STATISTICS:-

WEEK-1

DAY-1:-

ML

Prediction Model.

y → dependent variable
x → independent var.

Regression: $y = f(x)$
Nature of y here is Continuous values

Classification: $y = f(x)$
Here the nature of y is 0 or 1 (categorical).

n x p
no. of cols / records / fields / entries / features / instances / attributes / samples / headers / variables.

$y = \beta_0 + \beta_1 x$
↓ → Slope.
Intercept

β_0, β_1 → Model Parameters.
x → Predictor variable.

→ Historical data also known as Empirical data.
y → Dependent var / Target col / Outcome var.

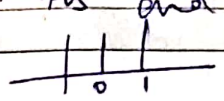
Entropy → Measure of uncertainty

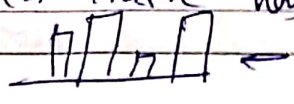
NOTE: There can be more than 1 outcome Var. Sunday 23
To predict y_1 and y_2 then you need to build one model to predict y_1 and 2nd model to predict y_2.

→ For a CATEGORICAL VAR you can make use BAR CHART or PIE CHART only. And you can only count the categorical variables.

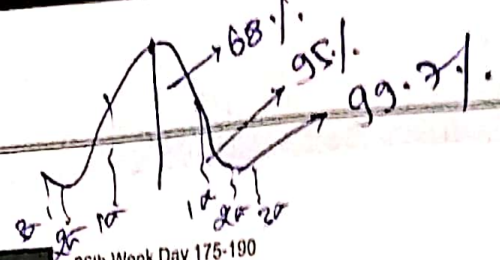
* The diff^n b/w BAR and HISTOGRAM is BAR takes discrete values on X-Axis and HISTOGRAM has cont^n values on X-axis is

BAR PLOT →





← Histogram



OUTLIER:- Any value greater than 3σ ($3 \times$ Std. Mean)

$|x - \bar{x}| > 3\sigma$

→ You need to know whether an outlier is an outlier or an EXTREME value.

OUTLIER
1. Remove
2. Replace with median
3. Resampling

EXTREME
1. Log TRANSFORM.

→ Extreme values are important
→ OUTLIERS are NOISE and unwanted

→ Exponential transformation is applicable when all of your data is in the range of 1σ. This is also not a good representation of data.

Inferencial Stats:-

→ When your sample data infers about the entire population.

Descriptive Stats:-

→ Talks about the columns in the data only.

TRUE MEAN:- You mean and my mean remains the same.

↓

Mean of the Population.

true mean → GOD knows.

N → Population Size
μ → population mean (true mean)
σ → Standard deviation

n → Sample size
x̄ or ȳ → Sample mean.
s → Standard devi. of sample

Note: To perform inferential stats the minimum sample must be n > 30 and has to be highly random.

Hypothesis Testing:
① Test of Mean
done for numerical data (T-test)

② Test of Proportion.
done for Categorical data. (Z-Test)

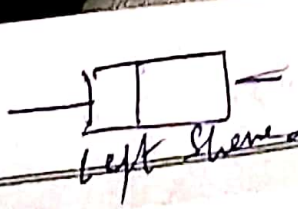
	Continuous	Categorical
Univariate	Histogram Box plot	Bar plot pie chart.
Bivariate analysis	→ scatter plot.	

df.corr().

Quantiles:
$q_2 = \frac{1}{2}$ (Sample) $= \frac{93}{2} = 46.5 \simeq 47^{th}$ element

$q_1 = \frac{1}{4}$ (sample) $= \frac{93}{4} = 23.25 \simeq 24^{th}$ elem.

$q_3 = \frac{3}{4}$ (Sample) $= \frac{3}{4} \times 93 = 69.75 \simeq 70^{th}$ elem



JUNE

26

Wednesday

26th Week Day 177-188

M	T	W	T	F	S	S	Wk Jun 2019
					1	2	22
3	4	5	6	7	8	9	23
10	11	12	13	14	15	16	24
17	18	19	20	21	22	23	25
24	25	26	27	28	29	30	26

06

SHALLOW COPY :-

```
A = [1, 2, 3]
B = A
B[1] = 20
print (B) = [1, 20, 3]
print (A) = [1, 20, 3]
```

Changes made in B are reflected in A.

