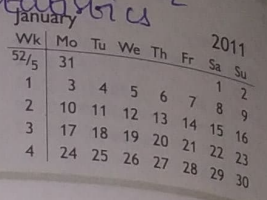
31

Monday

January 2011

* { descriptive value for population → parameter
" " " sample → statistics



Week 5 ▪ 31-334

#

Day-7.

Different aspects of python packages -
- Numpy
- Pandas
- Matplotlib
- Sikit-learn
- Scipy.

Day-8.

Introduction to World of statistics.

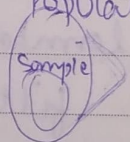
Statistics → It is a collection, organization, analysis & interpretation of data.

- Design of experiments
- Sampling.
- Descriptive statistics
- Inferential statistics
- Probability Theory.

Numerical way of analzying data ↓ Stats.

Why we want to learn stats.?
- Draw conclusion abt large sets.
- we can make forecasting about business actinty.
- Improve business process.

describe & presenting data clearly,



Types of stats.

Descriptive (Sumarization of data) — Mean, median, mode

Inferential. → Make prediction (using hypothesis, concluding result of whole popln.)
take small data → conclude on large data.

- Types of Variable. (Variables are predefined)
  - categorical (Predefined categories): Ex. Gender, Season,
  - continous.

(numerical variables) → measured in numbers.

Discrete	continous
(Finite)	(Infinite range)
Ex. No. of children	Ex sales, ... Real Numbers.

Method 1.

Measure of central tendency.

(1) mean, (2) median, (3) mode.

Sample mean: $\bar{x} = \frac{\Sigma x}{n}$

Population mean: $\bar{y} = \frac{\Sigma x}{N}$

Trimmed mean

Removing extreme value and finding mean. (outlier removal)

obtained by deleting a % of smallest & largest values from a dataset & then computing the mean of remaining values.

Geometric Mean.

When want to ~~count~~ check growth rate in financial data

calculated by finding nth root the product of n values.

Weighted Mean.

Find avg. wrt category wise.

$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i}$

w_i similar for $\left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)$

February 2011

Wk	Mo	Tu	We	Th	Fr	Sa	Su
5		1	2	3	4	5	6
6	7	8	9	10	11	12	13
7	14	15	16	17	18	19	20
8	21	22	23	24	25	26	27
9	28						

Note
arrange values in ascending order

(2) Median →

To find middle values.

- Observation → even.
  take two values & find mid pt.
- Observation → odd
  take the middle value.

(3) Mode

most frequent (repeated) value.

method 2:

Dispersion measures.

It describes the data spread ~~how~~ or how far measurement are from center.

- Range → max. value − min. value.
- Mean Deviation → How much value is deviated from mean.
- Percentile → Relative performance.

Ex. $i = (P/100)\,n$

Find 80th percentile → $(80/100) \times 70 = 56$

even no. (56) →

Avg. of 56 and 57 data value:

80th percentile = $(535 + 549)/2$

$= 542$

- Quartiles -

First Quartile → 25 percentile
Second Quartile → 50 percentile (Median).
Third Quartile → 75 percentile.

Ex. Third Quartile. → 75 percentile.

$i = (P/100)n = (75/100)\,70 = 52.5 \simeq 53$

Third Quartile (Q3) : = 525 ← value at 53 posn of data.

* Inter Quartile range.
dist. b/w 1st & 3rd quartile.

Q1	Q2	Q3	
25%	25%	25%	25%

Q3 − Q1 (interquartile range)

(3rd Quar.) Q3 = 525
(1st Quar.) Q1 = 445
(Inter Quar. range) Q3 − Q1 = 525 − 445 = 80

* Variance.
difference b/w the value of each observation (x_i) & the mean ($\bar{x}$ for a sample, μ for a population).
↓ used to find variation in data

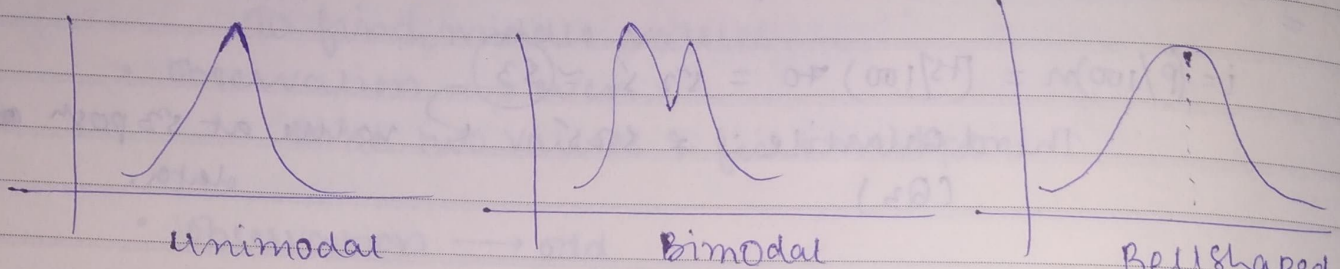
* Standard Deviation. → $\sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$
Use to measure the consistency.

