Statistics for Machine Learning :-

Learning pattern :-

Session 1 :-
Probability distribution → discrete distribution (Binomial and Poisson) & continuous distribution.

Session 2 :-
Continuous probability distribution (Normal distribution)
This concept will lead to CLT (Central Limit Theorem), it is the key topic to hypothesis testing | Range estimate of population parameter.

Session 3 :- Hypothesis Testing - start with one sample T test.

Session 4 :- Two sample T-test (we have paired and unpaired).

Session 5 :- Proportion test → Chi square test (There are four types)

Session 6 :- ANOVA (It is a type of regression analysis).
There is a strong connection between ANOVA & SLR.

Basic Probability Questions :-

1) A gambler cheats using a loaded die, with 'one' comes up 25% of time. What is the probability of getting '6' in that die?

Solution :-
It is a biased die here,

1 comes 25% of the time what is the prob of getting 6?

If it is a fair dice, answer is 1/6.

Remaining 5 numbers have 75% chance, all are equally likely,

so, 75%/5 = 15% (answer)

---

2) What is the probability of getting a sum = 3 while throwing two die (say white and black).

Solution:- Two dice are thrown, so the sample space will be $6^2 = 36$.

The list of combinations are:-

1 + 2 = 3
2 + 1 = 3 } only two combinations.

The probability is 2/36 = 1/18.

33:30

---

3) For the above experiment, what is the conditional probability of getting a sum = 3 when white die is already rolled and turns to be 'one'.

Solution:- white die is already rolled.

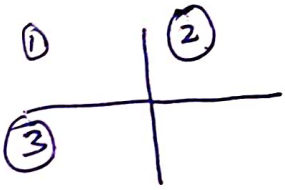
It is like a conditional probability.

$P(S=3/W=1)$ [what is the probability that sum is 3 with the condition of white die being 1

One die is already fixed, then the chances left for the second die is just six.

The solution is :- 1/6.

In probability we have four major types of calculations :-



① Mutually Exclusive events

Mutually exclusive event is going to calculate the probability of either A or B.

Example:- Head and Tail are mutually exclusive events. when Head happens tail doesn't occur and vice versa. Then P(H or T) has a 100% chance.

$$P(H \text{ or } T) = P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1$$

Therefore,

$$P(A \text{ or } B) = P(A) + P(B)$$

what is the probability of getting king or queen in pack of 52 cards :-

$$P(K \text{ or } Q) \Rightarrow 4/52 + 4/52 = 8/52$$

There is nothing common between H or T so,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B) = 0$

$$P(H \cap T) = 0$$

② Mutually Non-Exclusive Event :-

$P(A \cap B) \neq 0$ [need not be 0]

Example :- What is the probability of getting a K or ♡ in a pack of 52 cards.

$$P(K \cup ♡) = P(K) + P(♡) - P(K \cap ♡)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ where } P(A \cap B) \neq 0}$$

The above two rules are addition rules [union rules].

---

③ Mutually Independent Event :-

- It means both the events can take place together
- One event will not influence the other event.
- This is a product rule.

$$\boxed{P(A \cap B) = P(A) \cdot P(B)}$$

---

④ Mutually Non-Independent Event :-

- Here one event can influence the other event.

A influences B

$P(A \cap B) = P(A) \cdot P(B|A)$ ($P(B|A)$ means A has already happened).

we can rewrite it,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{--- (1)}$$

Now, B is influencing A, (B has already happened).

$P(A \cap B) = P(B) \cdot P(A|B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{--- (2)}$$

Equating (1) and (2), we get,

$P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

This is Baye's rule.

Naive Baye's rule works based on this philosophy.

Probability Distribution:-

•Let us take two coins, now the sample spaces are:-

{HT, TH, HH, TT}

X is defined as a random variable getting head.

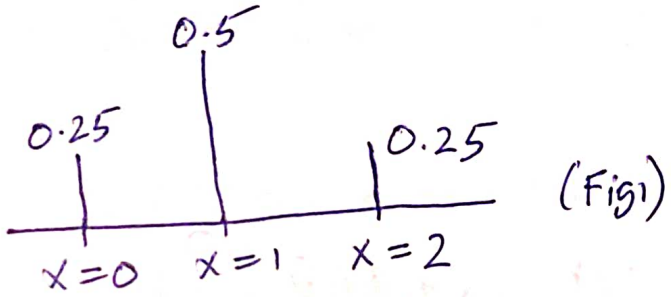
$P(X=0) = \frac{1}{4}$

$P(X=1) = \frac{2}{4} = \frac{1}{2}$

$P(X=2) = \frac{1}{4}$

$P(X=3) = 0$

Let us plot these values in a graph,

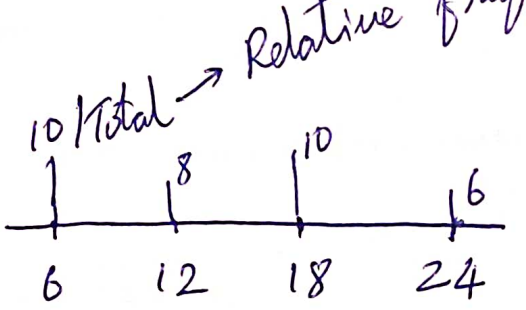


(Fig1)

This is the probability distribution

Let us take an example,

consider the number of months in the x-axis and number of people in the y axis.



(All these 10, 8, 10, 6 are frequencies)

(Fig2)

Your probability is nothing but the relative frequency.

Few observations about the above plots :-

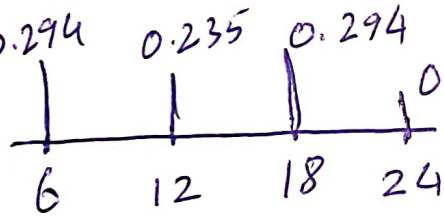
- Its nothing more than a histogram (with relative frequency).
- The x-axis is discrete.

It is not continuous, because we cannot calculate what is the probability of 1.5 head or 0.8 head -

- So, it is an example of discrete distribution.

How can we calculate the average experience in Fig2 :-

Let us replace the vertical lines with relative frequencies,



The average experience will be,

$6 \times 0.294 + 12 \times 0.235 + 18 \times 0.294 + 24 \times 0.176 = 14.1$

The average will be $\boxed{\Sigma x.P(x)}$

This is the global representation for any distribution.

- For continuous distribution we replace Σ with $\int$ symbol.
- $\int$ is the extended version of 'sum'.
- $\boxed{\text{Variance} = \Sigma (x-\bar{x})^2 . P(x)}$

Continuous, we do Integration.
Discrete, we do summation.

2.8 7, 8, 9, 10, 12, 13 10

$$\text{Variance} = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

$$\sigma = \sqrt{\Sigma (x_i - \bar{x})^2 / n}$$

$$\Sigma (x_i - \bar{x})^2 P(x)$$

$$(7-10)^2 P(x) + (8-10)^2 P(x)$$

$$\frac{(7-10)^2}{n} + \frac{(8-10)^2}{n}$$

Binomial Distribution & Poisson Distribution { Distributions - Descrete Distributions.

- Assume there is a T-shirt manufacturing company which is operating for last 5 years, and manufactured 2,00,000 T-shirts. The QA always finds the defective pieces, they observed,

① 5% of the T shirts are defective.

② Out of the defective T shirts, (10,000), they are counting the number of defects in each T shirt.

D → 3
D → 2
D → 4
D → 5
.
.
.
10,000 D → 4

} How can we communicate these defects.

$\frac{\text{Sum of Defects}}{10,000}$ will give the average number of defects.

The mean of the binomial distribution is given as np.

There is a function in python, pmf → probability mass function

stats.binom.pmf(x, n, p) { Binomial Distribution }

(This function is used to obtain the probability mass function for a certain value of x, n, p)

stats.binom.cdf will add everything (cummulative density function)

- Mean of the binomial distribution is np and the variance of the binomial distribution is npq.

- Formula for The Binomial Distribution is :- (we have another function stats.binom.cdf (cummulative distribution function))

$$P(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

where np(mean)

stats.binom.pmf(x, n, p)

- Formula for The Poisson Distribution :-

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where λ(mean)

stats.poisson.pmf(n, λ)

$\lambda \rightarrow$ The average number of events that happen in a period Δt.

(mean of the binomial distribution)

- Mean = np (where n is the number of events and p is the probability)

$\mu = np$

- The value you expect to get in a statistical experiment is the mean. It is the highest distributed value in the distribution

- Variance = npq where $\boxed{q = 1 - p}$

$\sigma^2 = npq$

- Standard Deviation $\sigma = \sqrt{npq}$

· If there are categorical variables then it is binomial distribution.

Defect / no defect → categorical column

men / women → categorical column

- Whenever we have a categorical information, always we communicate this information interms of proportion (or) percentage.

Let us look at three statements :-

① In a toll gate, during 8AM-10:30AM, average 78 cars are passing. (counting)
↘ numerical quantity (discrete)

② A customer care unit receiving on an average 150 calls. (counting)
↘ numerical quantity (discrete)

③ A bank data tells that 15% of customers default their loan EMI.
Default or Not Default (Categorical Variable)

All the above examples come under categorical variables with numeric information that is discrete.

- Whenever we have the percentage information for categorical variables, we can generate binomial distributions.
- For counting statistics (counting info) we can generate Poisson distribution.

- Whenever we express a categorical variable, we express interms of proportions (or) percentages. (Binomial)
- Whenever we address a numerical quantity (Continuous) then continuous distribution (normal distribution).
- If we are dealing with numerical quantity (discrete) then poisson distribution.