Probability :-

Used Mainly ← (1) Classical Probability :-
(2) Empirical probability (Historical Data).
(3) Subjective Probability.

Probability Distribution :-

→ When you have only 2 outcomes then it is a binomial distribution.

→ When you have many number of outcomes then its a Poisons distribution.

Calculating Probability in Bromial Distribution :-

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

$$\Rightarrow P(x=1) = {}^{50}C_{1}\,(0.05)^{1}\,(0.95)^{49}$$

binomial
poisson
normal

Formula for Poison Dist:-

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

$\lambda = 7$ in the previous case when on an avg no. of hotels in a shirtan? ..2

Normal Distribution:-

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu^2)}{2\sigma^2}}$$

where,
$f(x)$ is used to represent a probability density fun?
x is any value b/w $-\infty < x < \infty$,
$e = 2.7)$, $\pi = 3.14$.

Central limit Theorem:-
→ Majority of the data lies closer to the mean b/w -1σ and 1σ zone.

Standard Deviation for Normal Dist:-

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Binomial Dist:-

$$SD = \sqrt{np(1-p)}$$

Poison Dist:-

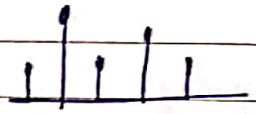
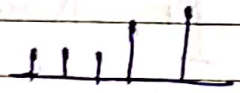
$$SD = \sqrt{\lambda}$$

28 Friday 28th Week Day 179-186

06

Random Variable :

Categorical	Discrete	Continuous
Yes / No	No. of Rooms in a House	Price of House
True / False	PClass in Titanic	Rate of Speed.
0 / 1	↓	↓
Def / ND.	You make use of POISONS DIST	You make use of NORMAL DIST
↓	↓	↓
You make use of BINOMIAL DIST	You get multiple OUTCOMES.	The values are Cont.
↓		
You get only 2 OUTCOMES		

Binomial		pmf (prob. mark fn")
Poisson		pmf (prob. mark fn")
Normal		pdf (prob. density fn").

NOTE: In a normal dist. if I wish to calculate $P(X = 45gm)$ is almost $\simeq 0$. Because in a Normal DIST you find the Density. (Area under the curve).

→ For one single point AREA under the curve is zero $\frac{1}{45}$

→ In normal dist you need two values to get the probability.

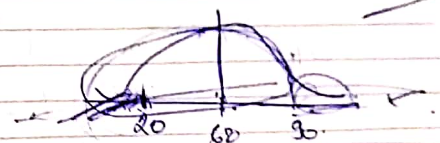
Unit less representation of the Normal Dist of const^n values

Standard Normal Dist :-

Scaling of data and making it ~~scale~~ unit less will make two changes mean = 0 and SD = 1.

Standard Error :-

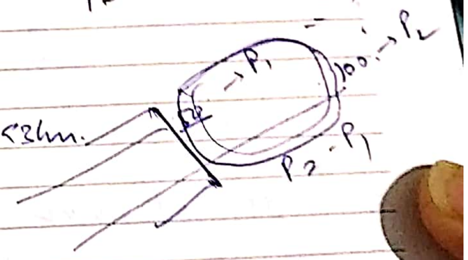
$$\frac{\sigma}{\sqrt{n}}$$



$y \pm \frac{1.96\sigma}{\sqrt{n}}$ → Spread of allowable error.

Covers 95% of data.

stats.norm.cdf(?)



Given Company claims delivery time ≤ 3 hrs.

Sample by agency:

$n = 50$

$\bar{x} = 2.8$ hrs

$s = 0.6$ hrs

Sunday 30

Now we need to infer the true mean (μ) which will be in a RANGE.

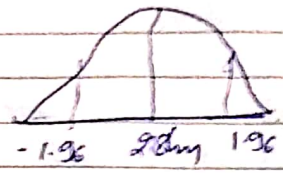
Confidence interval required is 95%.

For 95% confidence the range will be

$$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right]$$ → Confidence interval Range

for 99% $$\left[\bar{x} - \frac{2.58\sigma}{\sqrt{n}}, \bar{x} + \frac{2.58\sigma}{\sqrt{n}}\right]$$

for 98% → 2.33 for 90% → 1.645.