Small S.D → more consistent
Large SD → less consistent

4 Friday February 2011

Method 3.

Distribution of shape

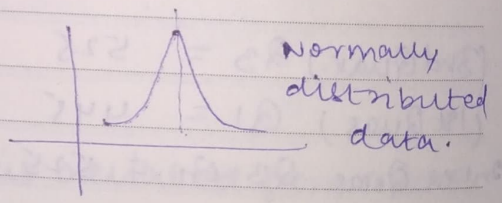


Analyze collection of Data.

- Skewness (Used to measure → symmetricness.)
- BoxPlot
- Kurtosis
- Scatter Plot
- Coeff. of variance.

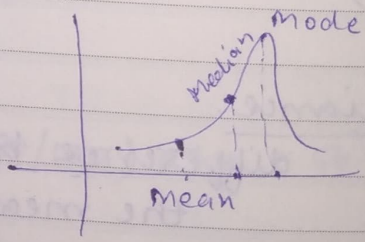
→ If skewness = 0,

Mean = Median = Mode



Skewness = -ve

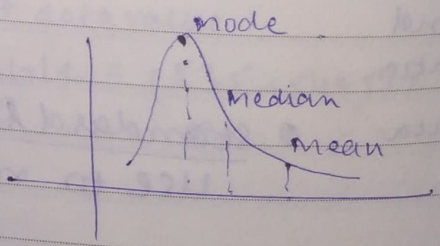
Mean < Median



all graphs are unimodal.

Skewness = +ve

Mean > Median



Skewness (S) = $\frac{3(\mu - \text{Median})}{\sigma}$

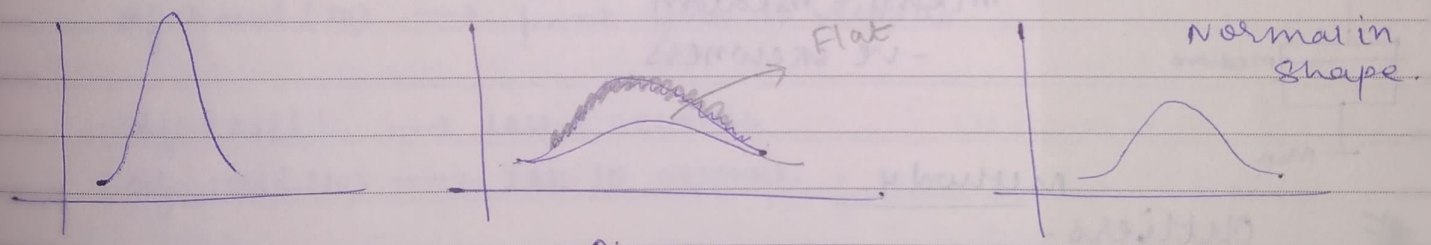
Mean < Median | Mean = Median | Mean > Median

Sx.

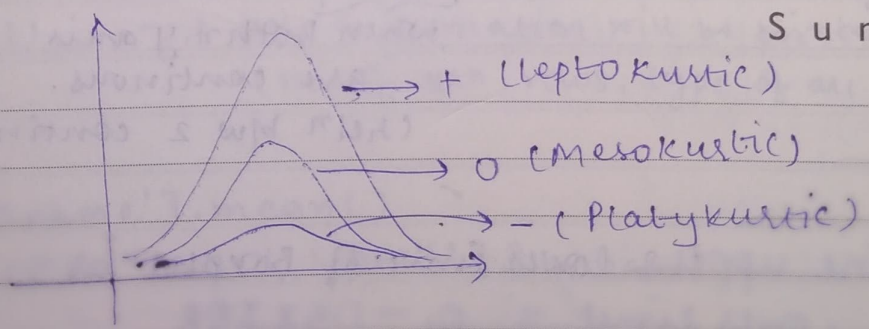
$\mu = 23$	$\mu = 26$	$\mu = 29$
$M_d = 26$	$M_d = 26$	$M_d = 26$
$\sigma = 12.3$	$\sigma = 12.3$	$\sigma = 12.3$
$S = \frac{3(\mu - M_d)}{\sigma}$	$S = \frac{3(\mu - M_d)}{\sigma}$	$S = \frac{3(\mu - M_d)}{\sigma}$
$= -0.73$	$= 0$	$= +0.73$
-ve skewed.	Neutral	+ve skewed.

Kurtosis

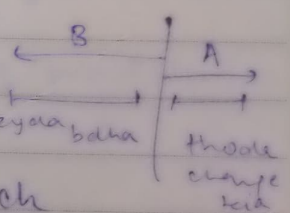
Used to measure peakness of data.



Leptokurtic (Peaky (High & thin)) | Platykurtic (Flat & spread out) | Mesokurtic (Bellshaped)



Measure distribution is flat or peaked.



Coefficient of Variation.

