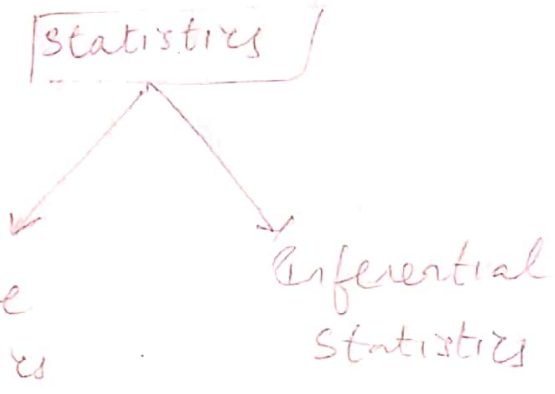
Statistics

* Statistics is the science of collecting, organizing & analysis data

* Data → facts or pieces of Information

Ex: Age of students in your class
30, 25, 60, 90, ... etc



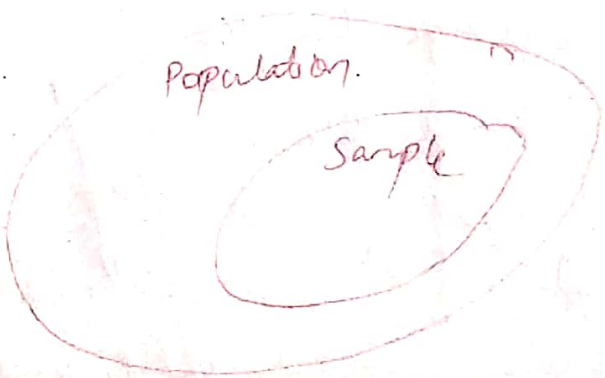
Descriptive: consist of organizing & summaring data

Inferential: consist of using data you've measured to form conclusions.

Population & sample

(N) Population: The group you're are intersted in studying

(n) Sample: a subset of population



Sampling methods

* Simple Random Sampling : Every member of population (N) has an equal chance of being selected for your sample

* Stratified Sampling : The population (N) is split into non overlapping groups (Strata) then simple random sampling is done on each group to form a sample (n)

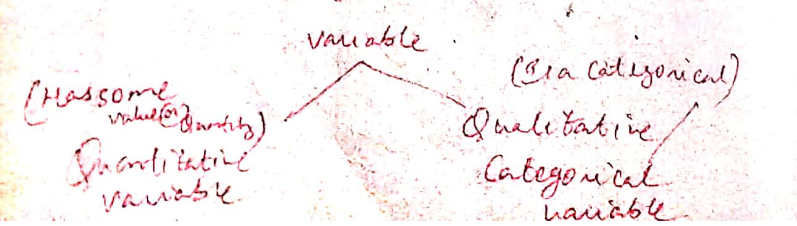
* Systematic Sampling : Every nth member of the population (N) is selected as a sample (n)

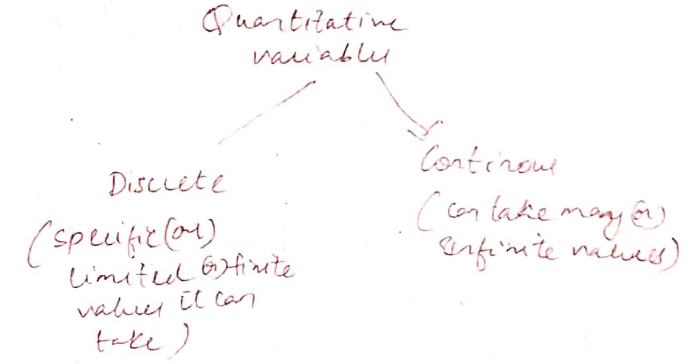
* Convenience Sampling : Easily obtained individuals from population (N) are placed in the sample (n) also called (voluntary response sampling)

Types of Variable

Variable : A property which take (or) many any value in

Ex: Age, Gender





Independent variable: It any variable that is being ~~Manipulated~~ Manipulated (or) is changing

Dependent Variable: It any variable that is being measured.

Variable Measurement Scales

Nominal
Ordinal
Interval
Ratio

↓ less Information

↓ More Info

Nominal Data → It data which is split into categorical types is called Nominal data

Ordinal Data → It data ie in which order matter but distance Between values ~~on~~ does not (or) can't be measured

Interval → Order matters, & Distance between it equal & meaningful → Natural zero is not present Ex: temperature.