∴ Confidence Interval $= \left[2.8 - \frac{1.9(0.6)}{\sqrt{50}},\ 2.8 + \frac{1.9(0.6)}{\sqrt{50}}\right]$.

$= [2.63, 2.96]$

The interval from 2.63 to 2.96 forms 95% confidence interval for μ. In other words we are 95% confident the avg delivery time lies b/w 2.63 and 2.96. Since all the values are less than 3hrs we have a strong evidence that the claims are correct.

→ There is 5% error this is called α-error.

→ To get 1% error lets compute Confidence for 99%)

∴ Range $= \left[2.8 - \frac{2.58(0.6)}{\sqrt{50}},\ 2.8 + \frac{2.58(0.6)}{\sqrt{50}}\right]$.

$= [2.58, 3.01]$.

To CONCLUDE which interval is true for our population then we need to perform HYPOTHESIS TESTING.

TEST of MEAN (or) TEST of PROPORTION.

t-test

*) $n = 36$, $\bar{y} = 573$, $s = 124$

Confidence interval 95% : $\left[573 - \frac{1.9(124)}{\sqrt{36}}, 573 + \frac{1.3(124)}{\sqrt{36}}\right]$.

$= [532.99, 613.5]$.

Prof:

*) When $n \geq 30$ according to Statistical theory

$s \simeq \sigma$

where $s \rightarrow$ Standard deviation of sample

$\sigma \rightarrow$ ——— of population.

more or ——.

*) t-distribution is same as normal distribution

It is used during hypothesis testing.

For normal dist $Z = \frac{x - \bar{x}}{s_D}$

For t-dist $Z = \frac{x - \bar{x}}{s/\sqrt{n}}$.

HYPOTHESIS TESTING:

$H_a \rightarrow$ Alternate Hypothesis (or) Researcher Hypothesis (or) Organ hypo

In H_a the claim is made on the population.

$H_0 \rightarrow$ Null Hypothesis. (or) No Difference Hypo.

for H_a: $\mu_0 < 3$hrs

H_0: $\mu_0 \geq 3$hrs.

H_a: $\mu_0 > 520$ (Right Tail)

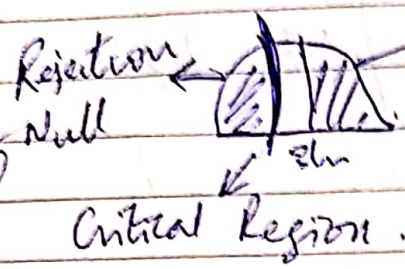
$\mu_0 \leq 520$. (left Tail).

JULY THIS is One Sample t-test.

03 Wednesday 27th Week Day 184-181

M T W T F S S | Wk Jul 2015
1 2 3 4 5 6 7 | 27
8 9 10 11 12 13 14 | 28
15 16 17 18 19 20 21 | 29
22 23 24 25 26 27 28 | 30
29 30 31 | 31 07

H_a : $\mu_0 < 3hr$ H_0 : $\mu_0 \geq 3hr$

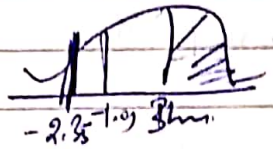


CRITICAL REGION is calculated as.

$$t_{stat} = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{2.8 - 3}{0.6/\sqrt{50}} = -2.35$$

RULE for REJECTING NULL HYPOTHESIS :

	95%	99%
t_{stat}	Should be > 1.96	Should be > 2.58
Pvalue	< 0.05	< 0.01



* Pvalue → probability of null hypothesis being TRUE.

* One Sample t-test :-

One Sample → (population ⊃ sample).

t-test → Test of Mean using t distribution.

JULY 04 Thursday

Two Sample t-test :-

$$t_{stat} : \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For 2 Sample t-test we have = or ≠ criteria.
only for 1 Sample t-test we have >, <, ≥, ≤.