When we change one variable, how much variation is there in other variable.

$$CV = \left(\frac{S}{\bar{x}}\right) \cdot 100\%$$

7 Monday

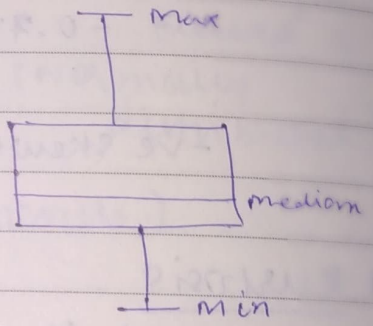
February 2011

Week 6 ▪ 38-327

Box plot

↓ visual representh of data

- Minimum
- Q_1
- Median
- Q_3
- Maximum



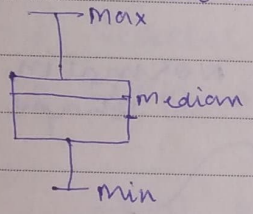
line is Below centre → +ve skewness

mean > median

line is above centre :

mean < median

-ve skewness



Method

Outliers.

↓ extreme values.

→ find by Boxplot

→ Or by Scatterplot

when Both x, y axis are continous.

(relⁿ b/w 2 continous variable)

Redefine upper & lower limit of Boxplot as :

Lower limit = $Q_1 - 1.5 \times IQR$

upper limit = $Q_3 + 1.5 \times IQR$

DAY-9.

File Name for today's case study.

Descriptive statistics.xlsx

```
import pandas as pd
df = pd.read_excel("Descriptive statistics.xlsx", sheetname = 0)
```

refers 1st sheet

if you have data on diff. sheet change no. acc to another sheet

You can also give sheet name.

```
Sheet_name = 'Descriptive statistics'
```

~~df.head~~

```
df.head()      ⟶ print first 5 records
df.head(10)    ⟶ print first 10 records.

df.tail()      ⟶ last 5 records
df.tail(10)    ⟶ last 10 records.

df.columns     ⟶ print all columns name.

df.info()      ⟶ all information will be printed.
                 (no. of coln, rows, type of all coln ... etc)

df['colnname'].mean()
df[['coln1','coln2','coln3']].mean()
```

for more than 1 var. use double sq. bracket.

print mean of 3 diff coln.

Similarly,

. drop duplicates
. isnull().sum()
. dropna
. duplicated()

df[['coln1','coln2','coln3']].median()

" . mode()

" . var()

" . std()

" . describe()

↓

in one shot you get all values mean, std, ~~var~~ min, 25 percentile, (Q_1), Q_2, Q_3, max...

↓

median

" . skew()

" . kurt()

How to find outliers?

1. Box plot
2. scatter plot.

plt.boxplot(df.coln)
plt.scatter(df.coln1, df.coln2)
plt.hist(df.coln)

compare two/multiple columns →

• pd.crosstab(df['Gender'], df['JobCat'])

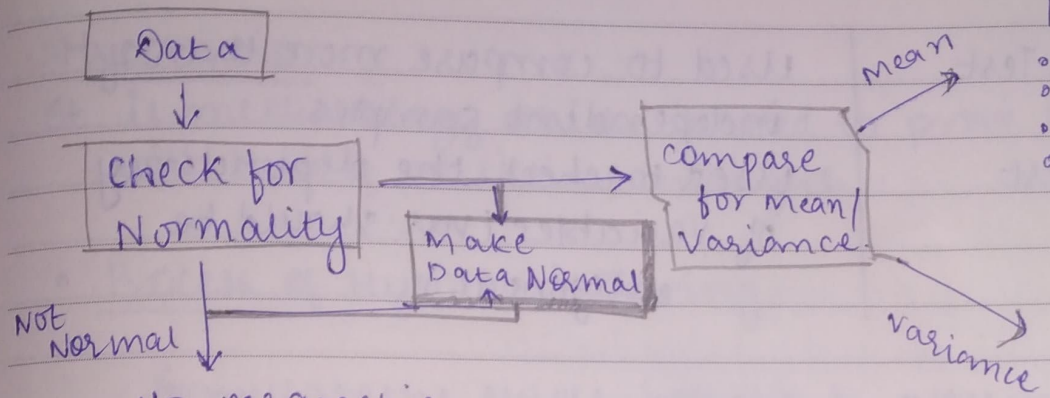
coln names.

Inferential

↓ Non-parametric (data is not normally distributed)

check by using skew, kurt etc

↓ Parametric (data is normally distributed).



Parametric Test.
- 1 sample T test
- 2 sample T test
- 2 sample paired Test
- 1 sample Test (ANOVA).

HOV (F-Test)

Non-parametric Test
- wilcoxon - sign Test
- Friedman
- mann-whitney
- Kruskal-Wallis
- chi-square.

9
10
11
12
1
2
3
4
5
6

Non-parametric.
- Wilcoxon - Sign Test
- Friedman Test
- Mann-whitney Test
- Kruskal-wallis Test
- chi-square Test

Parametric
- One-sample T-test
- Two sample Paired t-test
- Two sample Seperate t-test
- one sample Ftest

#one way (ANOVA)
- two sample Ftest

#two way (ANOVA)

Non-parametric

(1) wilcoxon - sign Test

(2) Friedman Test

(3) Mann-whitney Test

Purpose

It is used to compare two paired sample

It is used to compare more than two paired sample.