Ratio: Order matters, distance between is equal & meaningful, And a natural zero does exist

Frequency distribution: lists each measured category & number of occurences of occurences for each category

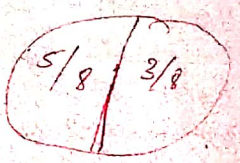
Cummilative distribution; Adding of occured Frequency frequencies is called cummilative frequency

Representation of Qualitative Data (Mostly discrete)

* Bar graphs: X axis → Character, Y → Occurence (no contact btw bars)
* Pie Charts.

Via circle divided into sector, each sector represent a category of data in proportion to the total amount of data collected

Ex: Pizza & pieces
Friend took 3/8
you took 5/8



Histogram Representation of Quantitative

* Histogram → Bars are in contact (used for discrete & continous)
* Stem & Leaf Plots

For continous histogram we convert into classes & set up a Interval.

Marks	Freq
1-2	1
3-4	2

1 → lower limit
2 → upper limit
Class width = upper limit − lower limit

To Decide classes Ex: we have data of 30 students marks ranging from 1 to 50

50/(Desired Classes (10)) = (5) → class width.

Arithmetic Mean = Sum of observation / Number of observations

Population Mean (μ) — Sample Mean ($\bar{x}$)

$\bar{x} = \frac{\Sigma x}{n}$ (Number of samples)

$\mu = \frac{\Sigma x}{N}$ (size of Population)

Median: 1, 1, 2, 2, 2, 3, 3, 4, 5, 5

$\frac{2+3}{2} = \frac{5}{2} = 2.5$

Mode: Most Repeated: 2

Bimodal : Two modes ; Multimodal : More than 2 Modes

Mean is effected by Outliers
Median should be used for central tendency. (or) Mode depending upon Question (or) we can remove outliers.

Measures of Dispersion

Dispersion refers to how spread out a dataset is about the mean

* Variance (σ^2)
* Standard Deviation (S.D) (σ)

Definition Formula Population Variance

$$\sigma^2 = \frac{\Sigma (x-\mu)^2}{N}$$

Definition Formula Population S.D (σ)

$$\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$

Computation Formula Population Variance (σ^2)

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad ; \quad \sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}}$$

Computational Formula Population S.D (σ)

Definational Formula Sample Variance (σ^2)

$$s^2 = \frac{\Sigma (x-\bar{x})^2}{n-1}$$

Definational Formula Sample S.D (σ)

$$s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

Computational Formula Sample Variance (σ^2)

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

Computational Formula Sample S.D (σ)

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

Percentile & Quartiles

Percentages = $\frac{\text{Meeting character of Interest}}{\text{Total no of observations}} * 100$

Dataset = 1, 2, 3, 4, 5

$\frac{2}{5} \times 100 = 40\%$

↑ Percentile is a value below which a certain percentage of observations lie

Ex: 2, 2, 3, 4, 5, 5, 5, 7, 8, 8 ... 10, 11, 11, 12

Percentile Rank of x = $\frac{\text{values below } x}{n} \times 100$

Percentile Rank of 10 = $\frac{16}{20} * 100 = 80\%$.

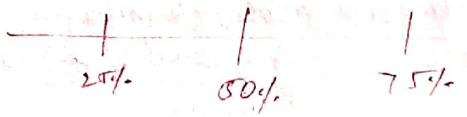
What value exists at the percentile ranking of 25%

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100}(20+1) = 5.25$$

5.25 so we take average of 5th & 6th

Quartiles:



First Quartile - 25% - Q_1

Second Quartile - 50% - Q_2

Third Quartile - 75% → Q_3

Five Number Summary: is a Method of summarizing a distribution of data

* Minimum
* First Quartile Q_1 (25%)
* Median (50%)
* Third Quartile (Q_3) (75%)
* Maximum.

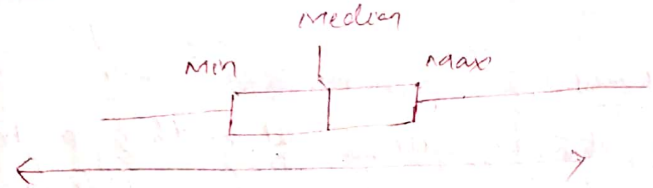
Lower Fence = $Q_1 - 1.5(IQR)$

Upper fence = $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

Outliers can be removed If they dont fall between Upper Fence & lower fence.