*) $n = 30$, $\bar{y} = 2.28$, $s = 0.65$

H_a : $\mu < 3$hrs
H_0 : $\mu \geq 3$hrs

$$t_{stat} = \frac{2.28 - 3}{0.65/\sqrt{30}} = \frac{-0.22}{0.65/5.36} = \frac{-0.22}{0.072}$$

*) $t_{stat} = -2.55$.

H_0 = $\mu_{hypothesized} = \mu_{actual}$ (or) $\mu_{actual} - \mu_{hypothesized} = 0$

*) $\mu_{hypothesized}$

Type I and Type 2 ERRORS:-

Type 1 error is α error
Type 2 error is β error.

DECISION	NULL HYPOTESIS H_0: TRUE	H_0: FALSE.
Reject H_0	(Wrong decision) Type I error α-error	Correct Decision $1-\alpha$.
Accept H_0	Correct Decision $1-\beta$	Type 2 error β-error

Type 2 error is a costly error.

JULY

05 Friday 27th Week Day 186-179

P value is the probability of NULL Hypothesis being TRUE.

n → 400 D → 100

		H_0 : TRUE	H_0 : FALSE
		0	1
Reject H_0	Healthy 300	(1-α) TN 280	FP 20 (α)
Accept H_0	Disease 100	FN 40 (β)	TP 60 → (1-β) → Sensitivity of the Model.

∴ Sensitivity of the model = $\frac{60}{100}$ = 60%.

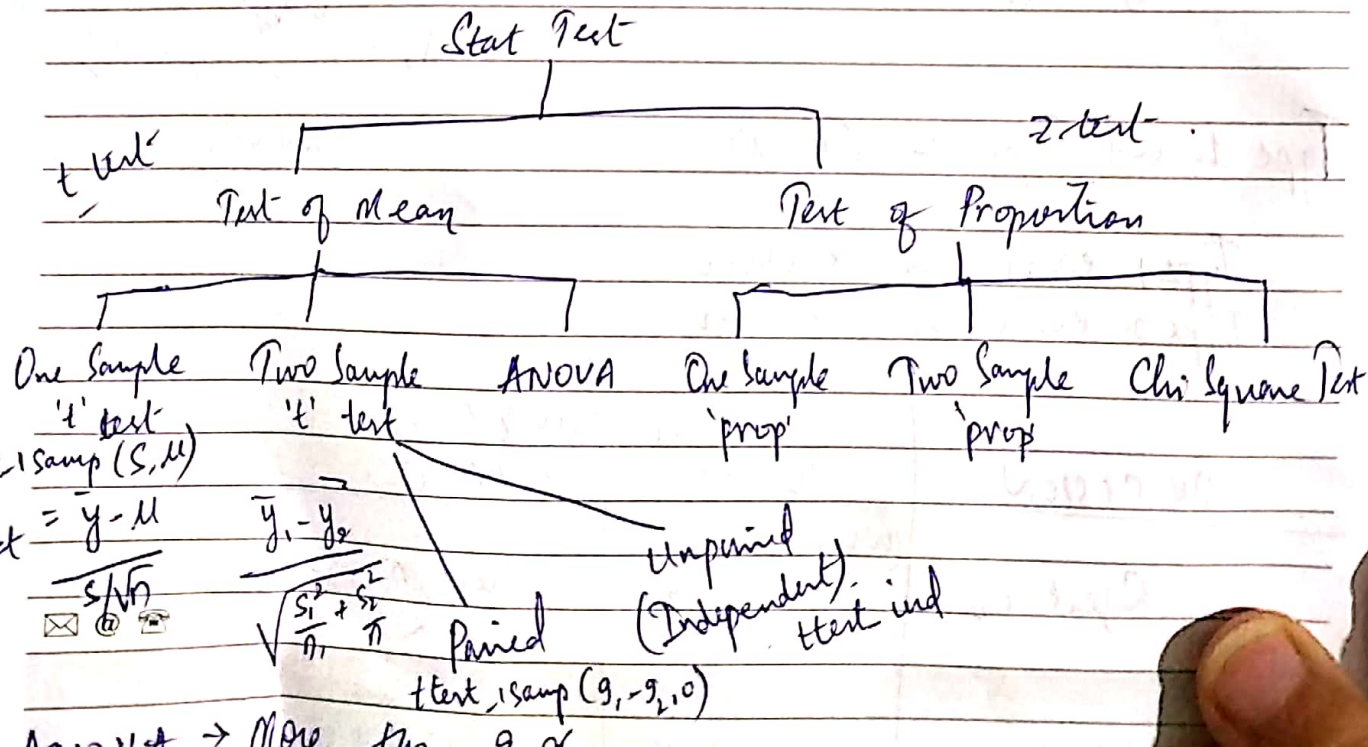
1-α → Specificity of the model = 280/300 ≃ 95%.