- In Poisson distribution mean is λ

General Rule:-

$\Sigma x \cdot p(x)$ (mean)

$\Sigma (x-\bar{x})^2 \cdot p(x)$ (Variance)

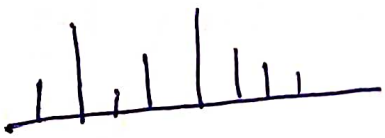
$\sqrt{\Sigma (x-\bar{x})^2 \cdot p(x)}$ (Standard deviation).

	Binomial	Poisson
mean	np	λ
S.D	$\sqrt{npq}$	$\sqrt{\lambda}$

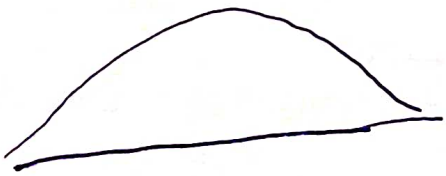
p → probability of the occurance of event happening.

n → sample size

λ → mean



Discrete Distribution



Continuous (normal) distribution.

- If we want to add multiple pmf's we can go for cdf's.
- Poisson Distribution is independent of n. (meaning mean doesn't effect the Poisson Distribution.

- Bernoulli → collection of two Binomials.

- Whenever discrete, we use the terminology mass, whenever it is continuous, we use the terminology density.
- Only in the normal distribution, we use the word standard normal distribution.
- In discrete distribution, we just talk about the counts. (no units mentioned)
- In continuous distribution, units are mentioned. (every variable has its own units).

- The characteristic of standard Normal Distribution is :-

$\bar{X} = 0$ } The mean of the variable becomes 0, SD becomes 1.
$SD = 1$ }

Inverse normal function :-

stats.norm.isf function is used in python.
It gives the cut off values (range) for the required area.

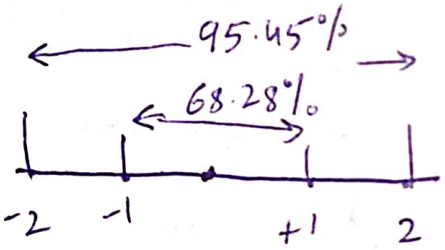
Central Limit Theorem is the foundation for inferential statistics.

We need to understand the population parameter in the inferential statistics.

- Standard Scalar, Z scale, Z-score (same)

In stats.norm.cdf function (_, μ, σ).

Isf (Inverse survival Function)



stats.norm.isf(0.025)

we use this if we want to calculate 95% of the area. Then 5% will be distributed as 2.5% on the left and right each. Give this value in the stats.norm.isf function.

95%	-1.96 to +1.96
90%	-1.64 to +1.64
99%	-2.58 to +2.58

Objective:- To estimate the population μ.

CLT will give the following rules in order to achieve the objective. (Estimate the population mean or proportion (categories))

1. Takes the samples randomly from everywhere.
- Take the mean values.
- Do one more round and take the mean values.
- Do for a multiple times (100 trials).
- So we got total 100 means.
- If we plot the distributions of 100 means, these 100 means will form the normal distribution.
- It will form the normal distribution irrespective of the population distribution.

$\bar{x} - 1SD$ to $\bar{x} + 1SD$ } 68% CI
(68% of data will lie within this range) 32% Risk

$\bar{x} - 2SD$ to $\bar{x} + 2SD$ } 95% CI
(95% of data) 5% Risk

$\bar{x} - 3SD$ To $\bar{x} + 3SD$ } 99% CI
(99% of data will lie within this range) 1% Risk

SD → Standard Deviation
CI → Confidence Interval
and Risk is the Error Risk.

Wider the Window
more Confident.

Narrow the window
less Confident.

· If there is no Inferential Statistics, there is no ML, then no DL.

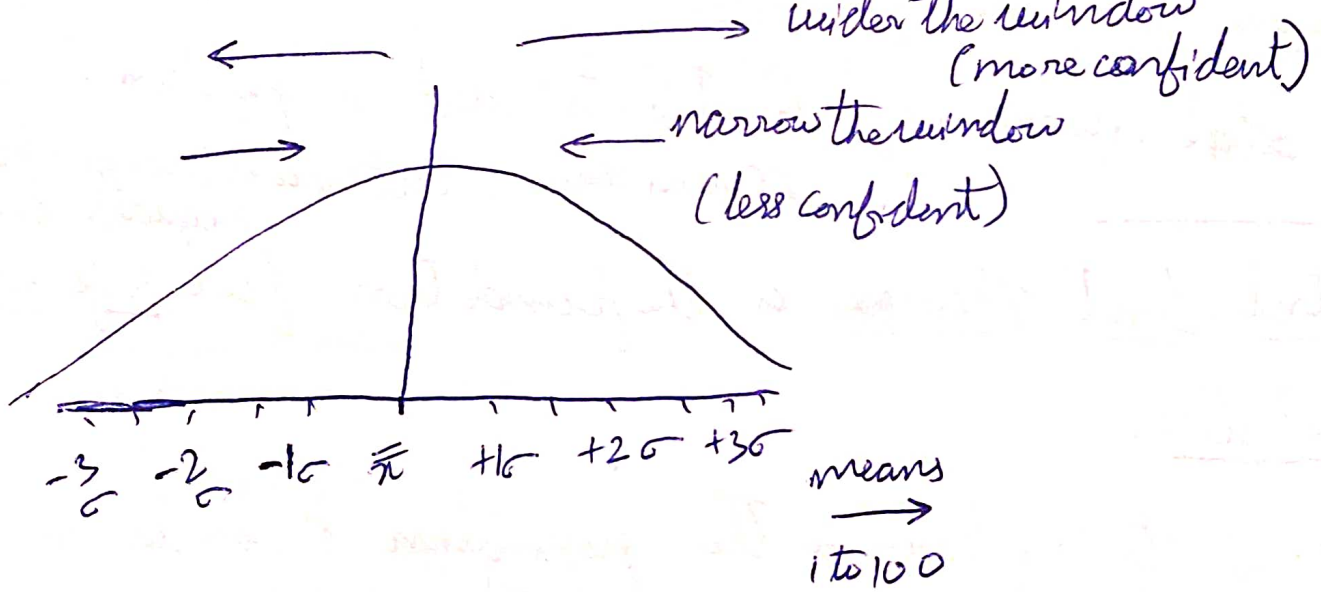
· CLT says that, if we plot the distribution of 100 means, these 100 means will form a normal distribution. irrespective of your population distribution. (what ever the shape is).

· The mean of the 100 means is $\bar{\bar{x}}$.

· Standard Deviation = 0 for the true mean

· SD = standard error of the mean

Each and every mean deviate with a certain amount from $\bar{\bar{x}}$



Examples for the type of distribution:-

- It is expected 20% of covid19 cases in India effecting children.
  Binomial
- It is also observed the avg age of covid affected is 38.45 years
  Continuous (Normal Distribution)
- The avg number of symptoms for the +ve cases is found to be '6'.
  Poisson

- From the before objective:

To collect the information of 50,000 Data Scientists, } 50,000

Take Average :- μ → This mean is known as true mean.

- True mean is unknown.
- We can only infer(or) predict the true mean.
- Point estimate is not at all possible in statistics.
- We can do a Range Estimate. (CLT is used).
- Objective is to Estimate the population mean (or) proportion if there are categorical variables.

Example :-

Assume $n = 40$, true mean is 24.6 years.

Let us say the range estimate is :-

[18 years to 50 years] Wider Interval, more confident.

- -0.5 to +0.5 is the acceptable range for skewness.