It is used to compare two independent samples.

(4) Kruskal-Wallis Test	Used to compare more than 2 independent samples.
(5) Chi-Square Test	Used to check the dependency of variables (var. should be categorical).

Parametric

(1) One-sample t-test	Used to compare sample mean with population mean
(2) two-sample paired t-test	Used to compare mean of two paired sample
(3) two-sample seperate t-test (independent) or	used to compare mean of two independent samples.
(4) One-sample F-Test or (one-way ANOVA).	Used when dependent var. is continous & independent var is categorical.
(5) two sample F-Test or (two-way ANOVA)	Used when dependent variable is continous & independent variable is categorical. in which there are 2 independent variable.

Non-Parametric

- Without using population parameters → you are analyzing data
- data → not normally distributed.
- NO-use of mean/median/mode.

Day 10.

Hypothesis.

It is methodology which is used to prove or disprove statement with help of statistical analysis.

Process of Hypothesis Testing.

Formulate the Null hypothesis & alternative hypothesis.
↓
Select appropriate test stats. 9
↓ 10
choose level of significance, confidence Interval, & degree of freedom 11
(α) (CI) (DF)
↓ 12
compute calculated test value of test statistics 1
↓ 2
compute table test value of test statistics 3
↓ 4
Compare calculated values & table values. 5
↓ 6
Make statistical decision & state the managerial conclusion.

1) Hypothesis formulation.

- **Null Hypothesis (H_0).**
  - What we are assuming.
  - → state claim or assertion to be tested.
  - Ex. Avg. no. of TV sets in US Homes equal to 3
    $H_0 : \mu = 3$
  - It is about population parameter, not about sample stats.

$H_0: \mu = 3$ ✓ $H_0: \bar{X} = 3$ ✗

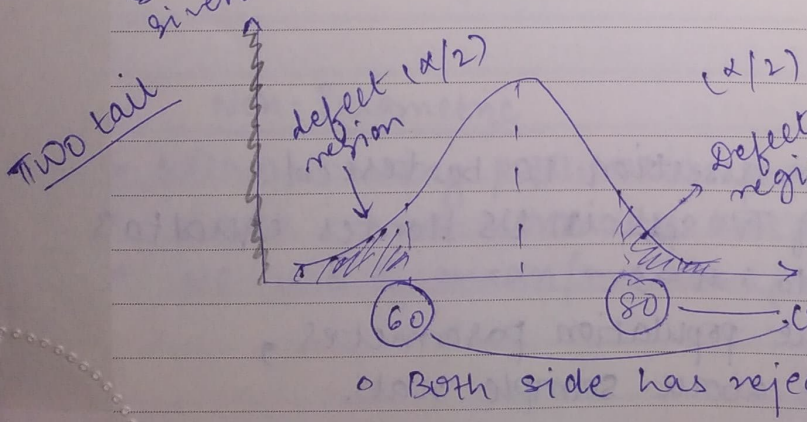
Always contain: ($=$ or $\leq$ or $\geq$) sign.

- **Alternative Hypothesis (H_1)**
  → opposite to Null Hypothesis.
  Ex Avg no. of TV sets in US homes is not equal to 3.
  $H_1: \mu \neq 3$
  - Never contain: ($=$ or $\leq$ or $\geq$) sign.
  - May or may not proven.

2.) Select appropriate Test:

Hypothesis Test Stat. for μ

- When Both side rejection is there → Two tail test
  - σ known (z test) — when σ is given
  - σ unknown (t test)
- When there is only 1 rejection is there → One tail test
  - Left Tail Test
  - Right Tail Test



- Both side has rejection area.

① Null hyp. should have equal sign

Ex $H_0: \mu = 3$
$H_1: \mu \neq 3$

Ztest (σ known).

→ populn SD is given

$$Z_{stat} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{X}$ → sample mean; μ → popⁿ mean; σ → popⁿ SD; n → sample.

t-test (σ unknown).

$$t_{stat} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

s → sample SD.

- Based on quality of data → α is defined. or in project client define 'α'.

Ex. Two tail test (σ Unknown).

$H_0: \mu = 168$

$H_1: \mu \neq 168$

∴ donot reject H_0. insufficient evidence that true mean cost is diff. than $168

Given: $n = 25$

$\bar{X} \rightarrow \$172.50$

$s \rightarrow \$15.40$

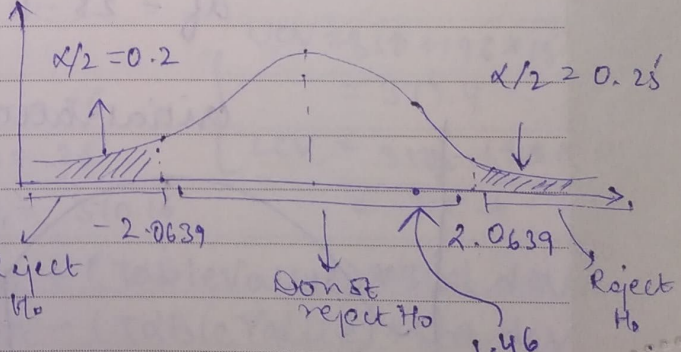
$\alpha = 0.05$ (error rate).

$$t \text{ test} \rightarrow \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

(fall inside boundary)

95% error & 95% → accurate.