Boxplot:

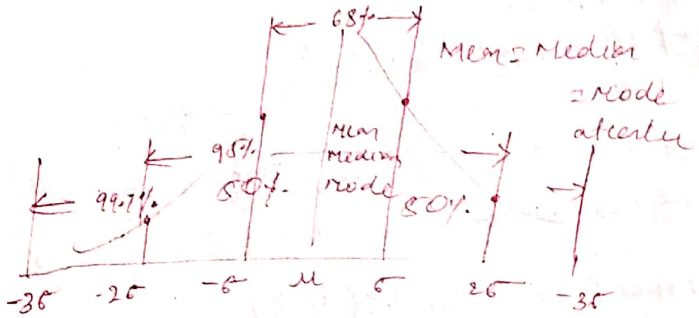


Skewness: How a distribution of data leans

Left skew ; Right skew, No skewness

Skewness is important to recognise because it has implication in hypothesis testing.

Distributions of most continous random variables will follow the shape of normal curve



Graph changes direction at inflection points

Empirical Rule:

$-\sigma$ to σ → 68% of data

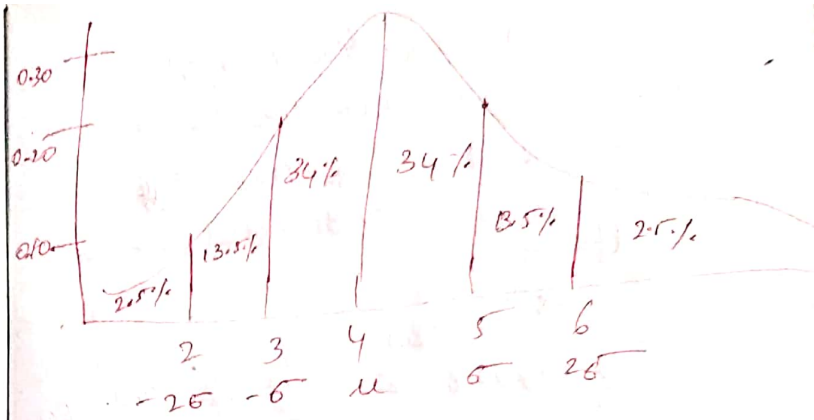
-2σ to 2σ → 95% of data

-3σ to 3σ → 99.7% of data

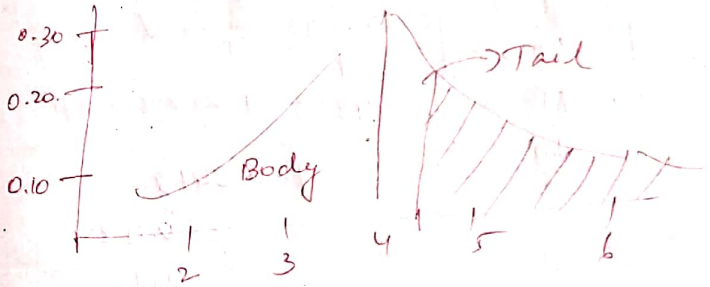
Z-scores

Are standardised values that can be used to compare scores in different distributions.

$$Z = \frac{x - \mu}{\sigma}$$



Let's say we have random variable 'x' distributed as such



What percentage of scores fall above 4.25?

$$Z = \frac{x - \mu}{\sigma} = \frac{4.25 - 4.00}{1} = 0.25$$

↑ distance from Mean

Area = 0.5987 in Body

$1 - 0.5987 = 0.4013$

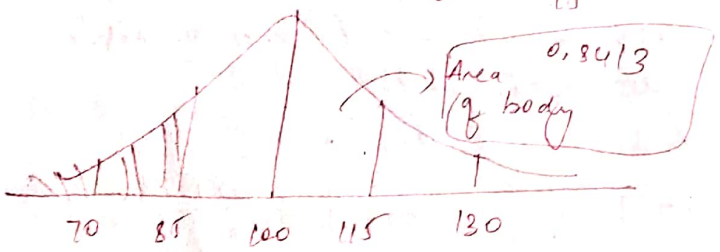
In united states Avg IQ is 100 with S.D of 15
What percentage of the population would you expect to have an IQ (a) lower than 85
(B) Between 90 & 120