1-β → Sensitivity of the model.

Notes:- First importance is given to sensitivity.

Sensitivity is also referred to as TRUE POSITIVE RATE

Specificity is also referred to as TRUE NEGATIVE RATE



ttest, 1samp (S, μ)

$t_{stat} = \frac{\bar{y} - \mu}{S/\sqrt{n}}$

$\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n}}}$

ANOVA → More than 2 groups.

If Normal → Parametric
Vice versa.

JULY 06 Saturday
27th Week Day 187-178

Aug 2019 Wk

	M	T	W	T	F	S	S
31				1	2	3	4
32	5	6	7	8	9	10	11
33	12	13	14	15	16	17	18
34	19	20	21	22	23	24	25
35	26	27	28	29	30	31	

08

Sometimes t test referred as z-test when population Standard Deviation is given to us.

ttest_1samp (g1, g2, 0), ttest_ind ( ) → parametric.
These are followed only if your data is normal or not.

SHAPIRO TEST :-
Test for normality. (Check if your data is normal or not).
from scipy.stats import ttest_1samp, ttest_ind, shapiro.

→ Shapiro will return the p value.
shapiro(age).
if p = 0.837.
Null hyp is accept the Ho.
Now if the data is normal then you can proceed with parametric test.

H_0: data = Normal
H_a: data ≠ Normal.

→ If it is not normal, then we will go with non-parametric test.

Sunday 07

Two Sample t test

Paired

Unpaired

Wilcoxson

ttest_1samp(g1,g2,0)

ttest_ind

Mannwitneyu.

Non parametric

Parametric

Parametric

Non-Parametric

JULY

08

Monday

28th Week Day 189-176

If all 3 cond^{n}s satisfy, then you go for parametric test.

Two Sample t test :-

1. Random Variable (Random Samples)
2. Data is Normal (Shapiro)
3. Test of Variance ($\sigma_1^2 = \sigma_2^2$).

↓

Go for parametric test.

- Independent: ttest_ind (g_1, g_2)
- Dependent: ttest_rel ($g_1 - g_2, 0$).

→ If any one cond^n fails, you do non-parametric

Non-Parametric

- Independent: Mannwhitneyu
- Dependent: Wilcoxon.

Test of Variance.

- Data is Normal: (Levene)
- Data is Not Normal: (Bartlett).

Here $H_0: \sigma_{g_1}^2 = \sigma_{g_2}^2$

$H_a: \neq$

Wilcoxon & Mann-- are used very OFTEN
→ used 90% of the times.

Aug 2019 Wk | M T W T F S S | JULY

31 | 1 2 3 4
32 | 5 6 7 8 9 10 11
33 | 12 13 14 15 16 17 18
34 | 19 20 21 22 23 24 25
35 | 26 27 28 29 30 31

08

ANOVA → Analysis of Variance

09

28th Week Day 190-175 | Tuesday

Test of Normality (Shapiro)	Test of Variance (Levene / Bartlett)	Test of Mean (ttest_ind / Mannwhitneyu)
H_0: Data = Normal	$H_0: \sigma_{g_1}^2 = \sigma_{g_2}^2$	$H_0: \mu_{g_1} = \mu_{g_2}$
H_a: Data ≠ Normal	$H_a: \sigma_{g_1}^2 \neq \sigma_{g_2}^2$	$H_a: \mu_{g_1} \neq \mu_{g_2}$

Test of PROPORTION :-

$$Z_{stat} = \frac{p_1 - p_2}{\sqrt{p_{pooled}(1-p_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}$

Two Sample Proportion test.

programmatically shall be done by using proportions_ztest( )

CHI SQUARE TEST :-

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

O → observed count E → expected count.