Normal Distribution

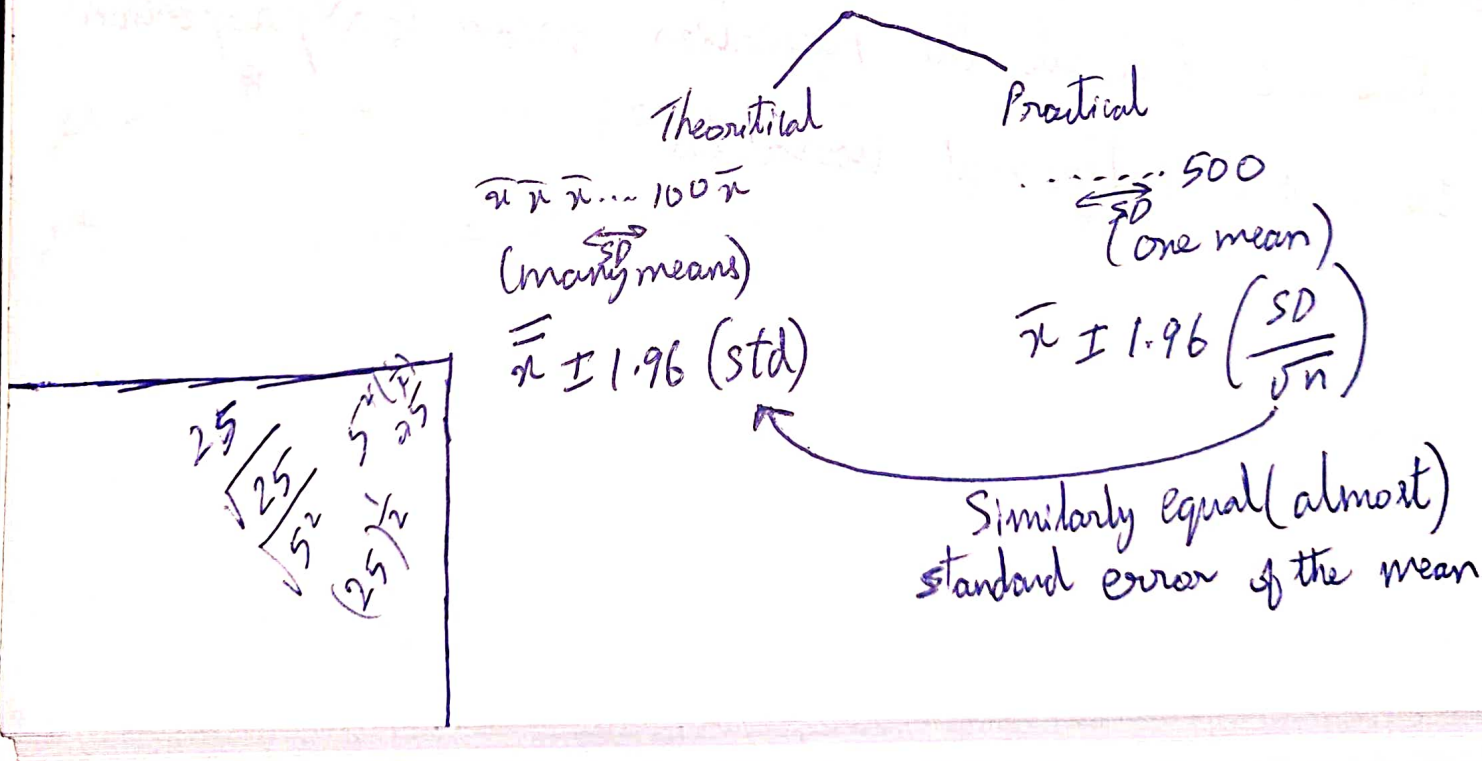
Area Under the normal Distribution

$\mu - \sigma \leq x \leq \mu + \sigma \rightarrow 0.6828$

$\mu - 2\sigma \leq x \leq \mu - 2\sigma \rightarrow 0.9545$

$\mu - 3\sigma \leq x \leq \mu - 3\sigma \rightarrow 0.9974$

- As we increase the sample size the error will reduce as the sample size will almost be same as population size.
- The CLT states that, if we take a number of means from different samples and plot them, they will be normally distributed, irrespective of the population distribution.
- With np.random.seed() same random numbers will be generated.



Lab Session 1 :-

• Poisson Distribution $\Rightarrow P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ where λ is the average number of events that happen in a period Δt.
Events in each of these Δt's are independent of each other.

• Binomial and Poisson is for Discrete Probability.

• Normal Distribution is for Continuous Probability. (distribution takes shape)

• Normal Distribution is a symmetrical bell curve, whenever the data is put together, it takes the shape of a symmetrical bell curve, where mean, median and mode lie at the same point. (which is the center of the curve).

For normal distribution, we perform stats.norm.cdf(x, μ, σ) where μ is the mean and σ is the standard deviation.

• Sampling Error :- Difference b/w Population mean and Sample mean.

• Standard Error :- σ (SD) $/ \sqrt{n}$ (sample size)
Standard Deviation (σ) of the sample divided by root over the n.

• stats.poisson.rvs is used to generate random variables.

• stats.poisson.rvs(loc, mu, size).
loc is the lowest x value, mu is the middle of distribution.

• np.random.seed() makes the random variables predictable.

• Confidence interval is sample mean ± (margin of error).

Statistics (Session 3)

1:18:40

• Range Estimate leads to one sample test and other tests.

• With skewness funtion, we can generate a skewed distribution.

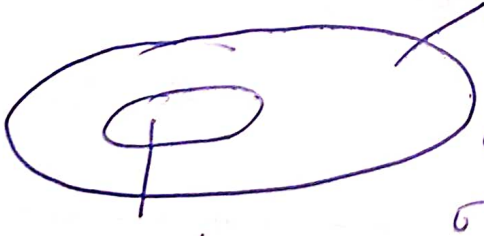
• In inferential statistics we are going to infer whether our sample is a representation of population. (we have to verify this with hypothesis testing)

• In inferential statistics, we can only infer about μ (or) proportion.

• In theoritical approach we take many trials to calculate mean of the mean.

• In practical approach we do only one round of sampling.

•



Population

μ → mean (True mean)

σ → SD

σ^2 → Variance

N → Sample size

Sample

$\bar{x}$ (mean)

s → standard Deviation

n → sample size

• Z critical value, margin of error value

$$\text{margin of error} = Z_{\text{critical value}} \times \frac{\text{Population SD}}{\sqrt{\text{Sample Size}(n)}}$$

Z critical value → we can find using isf(or) ppf anyone of them.

stats.norm.ppf(q), pass the q value where $q = 1 - p$

• Example:- The interval from 2.634 to 2.966 forms a 95% confidence interval for μ. In other words, we are 95% confident that the average delivery time lies between 2.634 and 2.966 hours.

• In both (theoritical and practical) the range contains the true mean.

• Tolerance level of skewness is ± 0.5.

- Null Hypothesis → No effect Hypothesis (H_0)
- Alternative Hypothesis → effect hypothesis (H_a)

- whenever we are checking equal or not equal, the equality goes to null hypothesis. (H_0)
- Something not equal to, will go for affected hypothesis. (H_a).
- If it is a one sided check, then it is a one tail sample.

- How the sample mean $\bar{x}$ is away from hypothesized mean μ. This formula $\Rightarrow \frac{\bar{x} - \mu}{s/\sqrt{n}}$. (This is generally used for 1 sample t test)

(using this formulas, we get the SDs)

(It is going to tell how many standard deviations the μ is away from $\bar{x}$).

- $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (This formula is used for two sample t Test).

Independent groups

We can find by how many SDs away with these formulas. (t_{stat})

- $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ (One sample t test)

- Confidence interval is also known as acceptance zone for H_0.

Let us take an example:- Swiggy claims that it can deliver the food within 50min,

$\mu_{avg} < 50min \rightarrow H_a$

But, the agencies, find out it is more than 50min.

$\mu_{avg} \geq 50min \rightarrow H_o$.

(The above is H_a because the delivery speed is effecting.

One Sample Mean Test

- The avg delivery time of hyderabad and chennai is equal and not equal.

$\mu_{avg(Hyd)} = \mu_{avg(Ch)}$

$\hookrightarrow H_o$

$\mu_{avg(hyd)} \neq \mu_{avg(ch)}$

$\hookrightarrow H_a$

Two Sample Mean Test

- $s/\sqrt{n}$ is the standard error of the mean

- ~~The minimal probability required for the null hypothesis being true is~~
- Null hypothesis being true is less than 5%
- whenever $P < 5\%$, we reject H_0.
- H_0 becomes true once it is greater than 5%.
-

To reject H_0

	t stat (dist. b/w two means)	P-Value (area)
95%	> 1.96	< 0.05
99%	> 2.58	< 0.01
90%	> 1.64	< 0.1

- A hypothesis is a testable statement made in support of a finding or a claim about something in the world around you.
- It should be capable of being tested, either by experiment or observation.
- Hypothesis testing is an objective scientific method that is used in making statistical decisions using observational or experimental data.
- Hypothesis testing is an assumption that we make about the population parameter, a premise or claim we want to test.

Two Tail Vs Two Sample

One Tail Vs One Sample

These terminologies are completely different.

Sample

One tail Test

Two Tail

left Tail

Right Tail

Two sample Test

Dependent Independent

$n = 500$

$\bar{x} = 7,10,000$

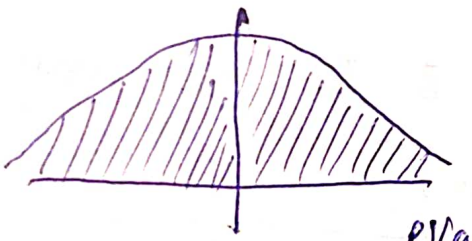
SD

Let us assume

$\mu = 7,10,000$

both are same.

$$\text{Stat} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 0.$$



7,10,000

pValue = 100%.

Only for one sample T Test or proportion test we have:

Two tail, left tail and right tail.

Two Tailed One-Sample 't' test :-

· Company claims they fill their cans with 300ml. We need to verify whether the expected population mean is 300ml.

First we need to do hypothesis,

We need to frame null hypothesis Ho and alternate hypothesis Ha.

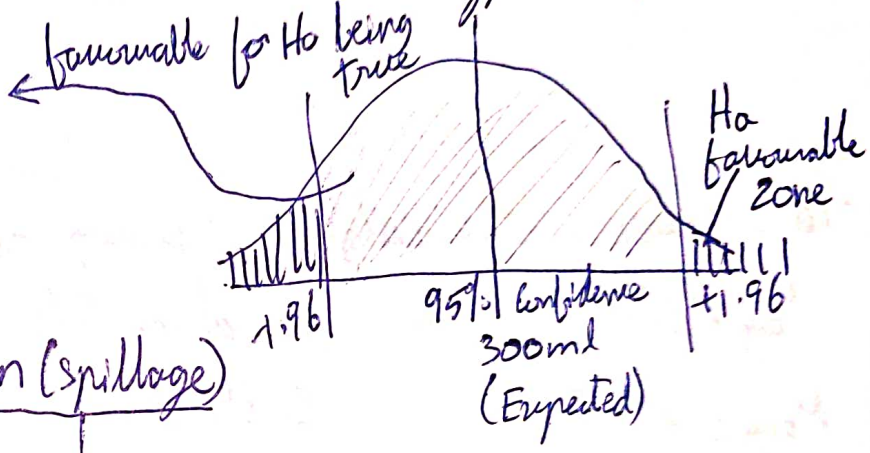
$H_o = \mu_{Pop} = 300ml$

$H_a = \mu_{Pop} \neq 300ml$

↳ what are the effects

→ They can overfill the can (spillage)

→ Under filling.



· When the value falls within the range, where null hypothesis is true, outside this range it is not true. When the value falls within the range probability percentage increase by 5%. when outside it is less than 5%.

$t_{stat} = 0$ (if sample $\bar{x}$ is 300ml) then the Pvalue is 100%. Meaning it correctly represents the μ. Null hypothesis is true (100%)

n=400 (Randomly Taken 400 samples)

$\bar{x} = 298.56$

$s = 1.5$

} Information given.