$\mu = 100$

$\sigma = 15$

(a)

$z = \frac{x - \mu}{\sigma} = \frac{85 - 100}{15} = -1.0$

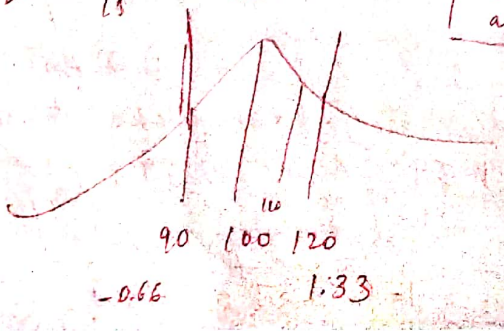


$1 - 0.8413$

$= 0.1587$

(b) $z_2 = \frac{90 - 100}{15} = \frac{10}{15} = -0.66 =$ Body area 0.7454

$z_2 = \frac{120 - 100}{15} = \frac{20}{15} = 1.33 =$ 0.9082 Body area



2, Area under Tail 25.46%

2, Area under Tail 9.18%.

Between area is 65.36%.

25.46
9.18
34.64

1.00
100 00
34.64
65.36

Probability : Is a measure of the likelihood of the event.

Prob = no of ways an event can occur / no of possible outcomes

(S)

Sample Space : Collection of All possible outcomes.

Probabilities will always be between 0 & 1
An event with a probability 0 is impossible
An event with a probability 1 is certain.

Addition Rule

Two events are "mutually exclusive" if they cannot occur at same time.

Ex: Rolling a dice and getting 6 & 1 at a time.

For mutually exclusive events (Addition Rule)

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

Probability of choosing a card that is heart or a king.

$$P(\text{Heart}) = \frac{13}{52} \quad ; \quad P(\text{king}) = \frac{4}{52}$$

$$P(\text{king \& Heart}) = \frac{1}{52}$$

For Non Mutually exclusive events (Addition Rule)

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

$\cap \rightarrow$ 'and' ; $\cup \rightarrow$ 'or'

$P(\text{Heart \& king}) = 0.308$

Multiplication Rule

Two events are independent if they do not effect one another.

Ex: Rolling a '5' & then Rolling a '6'

Dependent: Two events are dependent if they do affect one another.

Ex: Drawing a king & then drawing a queen from deck of cards without putting the king back.

What is probability of rolling a 5 & then a 3 with a normal six sided die?

$$P(A \cap B) = P(A) * P(B)$$

$$P(3 \text{ and } 5) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

What is probability of drawing a king & then drawing a queen from a deck of cards

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$P(\text{King and Queen}) = \frac{4}{52} \times \frac{4}{51} = \frac{16}{2652}$$

$$= 0.6\%$$

Multiplication rule for Dependent events is Conditional probability.

Permutation :

Ex: You visit a zoo with 6 animals & ask you to record the first 3 animals you see

Tiger, Lion, Monkey, Zebra, Walrus, Snake

$6 * 5 * 4 = 120$

With permutations order matters.

Total no of object $\leftarrow {}_nP_r = \frac{n!}{(n-r)!} = {}_6P_3 = 120$

$r \leftarrow$ no of objects you are picking.

Combinations : With combination order doesn't matter

Total no of objects $\leftarrow {}_nC_r = \frac{n!}{(n-r)!\,r!}$ ${}_6C_3 = 20$

$r \leftarrow$ no of objects you picking

Random Variable : Is a variable which has its value determined by a probability experiment.

Discrete Random Variable : Is a random variable which has a finite no of values.

Continous Random Variable : Is a random variable which has an infinite no of values

Probability Distribution : Displays probabilities associated with all possible outcomes

Ex:-
Probability Distribution for one roll of a six sided die?

Number	P(Number)
1	1/6 = 0.167
2	1/6 = 0.167
3	1/6 = 0.167
4	1/6 = 0.167
5	1/6 = 0.167
6	1/6 = 0.167

Probability Histogram : Is a histogram with possible values on the x axis, & probabilities on y axis.

x	P(x)
1	0.10
2	0.15
3	0.35
4	0.40

Right skew

Mean & Expected Value of Discrete Random Variable :-

Discrete Random Variable

x = Number of strokes to complete course

Below is probability distribution for a golfer on a par 3 hole

x * P(x)	x	P(x)
0.10	1	0.10
0.60	2	0.30
1.35	3	0.45
0.60	4	0.15

Mean formula :-

$\mu_x = \Sigma [x * P(x)]$

$\boxed{\mu_x = 2.65}$

The Mean we just calculated of 2.65 is an expected value.