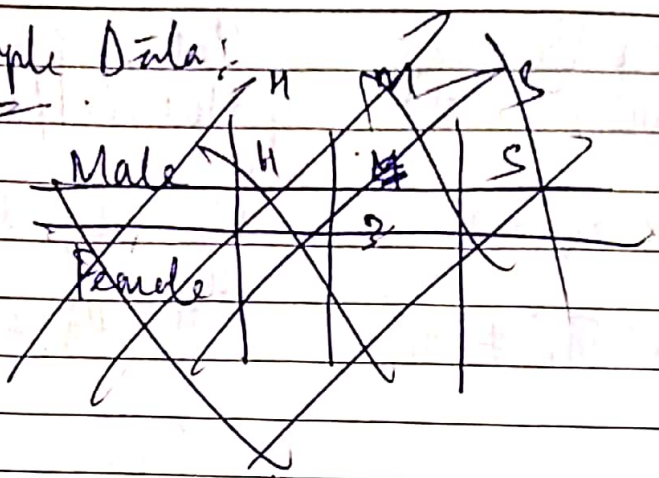
$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{Grand total}}$$

```
from statsmodels.stats.proportion import proportions_ztest
counts = np.array([150, 87])
obs = np.array([882, 588])
proportions_ztest(counts, obs)
```

10

Wednesday 28th Week Day 191-174

Example Data:-



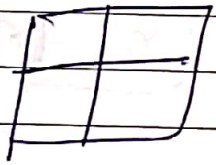
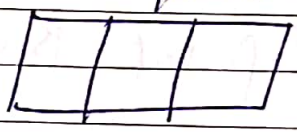
	H	M	S
Male	410	340	255
Female	95	85	70.

} Observed Count.

Expected Count =) 404 340 256
101 85 64

} Obtained using the formula for expected count.

Chi2. contingency chi2 contigency Chisquare.

✉ @ ☎

One Sample Prop. Test:-

Ho: 0.25 (25% of middle age women being diabetic)
Ha ≠ 0.25

Test of Proportion at a Glance

(*) For one sample proportion test → proportions_ztest (~~total~~ actual count, total_samples, claim proportion on population).

(*) Two sample proportion test.
proportions_ztest (counts, obs)
↓ ↓
actual count · [n1, n2]
[x1, x2]

(*) → 2 sample proportion test.

Chi2_contingency (crosstab).

(*) Chisquare (value_counts of single categorical variable) ↙ used majorly in sampling process.

	H	HD
E	127	54
NE	23	66

Ex vs HD.

n_1 → ~~total~~ exercise
n_2 → total not exercise
x_1 → ~~total~~ HD.
x_2 → HD.

proportions_ztest([54, 66], [181, 89]).

Here H_0 : $P_{HD}(E) = P_{HD}(NE)$

12

Friday 28th Week Day 193-172

07

STATISTICAL TEST

- Test of Mean
  - One Sample 't' test → t test_1samp → Parametric (t test_1samp) ($9_1 - 9_{2,0}$)
  - Two Sample 't' test
    - Paired (Dependent) → Parametric (t test_1samp) ($9_1 - 9_{2,0}$)
    - Paired (Dependent) → Non-Parametric (Wilcoxon)
    - Unpaired (Independent) → Parametric (t test_ind)
    - Unpaired (Independent) → Non Parametric (mannwhitneyu)
  - > Two groups w.r.t one categorical variable → f_oneway (ANOVA)
  - > Two groups w.r.t two or more categorical variable → 2 way_ANOVA → OLS library.
- Test of Proportion
  - One Samp Z test with single category (Proportions_ztest)
  - One Samp - Z test with multi category (chisquare)
  - Two Sample z test (Proportions_ztest)

	M	F
Y		
N		

  - > Two groups (Chi2_contingency)

	A	B	C
M			
F			

Everything at a Glance :-

(Q1) Whether heart disease is having any impact on the cholestrol.

here $H_0 : \mu_{CH} = \mu_{CNH}$ (Test of Mean)

$H_a : \mu_{CH} \neq \mu_{CNH}$

① Two Sample t-test.
② It is a independent hence unpaired
③ If not parametric go

(Q2) Effect of vessel on cholestrol level.