In jupyter notebook, we got,

$\bar{x} = 298.51$

$s = 1.54$

$t_{stat} = -19.18$

(Let us draw a diagram in the next page).

General formula for SD is $\sigma_{pop} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

- For sample data, SD is, $S_{samp} \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

- The numpy function np.std by default it divides by n.
- ddof = 1 means, the denominator is n-1.

ddof (delta degrees of freedom)

---

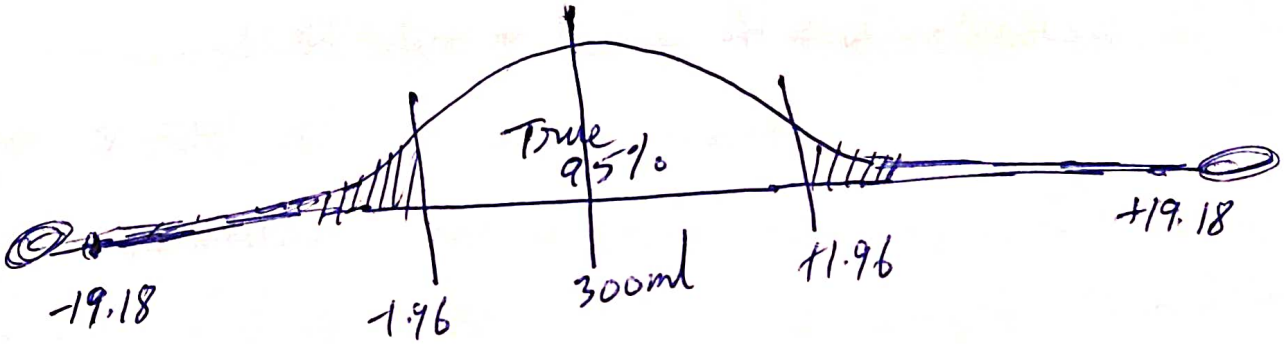
ttest_1samp (samp_array, 300) will give us the

t-stat score, p-value

---

df['vol'].std() by default divides by n-1.

np.std[] by default divides by n.



only 0.5 data lyes beyond 30.

The area will be obviously less than 5%.

Since the P value is $P < 5\%$, we reject the null hypotheses.

Letus do the one tail test.

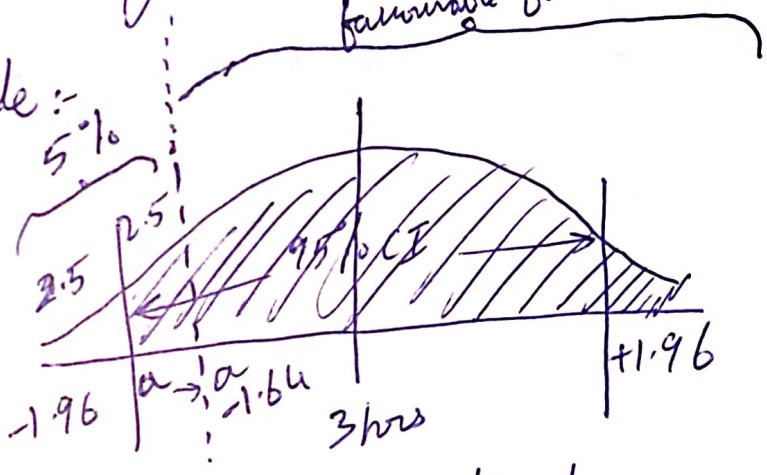
- From the Newyork Courier example :-

$\mu_{dtime} < 3hrs \rightarrow Ha$

$\mu_{dtime} \geq 3hrs \rightarrow Ho$

(This is left tail example)

(Right tail is similar)



This 95% will represent 3hrs

The whole shaded part is $95 + 2.5 = 97.5\%$. we need to move α to the right by 2.5%. Since it is a 1 tail test 5% will be on the left side. we use isf function to get -1.64.

- Every value has its own Confidence interval.
- ttest is the hypothesis testing formula.

- If null hypothesis is being rejected, then Ha is true.

- Let us Consider, a JEE college is claiming average marks score as :-

$\mu_{avg} > 98.5$ ... we always take equality sign as an advantage

$\mu_{avg} \leq 98.5$ to the H_0.

- we should always think what is H_a and H_0.

H_a always is ~~supportive~~ supportive and tells about the efficiency.

If H_a is supportive

$H_a >$ ___ (Right tailed)

$H_a <$ ___ (left tailed)

Two Sample T Test :-

Let us understand the type of testing in each statement :-

- Avg age of DSE at GL is > 25 years.

(Here the population is entire India and age is the variable. This is the test of mean. This is one sample because we are considering the entire country. This is right tail test.

Here, Ha is $\mu > 25$ years

Ho is $\mu \leq 25$ years)

- Average age of DSE student at Bangalore is more compared to Hyderabad.

It becomes two sample mean test.

Bbr: $n_1 = 1000$, $\bar{x}_1 = 26.5$, $s_1 = 2.3$

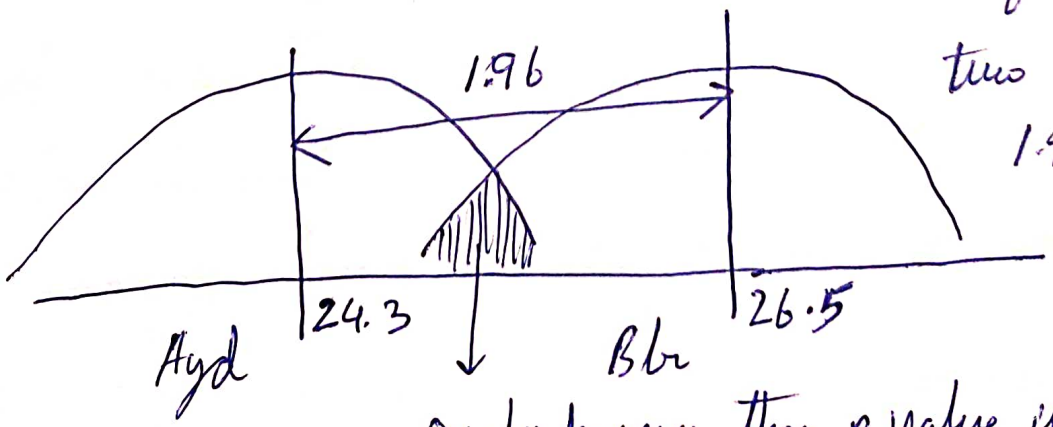
Hyd: $n_2 = 800$, $\bar{x}_2 = 24.3$, $s_2 = 1.8$

Here,

$H_o = \overline{age}_{Hyd} = \overline{age}_{Bbr}$

$H_a = \quad \neq$

If Ha must be true two means must be 1.96 SDs away.



Overlap happens, then p value increases.

Overlap doesn't actually happen for 1.96 standard deviations.

If we calculate, the two means are significantly different.
The t.stat score we got is 22.7640.

- There are two types of Two Sample Tests :-
  - → Independent
  - → Dependent

Independent (unpaired)
t_test_ind(g_1, g_2)

Dependent (paired)
t_test_rel(g_1, g_2)
(Here rel means relative)

g_1 and g_2 are two group arrays. Just upload the generated sample arrays into the function.

- Let us understand about Dependent (Relative groups)
- → Consider $n = 100$ people are taken randomly.
- → The number of antibodies (mean) before vaccination is μ_{ab} before vaccination.
- → After Vaccination it will be μ_{ab} after vaccination.

The Company claims that the antibodies will increase after vaccination.

This is Dependent type.

H_0 : μ_{ab} before vaccine $=$ μ_{ab} after vaccine.

H_a : $\neq$

Let us check about the independent :-

- Randomly we are giving covaxin to 15000 Indians.
- Randomly we have taken 10,000 African samples covaxin.
- After generating the covaxin shot, the average num of antibodies in the Indians and Africans is μ_{In}, μ_{Af}.

H_0 : $\mu_{In} = \mu_{Af}$ (Both are same)

H_a : $\neq$ (Both have a difference).

This is the example of an independent test.

Lab Session :-

- Our research question is always an alternate hypothesis whereas the null hypothesis is the status quo. Null hypothesis is the -ve version of our research question. Alternate hypothesis is what we are trying to question.
- There are Type1 and Type2 errors.

Type1 error :- denoted by α, rejecting the null hypothesis incorrectly.

Type2 error :- Incorrectly fail to reject the null hypothesis.

which means that we should have rejected the null hypothesis but we failed to reject the null hypothesis. Denoted by β.

- When the population standard deviation is not known to us, then we use T-test.
- If it is known we go for Z Test.

Session 5 :-

Summarizing the key statistical test :-

It is useful in machine learning model building.

Two cross two Quadrant :-

Whenever we talk about population parameter, either we check mean (or) proportion. When it is a continuous distribution we check whether the means are statistically equal (or) different.

- Whenever we are calculating mean for two samples (or) groups, we call it Two Sample 't' Test. (or) Two Sample Mean Test.
- For more than two groups we call it ANOVA.
- t test and z test are same for $n = 500$.

t-test $\Rightarrow \frac{\bar{x} - \mu}{s/\sqrt{n}}$ To the content of one sample,

z-test $\Rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

- According to statistics, when $n = 30$ and above, $s \simeq \sigma$ and

$t_{stat} = Z_{stat}$.

- In paired groups we have before and after effects.

For eg:- Before vaccination and after vaccination.