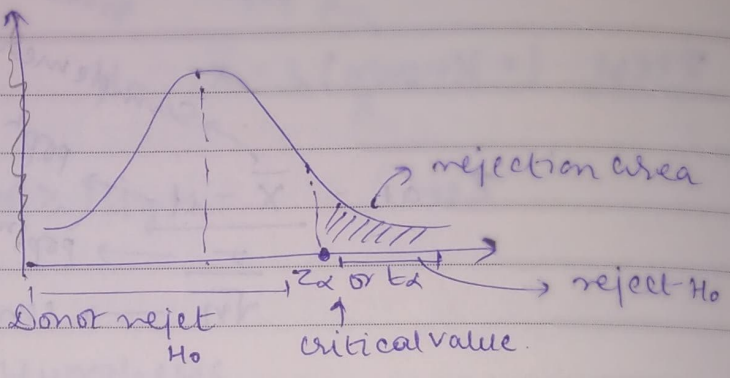
degree of freedom → $df = 25 - 1 = 24$ (no. of row − no. of col n)



(from table) critical value:

$\pm t_{24, 0.025} = \pm 2.0639$

- **One Tail test**
  only one side rejection is there.
  (either left side or right side)



Rejection area is in left side

Left tail test.

$H_0: \mu \geq 3$

$H_1: \mu < 3$

Lower-tail test since alternative hypothesis (H_1) is focused on the lower tail below mean of 3

Right tail test

$H_0: \mu \leq 3$

$H_1: \mu > 3$

rejection area is in right side.

Upper-tail test since the alt. hypothesis (H_1) is focussed on upper-tail above mean of 3.

Ex. $H_0: \mu \leq 52$ → Right-tail test

$H_1: \mu > 52$

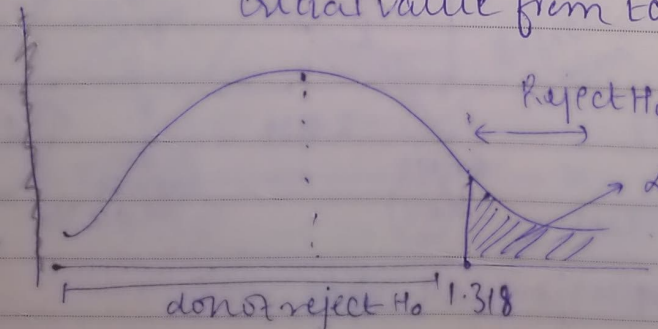
Given

$\alpha = 0.10$ (10% error → 90% accurate data).

$n = 25$

$df = 25 - 1 = 24$

critical value from table: ⁂ 1.318



Reject H_0 if $t_{stat} > 1.318$

* 95% → $\alpha = 0.05$ (~~α & χ test~~ table value = 1.96)

Thursday
February 2011
17

Note:
If α is not given ($\alpha = 0.05$ is default)

3.) Choose level of significance (α)

Level of significance is also called error Rate

$\alpha = 0.01$ (1% error, 99% accurate)

$\alpha = 0.05$ (5% error, 95% accurate)

$\alpha = 0.10$ (10% error, 90% accurate)

How much data is accurate.

Confidence Interval

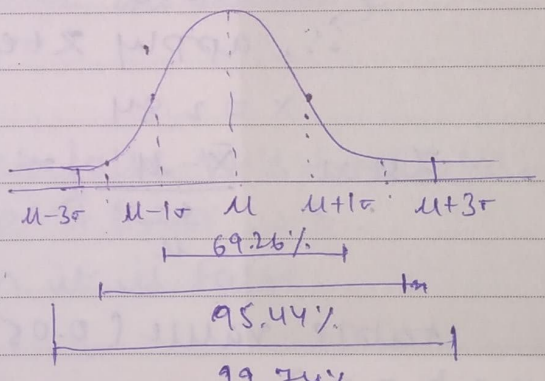
(limit) → Boundary defining (left / Right)

Method 1.

Empirical rule tells accuracy of model.

Ex.

mean consumption of various houses is 200 units,
SD = 20 units.



- 68.2% consume energy b/w 180 to 220 unit
- 99% have their energy consumption b/w 140 to 260 units

Method 2.

Using Formula.

$H_0 := 350$

$H_1 : \neq 350$

$\sigma = 75$

$n = 25$

$\bar{x} = 370.16$

UCV = 350 + 1.96 * 15
= 379.4

LCV = 350 − 1.96 * 15
= 320.6

* { • Upper confidence Interval = Mean + (Table Value) * Std. dev.
• Lower confidence Interval = Mean − (Table Value) * Std. dev

Degree of Freedom.

df = no. of row - no. of col^n.