Here $H_0 : \mu_{c(0)} = \mu_{c(1)} = \mu_{c(2)} = \mu_{c(3)}$.

$H_a : \mu_{c(0)} \neq \mu_{c(1)} \neq \mu_{c(2)} \neq \mu_{c(3)}$.

Clearly we have cholestrol as numerical and vessels has 4 Categories

Sunday 14

Therefore you have 4 groups.

Hence we use f-oneway ANOVA.

JULY

ONE NUMERIC with 2 CATEGORICAL is 2way ANOVA.

Migrain Date

15

Monday

29th Week Day 196-169

or more than 2 categorical

M T W T F S S | Wk Jul 2019

1 2 3 4 5 6 7 | 27
8 9 10 11 12 13 14 | 28
15 16 17 18 19 20 21 | 29
22 23 24 25 26 27 28 | 30
29 30 31 | 31

07

(*) Whether age is a factor influencing type of headache.

age → numerical
headche → 3types Categorical

Clear cut for Oneway ANOVA.

(*) age vs headache vs Gender

is 2way ANOVA.

CAR CASES

(*) Any Dependency b/w drive train and no. airbags in car.

Drive train → Categorical
Airbags → Categorical.

Test of Proportion.

H_0 : $P_{FN} = P_{FDO} = \ldots$

This is chi2-contingency. find the crosstab and pass it to chi2-contingency.

X! Note: If null hyp does not hold then you conclude that there is a dependency

V. Imp

Maths behind Anova.

MSTR - Between Sample Variability :-

$$MSTR = \frac{SSTR}{\text{degree of freedom}} = \frac{\sum n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$F_{stat} = \frac{MSTR}{MSE} = \frac{SSTR/df_1}{SSE/df_2}$$

$$MSE = \frac{\sum (n_i - 1) s_i^2}{n_t - k} = \frac{SSE}{df_2}$$

$s_i \rightarrow$ Standard deviation. $df_2 =$ no. of samples, n of groups

For Variance :- $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

In python its 'n' in the denominator
to correct it you need to pass np.var(A, ddof=1).
By default it is 'n-1' in 'R'

Covariance :- (Bivariate).

$$Cov(x, y) = E(x - \bar{x})(y - \bar{y})$$

$$\Rightarrow \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Cov (wt, ht) => 250 kg cm
Cov (wt, age) => 125 kg yrs
To remove the units you divide by SD.

$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$ → varies from -1 to 1.

JULY

17

Wednesday

29th Week Day 198-167

M T W T F S S | Wk Jul 2019

1 2 3 4 5 6 7 | 27
8 9 10 11 12 13 14 | 28
15 16 17 18 19 20 21 | 29
22 23 24 25 26 27 28 | 30
29 30 31 | 31

07

Linear Regression :-

$$y = \beta_0 + \beta_1 x$$

$$\beta_0 \rightarrow \bar{y} - \beta_1 \bar{x} \quad , \beta_1 = \frac{Cov(x, y)}{Var(x)}.$$

For Scikit learn Approach :-

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X, y)
y = model.predict(X).
```

Steps at a Glance :-

x_1	x_2	y

① `X = DF[['x1', 'x2']]`
`Y = DF['y']`

OLS Approach :-

```
model = ols('y ~ x1 + x2', DF).fit()
y_pred = model.predict(X).
model.params → will give you β0, β1, β2.
```

Sklean approach :-

X has to be in 2 Dimensions.

```
X = DF.drop('y').
y = DF['y'].
```

Framing the Hypothesis for linear Regression:

$H_0: \beta_1 = 0$ (β_1 = Slope)

$H_a = \beta_1 \neq 0$

$\beta_1 = 0$ ✗

$\beta_1 \neq 0$ ✓

Performance of Regression Model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Higher the R^2 higher the prediction Rate.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}}$$

Best Model Selection:

	R^2	RMSE
$y = \beta_0 + \beta_1 R$	35%	4.25
$y = \beta_0 + \beta_1 TV$	61%	3.24
→ $y = \beta_0 + \beta_1 TV + \beta_2 R$	89%	1.66

Best Model. Here the multivariate clearly outperforms the univariate.

✉ @ ☎