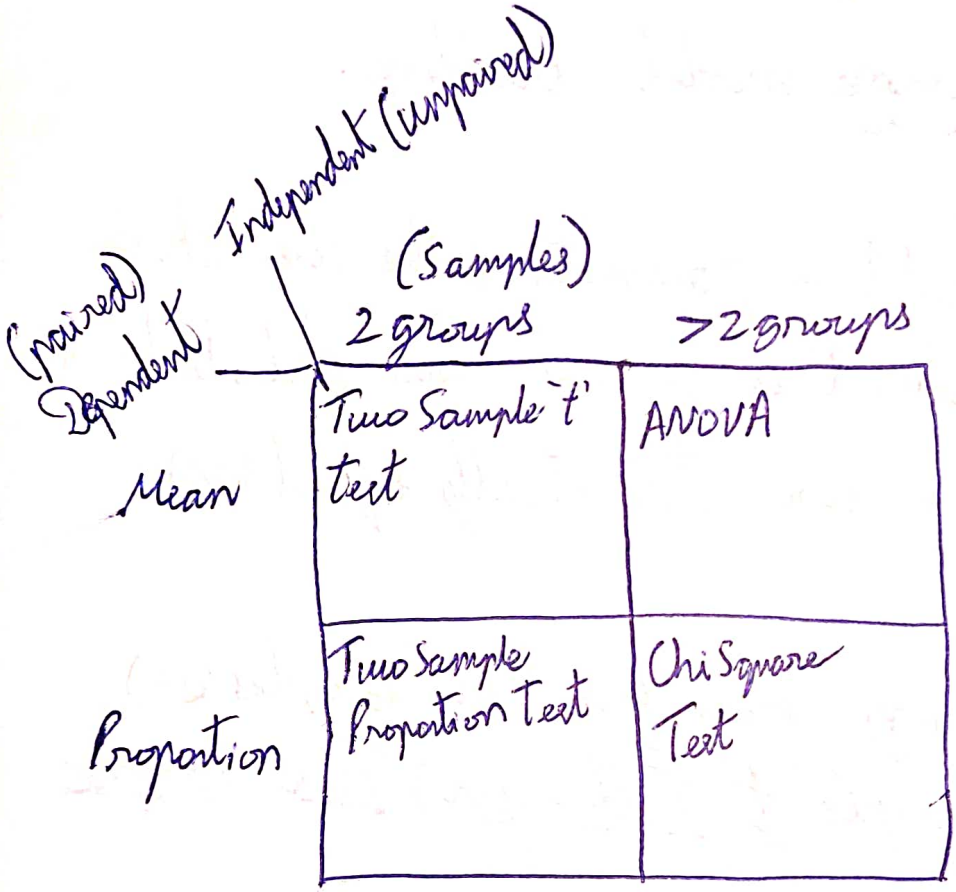
we are going to work with the same samples.

- Let us consider we have (Before $n = 50$) and (After $n = 50$) as g_1 and g_2.

First we will calculate $g_1 - g_2$ for all the 50 records. Each and every record you take the difference. ΣD will be the sum of total differences.

Another, column, take a squared difference $(g_1 - g_2)^2$. ΣD^2 will be the sum of squared differences.

- Unlike independent events n_1 and n_2, we have only one n in dependent groups.


Independent (unpaired)
(Samples)
(paired)
Dependent
2 groups
>2 groups
Mean
Two Sample 't' test
ANOVA
Proportion
Two Sample Proportion Test
Chi Square Test

• tstat for paired test ⇒ $\dfrac{\Sigma D/n}{\sqrt{\dfrac{\Sigma D^2 - \frac{\Sigma D^2}{n}}{(n-1)n}}}$.

tstat > 1.96 ⇒ Pvalue < 5% ⇒ Reject Ho

• Y is dependent variable.

In prediction model, we have combination of dependent and independent variables.

$Y = f[x]$ — airbags, bootspace, colour } Independent variables.

Dependent

The price of car is continuous numerical variable.

Regression Analysis.

• $Y = f[x]$ is y as a function of x.

• This kind of prediction model, we call it regression model.

• The outcome whatever we are predicting is continuous numerical model.

• ANOVA will form a foundation for regression.

· $Y = f[x]$ (age, bmi etc).

↓ Blood glucose level

[Instead of telling the blood glucose level, the model will tell Diabetic / Healthy.

- Dependent variable Y will have only categorical column which is Diabetic / Healthy.
- Model will predict whether diabetic / healthy. (Classification model).

Assume an example, our dependent variable 'y' is categorized into H (Healthy) → 0 and C (Cancer) → 1.

The independent variables are X1 (Age), X_2 (Gender), X_3 (Drinking)

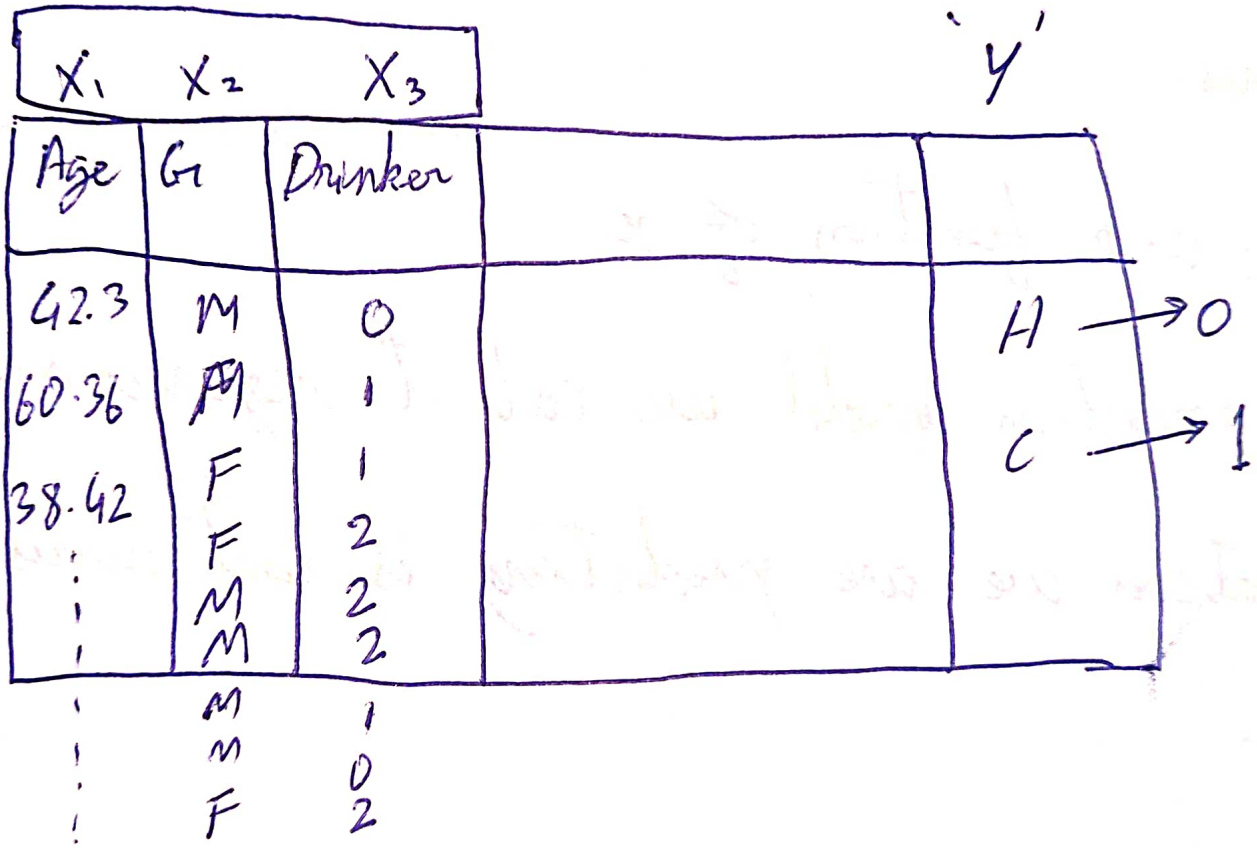
Assume

Age has continuous values → 42.3, 60.36, 38.42. (X_1).

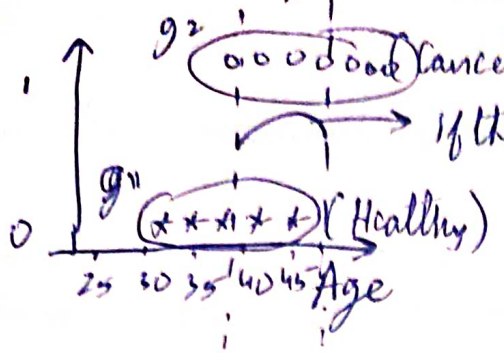
Gender has categories like M and F (X_2).

Drinking has categories like 0 (No drinking), 1 (Occasionally) and 2 (Severly).

We need to plot the independent and dependent variables now.


X1
X2
X3
'y'
Age
Gr
Drinker
42.3
60.36
38.42
M
F
F
M
M
M
M
F
0
1
1
2
2
2
1
0
2
A → 0
C → 1

Let us consider Dependent variable Y as on y axis.



If the dif btw mean is atleast >1.96 SD away.

(There is a lot of overlap in ~~normal~~ distribution)

(the means)

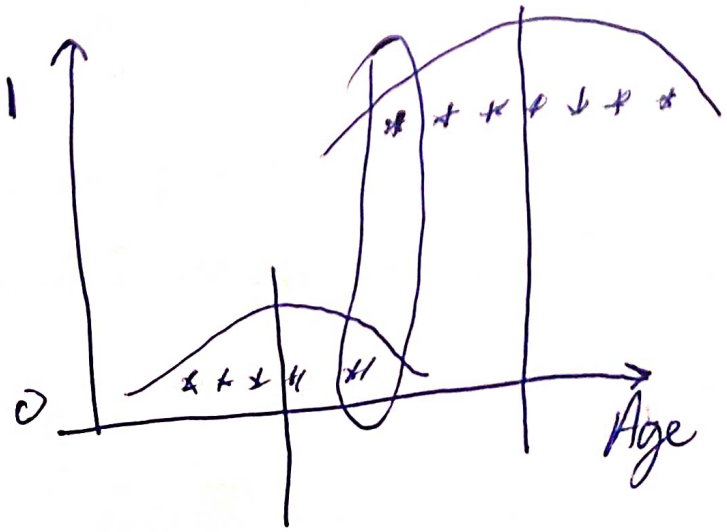
If the means difference is above 1.96, they are statiscally different.