4.) Critical value approach to Testing.

Ex.

$H_0: \mu = 3$

$H_1: \mu \neq 3$

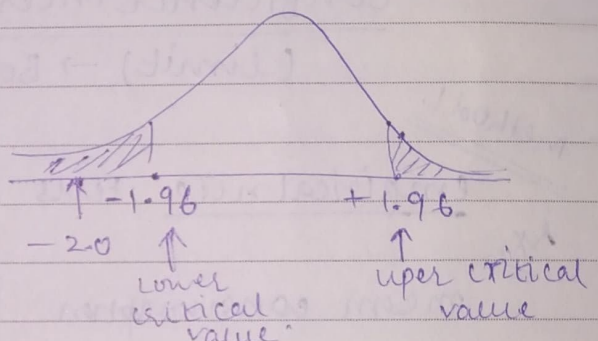
$\alpha = 0.05$

$n = 100$

$\sigma = 0.8$

∴, apply Z test

$\bar{x} = 2.84$



$Z_{stat} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = -2.0$ (fall outside the boundary).

∴, reject Hypothesis

table value ($0.05 \rightarrow \alpha$) $\rightarrow -1.96$

P-Value Approach (Prob. Value approach.)

- compare p-value with $\alpha = 0.05$

p-value < 0.05 (reject H_0)

p-value ≥ 0.05 (don't reject H_0)

If judgment is wrong, Type I or II error occur.

Decision based on sample	Truth about population: H_0 true	H_1 true
Reject H_0	Type I error	correct decision
~~Reject~~ Accept H_0	correct decision	Type II error.

True state \ Decision	Jailed	Set Free
Innocent	Type I	Correct
Guilty	correct	Type II

H_0 → "Person is innocent"

Q. How to control Type I & Type II error. ?
- increase α value,

- Type I error (α)
  - ★ Reject null hypo. (H_0) when it is True, considered serious type of error.
  - The prob. of a Type I error is α called level of significance of test.
  - set by researcher in advance.

- Type II error (β)
  - ★ Failure to reject false null hypothesis. The prob. of Type II error is β.
  - or Accept H_0 when it is False.

21 Monday

February 2011

Week 8 ▪ 52-313

February 2011

Wk	Mo	Tu	We	Th	Fr	Sa	Su
5		1	2	3	4	5	6
6	7	8	9	10	11	12	13
7	14	15	16	17	18	19	20
8	21	22	23	24	25	26	27
9	28						

if we want to find more var. relⁿ we go for regression.

Day 11.

Correlation.

Used to find the relationship b/w only 2 variables. (not for more than two variable).

how we find relⁿship b/w 2 variables.

→ By corelⁿ coefficient (r) or (R)

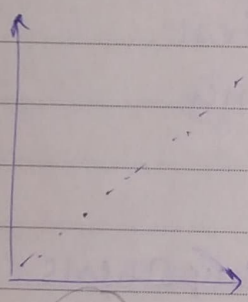
covariance

When we are changing one variable how much variation is there in other variable

standard dev.

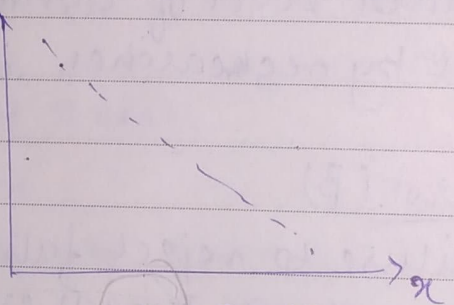
when want to measure consistency.

$$r = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})(y_i - \bar{y}) \right)}{\sqrt{\sum_{i=1}^{n} \left((x_i - \bar{x})^2 \cdot (y_i - \bar{y}) \right)}}$$



+ve corelⁿ

(directly proportional) to each other.



-ve corelⁿ

(inversely proportional) to each other

Graphically judge corelⁿ.

coefficient of Determination (R^2)

when we change one variable how much variation is there in another variable

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

Eg $r = -0.635$ $R^2 = .403$ (About 40% of variance is there in var.)

March 2011

Wk	Mo	Tu	We	Th	Fr	Sa	Su
9		1	2	3	4	5	6
10	7	8	9	10	11	12	13
11	14	15	16	17	18	19	20
12	21	22	23	24	25	26	27
13	28	29	30	31			

corelⁿ should always be considered with significance levels (p-values). If corelⁿ is not significant, it is of little use.

Tuesday 22 February 2011

Week 8 ▪ 53-312

Correlation Matrix.

find corelⁿ b/w one var. with another var. in the multiple combinations in one shot.

* same variable taken in row & colⁿ
diag always = 1.
Below diagonal is filled
above diag. is empty / get same result.

Corelⁿ Classification.

1. Product - Moment Correlⁿ
2. Partial correlation.
3. Non-metric corelation

1. Product moment corelation

→ Both var. are continous.