If we were to take a large enough sample of this golfer's performance on par 3 holes we expected his mean to approach 2.65

This is an Example of law of large numbers.

Variance & S.D of Discrete Random Variable

Variance Formula

$$\sigma^2 = \Sigma [x^2 * P(x)] - \mu_x^2$$

x	P(x)	x^2	$x^2 * P(x)$
1	0.10	1*1 = 1	0.10
2	0.30	4	1.20
3	0.45	9	4.05
4	0.15	16	2.40
			7.75

$\mu_x = 2.65$

$\sigma^2 = 7.75 - (2.65)^2$

Variance $\boxed{\sigma^2 = 0.73}$

S.D $\boxed{\sigma = 0.05}$

Law of large Numbers

As a probability experiment is performed many times, the observed value (mean) will arrive at the expected value (mean)

Ex: Suppose we toss a coin once & find probability of getting head is

$P(H) = \frac{1}{2} = 0.5 \rightarrow$ Expected mean value

If a coin is tossed ∞ times then the overall value of the probability of head will eventually reach 0.5 i.e which is expected

Binomial Distribution: An experiment is a binomial Distribution if

* It is repeated fixed number of times
* Trails are independent
* Trails have mutually exclusive outcomes either success (or) failure
* Probability of success is same for all trails.

$$P(x) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

In a recent survey we found 85% of houses have high speed Internet. If you take sample of 18 households, what is the probability that exactly 15 will have high speed Internet

$$P(x) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

$n = 18$; $x = 15$; $\boxed{p = 85\% = 0.85}$

$$P(15) = {}^{18}C_{15} (0.85)^{15} (1-0.85)^{18-15}$$

$$= 0.239$$

Atleast 15

$$P(X \geq 15) = P(15) + P(16) + P(17) + P(18)$$

$$P(x \geq 15) = 0.718$$

Mean & S.D of binomial Distribution

$$\mu_x = np \quad ; \quad \sigma_x = \sqrt{np(1-p)}$$

$$= (18)(0.85) \qquad \sigma_x = \sqrt{(18)(0.85)(1-0.85)}$$

$$\boxed{\mu_x = 15.3} \qquad \boxed{\sigma_x = 1.515}$$

Poisson Distribution

Is used when computating probability of a certain number of success within a specified interval.

An Experiment follows Poisson if:-

* Probability of two successes in a small enough interval is 0%
* Probability of a success is the same for any two intervals which share the same length
* Success are independent of successes in other intervals.

At a theme park, there is a roller coaster that sends an average of 3 cars through its circuit every minute between 6pm & 7pm. A random variable, X, represents the number of roller coaster cars to pass through the circuit between 6pm & 6:10pm

What is probability that 35 cars will pass through the circuit between 6pm & 6:10pm

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

(x: success; t: time interval; e: euler's constant)

$\lambda = 3$, $x = 35$, $t = 10$ minutes

e = constant = 2.718

$$P(35) = \frac{((3)(10))^{35}}{35!} \times e^{-(3)(10)}$$

$$P(35) = 0.045$$

Mean & S.D of Poisson Random Variables

Mean: $\mu_x = \lambda t$

S.D: $\sigma_x = \sqrt{\mu_x}$

x = number of success = 35

t = a length of time = 10 minutes

λ = avg no of success in interval = 3

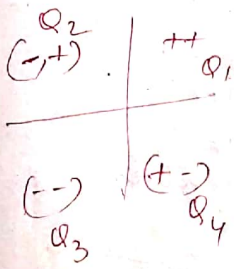
$e = 2.718$

$\mu_x = (3)(10)$

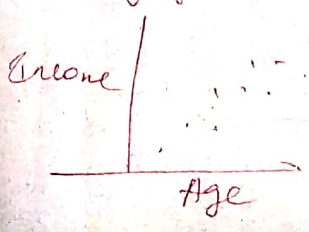
$\boxed{\mu_x = 30}$

$\sigma_x = \sqrt{30}$

$\boxed{\sigma_x = 5.477}$



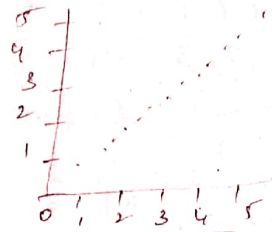
Scatter Plot: Are a method of graphically displaying bi-variate data.