If the total score is 1.12, then there is no relation between the age and cancer.

If the means are within 1.12 standard deviations, there is high Pvalue $P > 5\%$, null hypothesis holds good.

$H_0 : \overline{age}_0 = \overline{age}_1$

$H_a : \neq$



In this graph there is very less overlap.

$P < 5\%$

we need to understand between X_1 and Y what statistical test is performed, two sample T test is done to understand whether X_1 is the potential variable in predicting the disease or not.

- whenever we get $P > 5\%$, we will not use the feature in the model.
- If at all we need to use the feature, $P < 5\%$ which means $P < 0.05$, they are statistically significant, we can use in the prediction model.

Now let us consider with X_2 (Gender). We will use the cross tab function now,

pd.crosstab(df[Gi], df['Y']).

	H	C
100 M	50	50
80 F	40	40

$H_0 : P_{MC} = P_{FC}$

(Proportion of male having cancer is equal to proportion of female having cancer).

$H_a : \neq$

will the gender effect cancer?

Answer :- No

$\frac{50}{100} = 50\%$, $\frac{40}{80} = 50\%$

• Let us consider,

		C	
n_1 100M	45	55	x_1
n_2 80F	30	50	x_2

then $P_1 = 0.55$

$P_2 = 0.63$

P_{pooled} is total proportion of cancer irrespective of the gender.

$$P_{Pooled} = \frac{105}{180}$$

• Proportion_ztest formula is $$\frac{P_1 - P_2}{\sqrt{P_{Pooled}(1-P_{Pooled}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

we have a python function,

Proportions_ztest $\left([x_1, x_2], [n_1, n_2]\right)$.

• A travel agency wanted to study the effect of holiday choice among men and women

Independent	Dependent: Beach	Cruise
men	209	280
women	225	248

Travel agency wanted to study whether the gender effects preferred holiday.

• we can formulate the hypotheses in two ways, to the content of beach (or) to the content of cruise. If any one side of the story is satisfied, automatically the other side is satisfied.

We will formulate hypotheses w.r.t beach first.

$H_0 : P_{MB} = P_{WB}$

$H_a : \neq$

Then, w.r.t @ cruise holiday,

$H_0 : P_{MC} = P_{WC}$

$H_a : \neq$

Let us solve w.r.t beach first,

$P_1 = \frac{209}{489}$ $P_2 = \frac{225}{473}$ $n_1 = 489$ $n_2 = 473$

$P_{pooled} = \frac{434}{(489+473)}$

In the interview, they are two contents, t-stats and z-stats,

For continuous variable we say,

tstat (s) In real life we calculate $n \geq 30$, when n

Zstat (σ) approaches 500, $n = 500$, then

tstat = Zstat.

- Two sample proportion test, popularly called as Zstat (or)

Z data. Proportion_ztest (two sample proportion test).

• Let us look at a problem statement :-

The number of Covid 19 positive cases in India which is effected for women, expertion is 60%

$P_{prop\ C19\ W} = 60\% \rightarrow H_0 \rightarrow$ Zstat = 0 for this

$\neq 60\% \rightarrow H_a$ P-value = 100%.

To verify this we randomly collect 1500 samples, Male are 650 and Female are 850.

What proportion of females are effected :- 850/1500, $P_f = 56\%$.

Sample proportion is 56% but expected is 60%

This above problem comes under <u>one-sample proportion test</u>.

Proportion_ztest ([x_1, x_2], [n_1, n_2])

Whether it is one sample or two sample proportion test we have common function.

We have only one set of information here, then the arguments are:-

Prop_ztest (850, 1500, 0.60)

1:09:31

Chi Square Test

one sample mean test

two sample mean test

one sample proportion test

two sample proportion test

Test of Mean :-

- One-sample mean Test
  - ttest_one samp(samp, μ)
- Two sample mean test
  - paired (dependent)
    - ttest_rel(g_1, g_2)
  - unpaired (independent)
    - ttest_ind(g_1, g_2)

Chi square Test for Goodness of fit of multinomial data :-

$\chi^2_{data} = \sum \frac{(O-E)^2}{E}$ E is expected, O is observed

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

	Healthy	Mild	Severe
Male	10	5	5
Female	12	6	6

← This is observed count

- 50% of male and 50% of female are healthy.
- 25% of .. .. 25% . . . . . mild
- 25% of .. .. 25% . . . . . Severe.

Similar pattern is observed between the male and female. then only the null hypothesis holds true

Let us change the observations :-

(Expected count is male proportion should match the female proportion.

How to arrive the count logically.

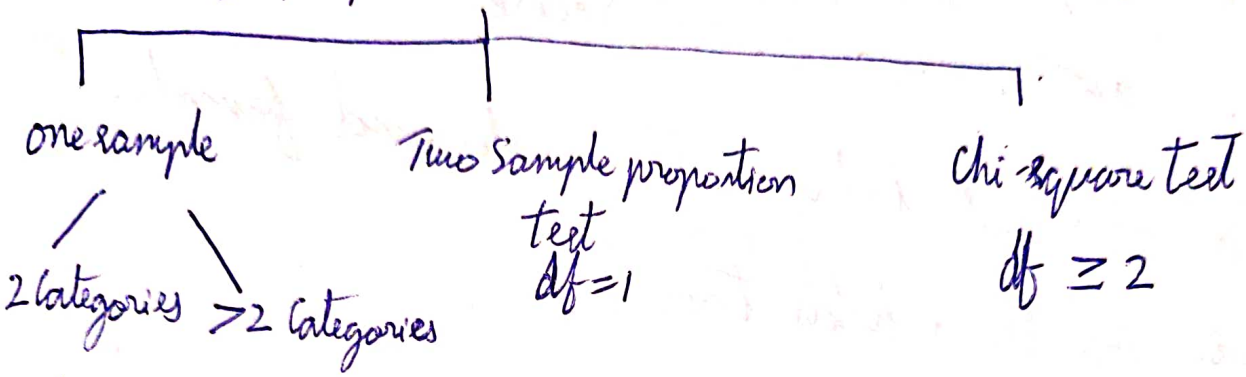
		H	M	S
20	M	8 (40%)	8 (40%)	4 (20%)
24	F	12	6	6
		50%	25%	25%

Male healthy expected count is $\frac{(\text{Total male}) \times (\text{total healthy people})}{(\text{Total male and female})}$ (20, 20, 44)

$$\Rightarrow \frac{20 \times 20}{44} = 9.09$$

We have the four flavours of proportion test :-

→ Test of_proportions



- Chi square :- we have two groups with degrees of freedom 2 and above
  Chi2-Contingency() is used.
- Two Sample proportion :- prop_ztest([x_1, x_2], [n_1, n_2])
- One Sample (2 Categories) :- prop-ztest(x, n, expected proportion)
- One Sample (>2 Categories) :- Chi square([ ], [ ])
  we should put either numerical list or numpy array

Similarly, the expected count table will be,

9.1	6.3	4.5
10.9	7.63	5.45

Then the proportion percentage will be

45.5%	31.5%	22.5%
45%	31.7%	22.7%

Let us draw the two tables once again:-

Observations Table

	Healthy	Mild	Severe
20 M	8 (40%)	8 (40%)	4 (20%)
24 F	12 (50%)	6 (25%)	6 (25%)

Expected Count

9.1 (45%)	6.3 (31.5%)	4.5 (22.5%)
10.9 (45%)	7.63 (31.7%)	5.45 (22.7%)

Chi-square value is defined as:-

$$\frac{[(\text{Observed Count}) - (\text{Expected Count})]^2}{\text{Expected Count}}$$

+ (for all the six entries).

$$\Sigma \frac{(8-9.1)^2}{9.1} + \frac{(8-6.3)^2}{6.3} + \frac{(4-4.5)^2}{4.5} + \cdots$$

- There is an effect if chi-square value is large.
- If there is correct proportion, then the observations match with the expected count.

Then the chisquare value becomes zero.
So the lower the chi-square value, it ends up with high p value, the null hypothesis holds good.
Large chi-square value, we end up with small p value, we reject the null hypothesis.
· The chi-square distribution is square of the standard normal distribution.

For Z score the cutoffpoints are

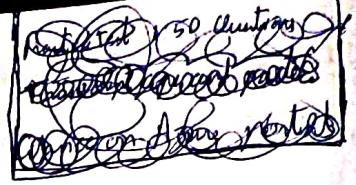
95%.	1.96
99%.	2.58
90%.	1.67

If p-value is less than 0.05 we reject Ho.

- Chi 2 - Contingency is the function for chi-square.
  Chi 2 - Contingency (CT) is the parameter where CT is the name of the table.
- post hoc analyses :- we are going to convey story about the Ha statement.

- If a dependent variable is continuous, it is a regression problem.
- If a dependent variable is categorical, it is a classification problem.

Great learning (Statistics Course) session 6 (ANOVA) :-



- ANOVA means we are checking the mean for more than two groups.
- If it is two groups, the degree of example is 1, we do two sample mean test.
- But in the example, the degree of freedom is 2 (no. of groups − 1), whenever we have two degree of freedom and we are checking the mean, it is an ANOVA test.
- The below example is age feature w.r.t three categories.

