Labelled data -> ANN random weights->cost function J->Use GDA(gradient dissent) to tune weights to min J->Replace the old weights with new weights.-->**point 2**(ie ANN random weights -> with updated weights)
* From above you will get trained model with best weights.
* Feed sample input (x data with new features) to the above given trained model we will get predicted label (Y^ predicted).
* 
* 
* We have 2 activation functions

1. Linear activation function (Regression type)
2. Nonlinear activation function (classification )

* Non Linear activation function:

1. Monotonic : logistic sigmoid function can cause a neural network to get stuck at the time of training.
2. Sigmoid function: classification
3. Softmax function is more generalized logistic activation function used for multi classification

* Sigmoid function