Pearson's Correlation (r)

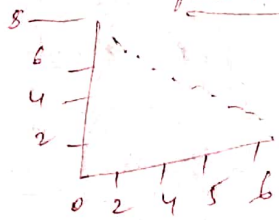
It measures strength of linear relationship between two variables

Pearson's (r) is always between -1 & 1



r = 1.0 → Perfect +ve correlation

r = -1.0



Perfect -ve correlation

r = 0

→ No correlation

Pearson's r

subject	(x) Age	(y) Income	(x)×(y)
1	18	15,000	270,000
2	25	29,000	725,000
3	57	68,000	3,876,000
4	45	52,000	2,340,000
5	26	32,000	832,000
6	64	80,000	5,120,000
7	37	41,000	1,517,000
8	40	45,000	1,800,000
9	24	26,000	624,000
10	33	33,000	1,089,000
	369	421,000	18,193,000

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

$$r = 18193000 - \frac{(369)(421000)}{10}$$

$$\sqrt{15629 - \left(\frac{36161}{10}\right)} \cdot \sqrt{21289000000 - \left(\frac{177241000000}{10}\right)}$$

$$= 0.99$$

Strong correlation

Pearson's 'r' Hypothesis testing ('r')

Researcher collect 10 individual data to test if age & Income are related

$\alpha = 0.05$

S.no	Age	Income
1	18	15k
2	25	29k
3	57	68k
4	45	52k
5	26	32k
6	64	80k
7	37	80k
8	40	41k
9	24	26k
10	33	33k

- $H_0 : \rho = 0$
- $H_1 : \rho_1 \neq 0$

$\alpha = 0.05$

$df = n - 2$

$10 - 2 = 8$

using r table

using df, α → column

↓ Row

$r_{0.05} = 0.632$

Using 'r' formula used find 'r' = 0.99 So we reject null Hypothesis

Spearman's Correlation (r_s)

So far Pearson's 'r' which measures the relationship between two continous (interval (or) ratio) variables

Spearman's correlation is used when

* Measuring the relationship between two ordinal variables
* Measuring relationship between two variables that are related but not linearly.

X	Y
2	21
5	17
8	14
11	10
15	5
16	3

So lower score is X has lowest rank & similarly in Y lowest has lowest rank

* To calculate spearman's correlation, we must first rank scores

X	Y
1	6
2	5
3	4
4	3
5	2
6	1

Now we find 'r's

$$r_s = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\sqrt{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} * \sqrt{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}}$$

$$r_s = \frac{56 - \frac{(21)(21)}{6}}{\sqrt{91 - \frac{(21)^2}{6}} \cdot \sqrt{91 - \frac{(21)^2}{6}}}$$

$$\boxed{r_s = -1.00}$$

Perfect -ve Relationship

The Reason we find correlation is to find Linear Regression.

Linear Regression

If we know that two variable are strongly correlated we can use one variable to predict other

Linear Regression Function $\leftarrow \boxed{\hat{y} = \beta_1 x + \beta_0}$ ↑ y intercept & Regression line

slope eq Regression line → $\boxed{\beta_1 = r \frac{S_y}{S_x}}$ $\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$

$S_y \rightarrow S.D(\sigma_y)$

$S_x \rightarrow S.D(\sigma_x)$

Stronger your correlation (that is, the closer r is to -1 (or) 1), the more accuracy your prediction will be.

S.No	x	y
1	18	15k
2	25	24k
3	57	68k
4	45	52k
5	26	32k
6	64	80k
7	37	41k
8	40	45k
9	24	26k
10	33	33k

Sol from previous Karl Pearson we get $r = 0.99$

$$S_x = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1}} = \sqrt{\frac{15629 - \frac{13616}{10}}{10-1}} = 14.96$$

$$S_y = \sqrt{\frac{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}{n-1}} = 19902.26$$

$$\beta_1 = (0.99)\frac{19902.26}{14.96} = 1317.06$$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$ ($\bar{y}$ → y Mean, $\bar{x}$ → x Mean)

$\beta_0 = (42100) - (1317.06)(36.9)$

$\beta_0 = -6499.51$

$\hat{y} = 1317.06x - 6499.51$

For Ex)

Predict yearly Income of someone who is 33 yr old

$\hat{y} = (1317.06)(33) - 6499.51$

$= 36963.47$

Correlation vs. Causation

Causation mean that one variable causes something to happen in another variable.