Sample ages for Groups A, B and C

Group A	Group B	Group C
30	25	25
40	30	30
50	50	40
60	55	45

Sample ages for Groups D, E and F

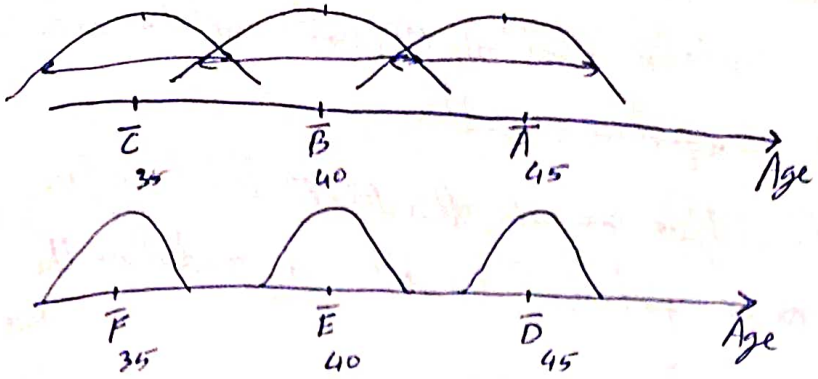
Group D	Group E	Group F
43	37	34
45	40	35
45	40	35
47	43	36

- The groups A, B, C has large variations in their values.
- The groups D, E and F has less variations in their values.
- $\bar{A} = 45$, $\bar{D} = 45$ } The means are same but the spread is large.
- $\bar{B} = 40$, $\bar{E} = 40$
- $\bar{C} = 35$, $\bar{F} = 35$
- Let us understand these values diagramatically.

$H_0: \bar{A} = \bar{B} = \bar{C}$

$H_a: \neq \neq$

Here we are checking $45 = 40 = 35$



The null hypothesis will be, $H_0: \bar{D} = \bar{E} = \bar{F}$

$H_a: \neq \neq$

Here also we are checking $45 = 40 = 35$

- In the first case H_0 holds true.
- In the second case H_a holds True.

- In the first one we can see many overlaps, due to overlaps, p value will be high ($P > 0.05$)

If P value is high, H_0 holds good.

In the second case, statistically, one confidence interval will not overlap other confidence interval. So, the p value will be near to zero.

$P \simeq 0$, so we reject the null hypothesis, alternate hypothesis H_a holds good.

- Whenever we get P value high, the feature is a useless feature.
- There is no significant difference in the mean (first case), hence we have useless features.
- In the second one, we have significant difference in the mean and the features are useful.
- Now we are going to prove this mathematically.
- Let us take an another scatterplot representation.

*In this example, we have taken continuous variable age as an independent variable. and the categories are SD(severe diabetes), MD(mild diabetes) and H(Healthy). Its something like a classification problem.

For classification problem, definitely we can do this ANOVA test to, infer whether we can infer age as an potential variable in predicting the category of the disease.

Example: Suppose we three different rice grain variety.

Rice1	14.3mm
Rice2	18.6mm
Rice3	22.6mm

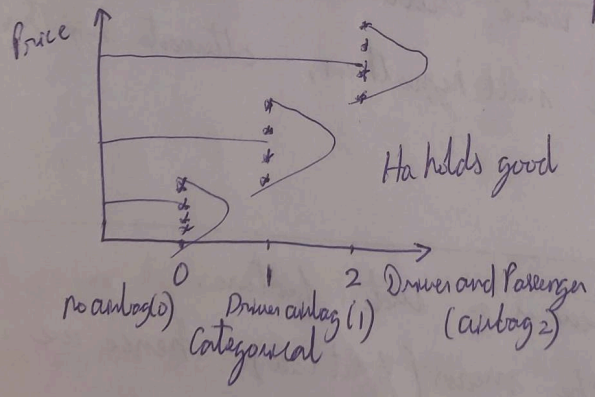
This is a classification problem

Different rice varieties have different shapes. (eccentricity)

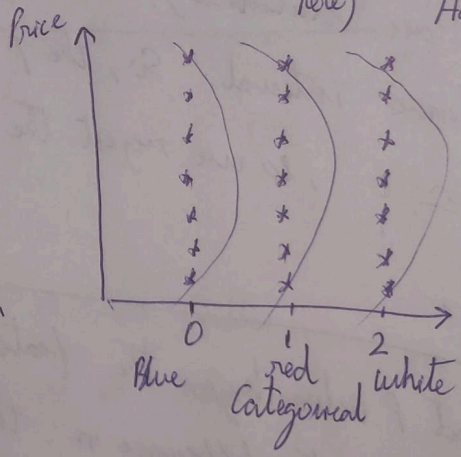
To predict the type of rice, whether the mm feature is useful or not,

If means are significantly different, we can use the feature to categorise.

Assume the price of the car.



(The means are very close here) Ho holds good



Price = f(carbags)

Price = f(Color of the car)

For ANOVA we have Fstat formula,

$$F_{stat} = \frac{MSTR \text{ (mean square treatment)}}{MSE \text{ (mean square error)}}$$

In this example, we have taken continuous variable age as an independent variable. and the categories are SD (severe diabetes), MD (mild diabetes) and H (Healthy). Its something like a classification problem.

For classification problem, definately we can do this ANOVA test to, infer whether we can infer age as an potential variable in predicting the category of the disease.

Example :- Suppose we three different rice grain variety.

Rice1	14.3mm
Rice2	18.6mm
Rice3	22.6mm

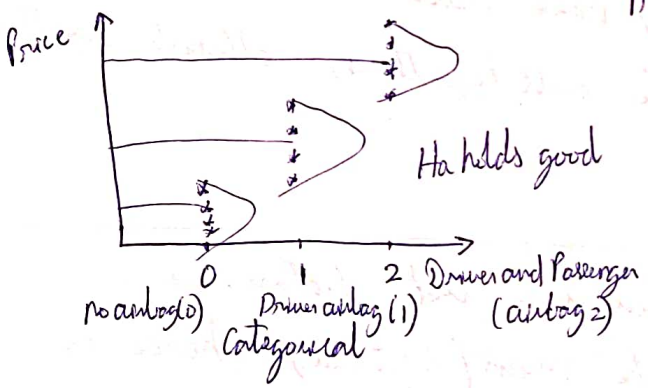
This is a classification problem

Different rice varieties have different shapes. (eccentricity)

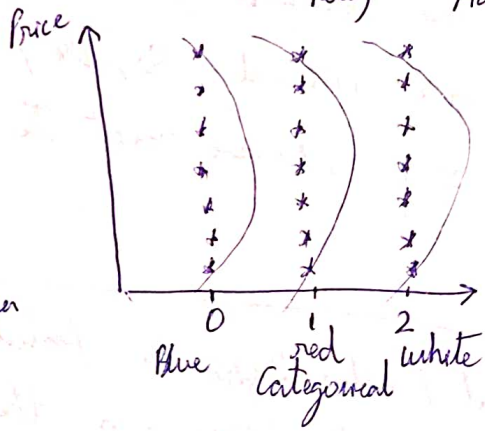
To predict the type of rice, whether the mean feature is useful or not,

If means are significantly different, we can use the feature to categorise.

Assume the price of the car.



(The means are very close here) Ho holds good



Price = f (airbags)

Price = f (Color of the car)

For ANOVA we have Fstat formula,

$$F_{stat} = \frac{MSTR \text{ (mean square treatment)}}{MSE \text{ (mean square error)}}$$

•MSTR is going to talk about between group variability. (Between samp variability)
•MSE means within group variability (within samp variability)

In order the feature to be significant, Ha holds good. P should be less than 0.05.

• If Ha holds good, Fstat should be high.

Fstat will be high when Numerator is large and denominator is less.

Higher the value, the means are far apart.

Smaller the values, the means are closer

• Whenever the Fstat is very high, the P value is very small and vice versa.

• Between sample variability : Finding the variance of K-sample means, weighted by sample size and expressed as Mean Square Treatment (MSTR).

$$MSTR = \frac{\sum n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

where n_i is the number of samples in each group. These
Numerator of MSTR is the sum of the squares treatment (SSTR) and
denominator term $K-1$ is the degree of freedom.

Here,
n_i → number of records in each group.
$\bar{x}_i$ → one group mean value.
$\bar{\bar{x}}$ → global mean

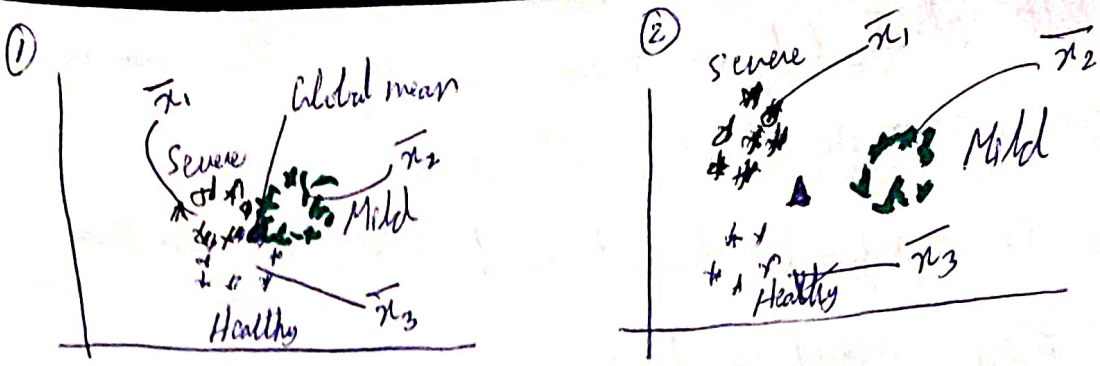
• Larger the MSTR value, good the separation is.