Ex (Attitude Vs Duration)

$$r = \frac{\text{covariance } x,y}{\sqrt{SD_x \cdot SD_y}}$$

covariance formula:

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

Standard deviation of x

$$\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)$$

y.

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

- if r=0, there is no relationship b/w variables.
- if r=1 which means if one value ↑, the other ↑ in same proportion.
- if r=-1 which means, if one value ↓, the other ↑ in same proportion.

r → Pearson correlⁿ coefficient or corelⁿ coe

strong ← weak | weak → strong

+1.0 | 0 | +1.0

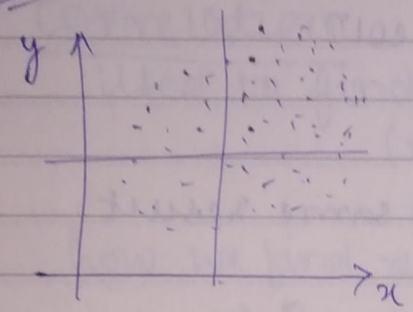
-ve corelⁿ | +ve corelⁿ

$-1 < r < 1$

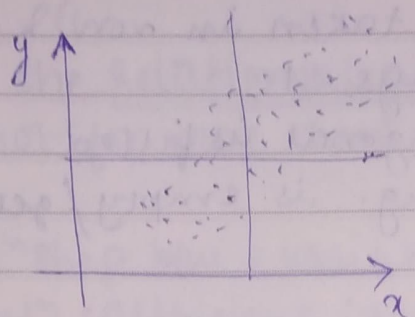
perfect -ve colinearity. | perfect colinea

NOTE { +ve corelⁿ → graph in upward direct^n
-ve corelⁿ → " " downward "

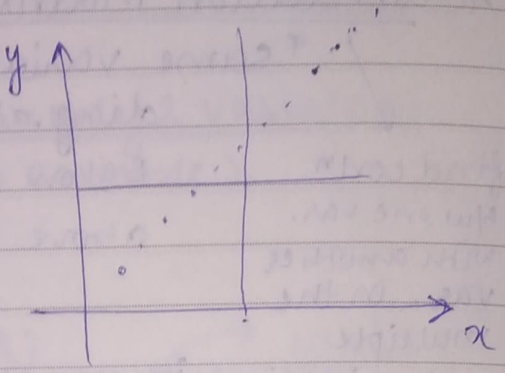
Ex. of r.



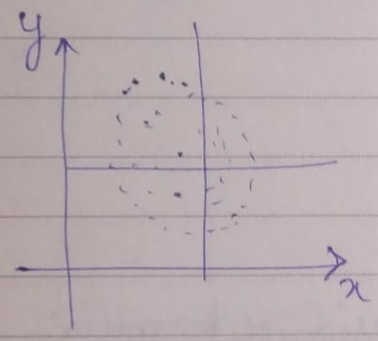
r = 0.30
low +ve corelⁿ



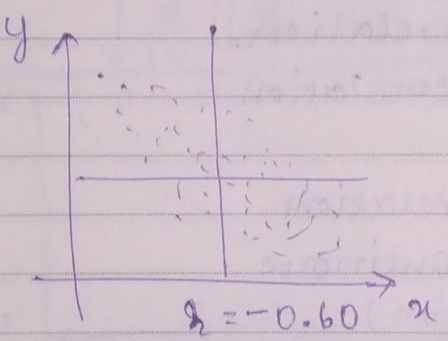
r = 0.80
High +ve corlⁿ



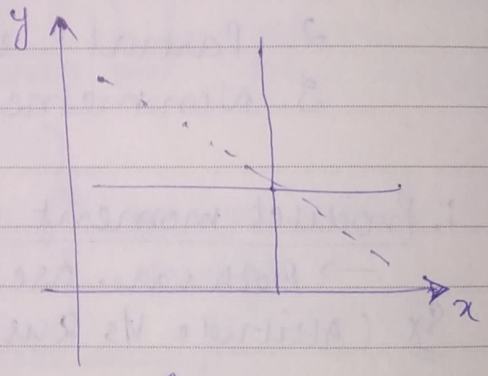
r = 1.0
Perfect +ve corlⁿ.



NO corelⁿ
r = 0



r = −0.60
Moderately
-ve corelⁿ



Perfect -ve
corelⁿ

* Null hypothesis → always be the -ve statement.

Ex. $H_0 : \rho = 0$

$H_1 : \rho \neq 0$ → $\rho^2 = 0$

1.) There is no corelⁿ b/w attitude & Duration

2.) It is two tail test.
- there is no population SD given ∴ it is t-test.

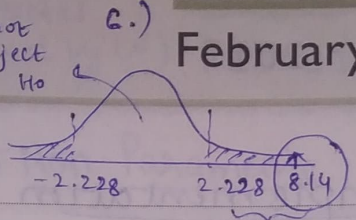
3.) $\alpha = 0.05$ (5% error, 90% accurate data)

df = row - colⁿ → 12 − 2 = 10

Thursday
February 2011
24
Week 8 ▪ 55-310

compute calculated test.