To say that two things are correlated is to say that they share some kind of relationship

In order to imply causation, a true experiment must be performed where subjects are randomly assigned to different conditions.

Parameter, statistics & sampling error

Parameter: A characteristic that describes a population

Because it's not possible to measure whole population. parameters are most often estimated.

Ex Population mean μ, Population variance σ^2

For instance we have to measure mean of height of people on earth that could be a parameter but that's impossible ; ie why we measure statistic.

A characteristic that describes a sample is called a statistic

Statistics are most often used to estimate the value of unknown parameters.

Ex Sample Mean $\bar{x}$; Sample Variance s^2

Sampling error: It any difference that exist between statistic & its corresponding parameter.

EX: We use statistic to find an estimate population mean ie 68 However when do manually we find it is 70 so this dispersion is called sampling error. $70 - 68 = 2$ → Sampling error is 2

Distribution of Sample Mean (statistic)

Is a probability distribution for all possible values of a sample mean, computed from a sample of size of n'

Ex: A statistics class has 6 students, ages displayed below. Construct a sampling distribution of the mean of age for samples ($n=2$)

Ages: 18, 18, 19, 20, 20, 21

Sample	Sample Mean
18,18	18
18,19	18.5
18,20	19
18,20	19
18,21	20
18,19	18.5
18,20	19
18,20	19
18,21	19.5
19,20	19.5
19,20	19.5
19,21	20
20,20	20
20,21	20.5
20,21	20.5

All possible combinations of samples of size $n=2$

Sample mean	Frequency	Probability
18	1	1/15
18.5	2	2/15
19	4	4/15
19.5	3	3/15
20	3	3/15
20.5	2	2/15
	15	

Mean $= \frac{18+18+19+20+20+21}{6}$

$\mu = 19.33$ → So, its accurate

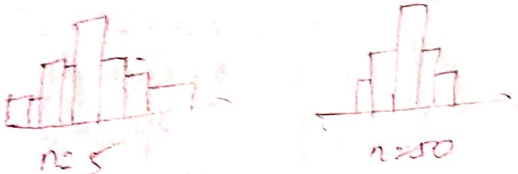
From above most sample Mean method we did we will get population mean close to 19 & least population Mean is 18 so, 1 sec mean

Remember large our sample size is the closer our sample mean should be to the population mean. This is Law of large numbers.

Large sample better population estimate.

Standard Error of Mean

There are sampling distribution of Mean Below



As sample size increases, S.D decreases ($\sigma_{\bar{x}}$)

Large sample size means less dispersion

Standard deviation of sampling distribution is also known as the standard error of the mean

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ → S.D of population

$\sigma_{\bar{x}}$ → Standard error of mean

n → sample size

Central Limit Theorem (CLT)

It states that regardless of the shape of the population distribution, the distribution of sample means will be approximately normal.

The distribution of sample mean will become more normal as its sample size increases

Good rule of thumb: sample distributions will be approximately normal if their sample size is $n = 30$ (or) larger

i,e Sample Distribution always will be normal distribution. → (CLT)

Sample Proportion

Let say we want to measure how many left handed exist in a population obviously it is impossible so we take a sample of 500 people & create a proportion.

In our sample, we found 75 are left handed

$\mu_{\hat{p}} = \hat{p} = \frac{x}{n} = \frac{75}{500} = 0.15 = 15\%$ of population is left

Distribution of Sample Proportion

We can find out the distribution of the sample proportion if the sample size is less than 5% of the total population size

If $np(1-p) \geq 10$, the distribution of sample proportion is approximately normal

S.D of Sampling Distribution

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Ex: In a sample of 500 individuals, 75 are left handed Describe distribution of the sample proportion

Sample size

Yes 500 is less than 5% of entire world population.

$$500(0.15)(1-0.15) \geq 10$$

$$63.75 \geq 10 \text{ so normal}$$

S.D

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.15)(1-0.15)}{500}} = 0.016$$

$\mu_{\hat{p}} = 0.15$ $\sigma_{\hat{p}} = 0.016$

Confidence Intervals about the Mean, Population, S.D known :-

Point estimate

Value of any statistic that estimates the value of a parameter is called a point estimate

Ex:

$\bar{x} \longrightarrow \mu$

$\bar{x} = 2.95$ $\mu = 3.00$

We rarely know if our point estimate is correct because it is merely an estimation of the actual value.