Substituting values into the formula,

$$\frac{4*(45-40)^2 + 4*(40-40)^2 + 4*(35-40)^2}{2} = \frac{100+100}{2} = 100$$

$MSTR = 100$

If we plot the records into a scatter plot,



$\bar{A} = \bar{D} = 45$

$\bar{B} = \bar{E} = 40$

$\bar{C} = \bar{F} = 35$

- For (2) MSTR will be large and for (1) it is small.

- Within sample variability :- Finding the weighted mean of sample variances, expressed as Mean Square Error (MSE).

$$MSE = \frac{\Sigma (n_i - 1) s_i^2}{n_t - k}$$

$$F_{data} = \frac{MSTR}{MSE}$$

- Numerator of MSE is the sum of the squares error (SSE) and denominator term $n_t - k$ is the degrees of freedom.
- The total number of squares (SST) is the sum of SSTR + SSE.
- ANOVA table is the convenient way to display all these parameters.

Here,

$n_i - 1$ is each group number of records -1

s_i is the variance of groups. (s_1^2, s_2^2 and s_3^2)

n_t means the total number of records

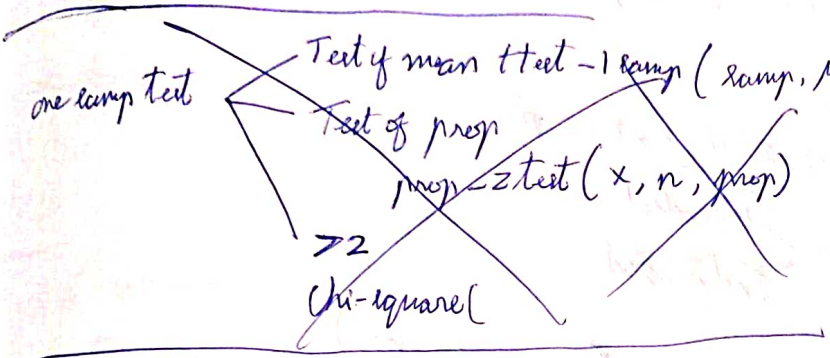
k is the total number of groups.

Whenever we reject H_0, we have some interesting story to tell from H_a.

That is the post hoc analysis.

In the reality :-

stat. Test	H_0 : True	H_0 : False	If no one knows there in the reality.
Reject H_0	This is error (X) Type I (or) α	whenever H_0 is false we reject H_0. Right decision ✓	
Accept H_0	whenever H_0 is true, we are accepting H_0 Right ✓	This is error (X) Type II or β	



~~Summary of learning~~

To check the normality of the data :-

- Acceptable range of skewness is -0.5 to 0.5.
- from scipy.stats import shapiro.

shapiro (df['age']), it is going to return the p value.

H_0 : Data = Normal

H_a : Data ≠ Normal ← This is skewness effect

If ~~shapiro~~ when we do shapiro of our numerical column, it returns, for example :-

$P \rightarrow 0.000138$ (Data is skewed, ~~it~~ because we reject the null hypothesis)

If $P = 0.348$, then H_0 holds good, the data is normal.

Summary of Learning:

- Understanding probability distribution
  - Discrete
  - Continuous
  - Normal
- Central Limit Theorem → Theoritical and practical approach
  In practical approach what approximation we do, $\frac{s}{\sqrt{n}}$, ie standard error of mean.
- Range estimate for the confidence interval
- One sample Test
  - Test of mean — ttest_one samp (samp, μ)
  - Test of proportion — prop_ztest (x, n, prop)
- Two sample Test
  - Test of mean
    - Dependent
    - Independent
  - Test of proportion — prop_ztest ([x_1, x_2], [n_1, n_2])

One sample Test
- one sample proportion test with more than two categories.
- >2 Categories.
- Chi-square (observed count, expected count)

• Two sample proportion test doesn't have any flavours.

- If we are checking more than two group test.
  - Test of mean — ANOVA
  - Test of proportion — Chi2 - Contingency

Lab (statistics)

- what are the applications of chi-square.

→ Goodness of fit test

Example: Suppose you have a data, and you want to test whether the data supports poisson distribution. Such kind of tests are called goodness of fit.

→ Independence Test

The relationship between two features, for example x and y are independent or dependent.

→ Homogeneity test

To test how uniform the data is.

- If the calculated chi-square is exceeding the chi-square critical, then we reject the null hypothesis. (like any other test).

Contingency Table

means the cross tabs

Example: A dice is rolled 132 times, find whether it is biased or unbiased.

chi-square critical value $\chi^2_{crit} = 11.07$.

1	2	3	4	5	6
16	20	25	14	29	28

Expected should be 132/6 (equal chances)

	f_o (frequency observed)	f_e (expected)	O-E	$(O-E)^2$	$(O-E)^2/E$
1	16	22	-6	36	36/22
2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\chi^2 = \frac{\Sigma (O-E)^2}{E} = \frac{36+49+64+9+4+36}{22} = 9.$$

- Chisquare calculated $< \chi^2_{crit}$, so we will fail to reject the null hypothesis.

- This is the goodness of fit example.

Let's talk about Test of Independence,

How do we calculate the expected frequency when we have two features and you are trying to look at the proportions of the features w.r.t each other.

$$f_{expected} = \frac{\text{Column Total} \times \text{Specific row total}}{\text{Overall Total } (N)}$$

Example:- A total of 10,000 people voted, out of which some voted and some did not vote. Out of the people who voted 2792 men and 3591 women. and out of the people who didnt vote 1486 were men & 2131 women we need to check if the gender and voting are independent to each other or not.

Solution:-

$\chi^2_{crit} = 3.841$

Observed data

	Men	Women	
vote	2792	3591	→ 6383
no vote	1486	2131	→ 3617
	↓ 4278	↓ 5722	

Expected data

	Men	Women
vote	$\frac{4278 \times 6383}{10,000} = 2731$	$\frac{6383 \times 5722}{10,000} = 3652$
no vote	$\frac{4278 \times 3617}{10,000} = 1547$	$\frac{5722 \times 3617}{10,000} = 2070$

Chisquare $\chi^2_{cal} = 6.6$.

Since Calculated value is exceeding the critical value, we reject the null hypothesis. which means the gender and voting are dependent on each other.

Example: A sample of $n=60$ size is taken, and the observations that are seen are

0	1	2	3
32	15	9	4

Does the data follow poisson distribution or not.

Solution:

Hint: To calculate the expected frequencies, we should do the (probability distribution function) value $\times$ (sample size) for each case.

$fe = PDF \times n$ $\qquad \chi^2_{crit} = 7.814$

$= f(x) \times n$

- Because, this is a poisson distribution, we need to know, what is lambda.
- λ is the average : $\lambda = \frac{(0\times32)+(1\times15)+(2\times9)+(3\times4)}{60} = 0.75$

Exp data Calculation $\qquad f(x) = \frac{e^{-\lambda}\cdot\lambda^x}{x!}$

$$f(0) = \frac{e^{-0.75}\times(0.75)}{0!} = 0.47$$

$$f(1) = \frac{e^{-0.75}\times(0.75)^1}{1!} \ldots\ldots$$

$$fe_0 = 0.47 \times n = 28.34 \ldots\ldots\ldots$$

The chisquare calculated $= 4.47$

So, it is lesser than χ^2_{crit}. we fail to reject the null hypothesis.

Data follows Poisson distribution.

- This is a goodness of fit example.

ANOVA → Analysis of Variance

- we are analyzing the variance between the samples.

Why ANOVA:

- To this point we have been comparing two populations
  - → Independent samples t-test (random).
  - → Matched-sample t-test (paired).

- Of course limiting ourselves to the comparison of two populations is well ... limiting.
- What if we wish to compare the means of more than two populations?
- What if we wish to compare populations each containing several levels or subgroups?
- Enter ... ANOVA; ANalysis Of VAriance.

Things solved during ANOVA test,

- Suppose we want to compare three sample means to see if a difference exists somewhere among them.
- Is one mean so far away from the other two that is likely not from the same population?
- What we are asking is:
  - Do all three of these means come from a common population?
  - Or are all three so far apart that they ALL likely come from unique populations?

Null hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$

We have one way ANOVA

Two way ANOVA with repetition

Two way ANOVA without repetition

Suppose we have three samples, S_1, S_2 and S_3 and their respective means are $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$.

How would you calculate the f-ratio.

· Lets talk about one way ANOVA first,

· we need F value and compare it to the F critical

· If F_{value}^{cal} exceeds $F_{critical}$ then reject the null hypothesis.

$$F_{cal} = \frac{MSC}{MSE}$$

$$SSC = (\bar{x}_1 - \mu)^2 + (\bar{x}_2 - \mu)^2 + (\bar{x}_3 - \mu)^2$$

(between)

$$MSC = \frac{SSC}{df_c}$$

df_c = num of columns − 1

F_{cal} is Fcalculated.

df_c is degree of freedom of columns.

SSC is sum of squares of the columns.

$$MSE = \frac{SSE}{df_e}$$

df_e is degree of freedom of error
= total population (N) − num of columns (C)

· SSE is within and SSC is between

· within means variance within each sample

Basically the variance of each data point in a sample from its respective mean

$SSE = SST - SSC$ where SST is total sum of squares $\sum (x - \mu)^2$

Problem statement

· Twenty one students at the autonomous university of madrid (AUM) in Spain were selected for an informal study about student study skills; 7 first year, 7 second year and 7 third year undergraduates were randomly selected.

The students were given a study-skills assessment having a maximum score of 100. As researchers we are interested in whether or not a difference exists somewhere between the three different year levels. We will conduct this analysis using a one-way ANOVA technique.

$$F = \frac{MSC}{MSE}, \quad MSC = \frac{SSC}{df_c}, \quad MSE = \frac{SSE}{df_e}$$

$df_c = C - 1$

$df_e = (C-1)(B-1)$

$MSB = SSB / df_B(B-1)$

$df_T = N - 1$