6.)



4.) $$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2}$$

$$= 0.9361\left[\frac{12-2}{1-(0.9361)^2}\right]^{1/2}$$

$$= 8.144$$

(r lies outside boundary).

- hence null hypothesis of no relnship b/w X & Y is rejected.

5.) calculate table value by using df & $\alpha = 0.05$.
→ 2.228

7.) H_1 is accepted.

∴, there is significant corelⁿ b/w attitude & Duration.

Partial corelation

one var → continous
other var → categorical.

Ex sales V/s Product
(contin) (category)

Non-metric corelation

Both var → categorical

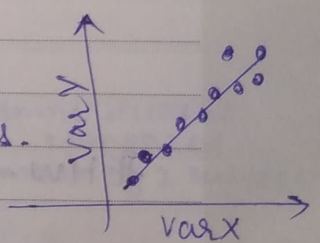
Ex gender V/s Product

Regression

estimating magnitude of relⁿ b/w variables.

more than one variable.

- relⁿ is defined as:
  var1 = function(var2).
- Based on this relⁿ, value of one var. corresponding to a particular value of other var. can be estimated.



Practical Implementation in Python.

~~dat~~ df = pd.read_excel('correlation.xlsx', sheet_name=1)
df → (new dataset).

from scipy.stats import pearsonr → fn name for corln.

stats, p = pearsonr(df.Attitude, df.Duration).

print(stats, p)

0.9360778..., 7.545161167...e-06

r value — p-value

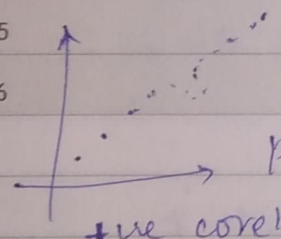
check

7.545161...e-06 - 6

0.0000075 which is less than 0.05

p < 0.05

∴, Ho is rejected
H1 is accepted



plt.scatter(df.Attitude, df.Duration)

df.corr() # Make corelⁿ matrix.

HW → find corelⁿ with Attrition & other variables.

Statistical Test

(diagram + definations → written in day9).
refer day9 Notes.

March 2011

Wk	Mo	Tu	We	Th	Fr	Sa	Su
9		1	2	3	4	5	6
10	7	8	9	10	11	12	13
11	14	15	16	17	18	19	20
12	21	22	23	24	25	26	27
13	28	29	30	31			

Check examples from notes (provided by LU)

Saturday 26 February 2011

Week 8 ▪ 57-308

Non-Parametric

- **Wilcoxon-Sign Test**
  Ex Same people → 2 different situation.
- **Friedman Test** → compare more than 2 paired samples
  Ex same people → 3 different situation (more than 2.)
- **Mann Whitney Test** → compare 2 independent samples.
  20-Mrng, 20-Evening
  20 → Mrng
  20 → Afternoon
  20 → Evening.
  } more than 2 indepen.
  - **Kruskal-Wallis**

9 10 11 12 1 2 3 4 5 6

- **Chi-square** → applicable only for categorical variable. Gender: M, F
  Smoker: Postsmoker, Current smoker, Non smoker.

Parametric

Sunday 27 58-307

- **One Sample Test**
  compare sample mean with population Mean.
- **Sample paired t test.**
  compare mean of 2 paired samples
  Ex = same student compared mean of 2 subjec
- **Sample independent t-test.**
  compare mean of 2 indepⁿ samples
  Ex Marathon Ex.
  = Mean of Athlethe & Non-Athlethe. duration.

28

Monday

February 2011

Week 9 ▪ 59-306

February 2011

Wk	Mo	Tu	We	Th	Fr	Sa	Su
5			1	2	3	4	5	6
6	7	8	9	10	11	12	13
7	14	15	16	17	18	19	20
8	21	22	23	24	25	26	27
9	28						

Note. Stats value → critical value.
(every test has critical value & Table value).
calculated value → critical value.

Day 12.

Practical implementation of Non-parametric.

Wilcoxon Test.

H_0 : There is no significant difference in calcium level of patient initial & after 2 weeks.

```
df = pd.read_excel('Wilcoxon.xlsx', sheet_name = 0)
df.head()
```

```
from scipy.stats import wilcoxon
stats, p = wilcoxon(df.TOTALCIN, df.TOTALCW2)
print(stats, p)
```

↓ ↓

29.5 0.0025

p < 0.05 → reject H_0 accept H_1 ✓✓

Friedmann Test.

```
from scipy.stats import friedmanchisquare
stats, p = friedmanchisquare(df.TOTALCIN, df.TOTALCW2,
                             df.TOTALCW4)
print(stats, p)
```

↓ → 8.62133 e-07

27.9277

e-07 : exponential power

p < 0.05 ∴, H_0 rejected, H_1 accepted

Mann-Witney Test

```
df1 = pd.read_excel('manwhitney.xlsx', sheet_name = 1)
df1.
```