Confidence Intervals

Because of this discrepancy, we construct confidence intervals to help estimate what the actual value of the unknown population mean is → Point Estimate ± Margin Error

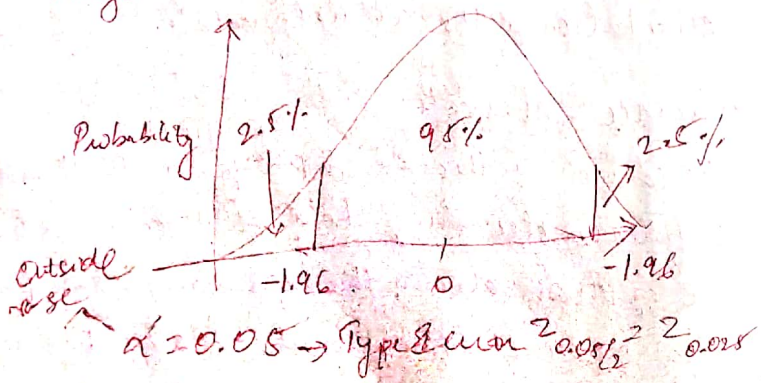
first of all how to

* How confident we want to be with our assessment
* Population Standard deviation
* How large our sample is.

Confidence Intervals about the mean

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

let say we have a distribution like this & we want 95% to be confidence & a we take a chance of 5% of it to be outside the range



outside area

$\alpha = 0.05 \rightarrow$ Type I error $z_{0.05/2} = z_{0.025}$

Ex: On the verbal section of the SAT, S.D is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% confidence interval about the mean

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad z_{\alpha/2} = z_{0.025}$$

$\sigma = 100$

$\bar{x} = 520$

$$520 \pm (1.96) \frac{(100)}{\sqrt{25}} = 480.8, \; 559.2$$

480.8 ↓ Lower Bound; 559.2 ↓ Upper Bound

ie I'm 95% confident that mean SAT score is between 480.8 & 559.2

Calculating Required sample size to Estimate Population Mean :-

We can calculate what sample size we will need in order to have a certain margin of error

$$n = \left(\frac{(z_{\alpha/2})(\sigma)}{E} \right)^2$$

Ex: On the verbal section of the SAT s.d is known to be 100. What size sample would we need to construct a 95% confidence interval with a margin of error of 20?

$$n = \left(\frac{(1.96)(100)}{20}\right)^2 = 96.04$$

A sample size of 97 is needed to create a 95% confidence interval with a margin of error of 20.

Student's t- Distribution

When performing any type of test (or) analysis using z-score it is required that σ (population S.D) is already known.

In real life, this is hardly ever the case. It is almost always impossible for us to know the S.D of the population from which our sample is drawn.

So HO we perform an analysis when we don't know S.D of population so we use Student's t- Distribution

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad ; \quad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Degree's freedom

As you can see formula is same major difference comes from Degree of Freedom

Student's t-Distribution has n-1 degree of freedom

Remember we are no longer given population standard Deviation. Instead we must estimate it with Sample S.D

Sample S.D is itself is a random variable. by calculating Sample S.D, it is given a fixed value & thus one less value is free to vary.

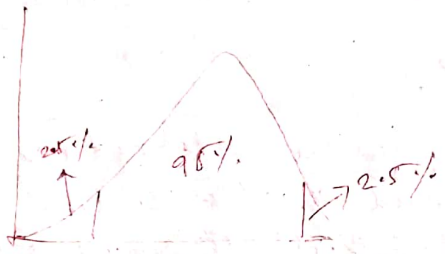
D.F change how the probability distribution looks. Probability distribution of t has more dispersion than normal distribution

so we t-Table

Confidence Intervals about Mean, Population S.D unknown:

On the verbal section of the SAT, a sample of 25 test takers has a mean of 520 with S.D of 80. Construct 95% Confidence interval about the mean.

$$\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)$$



$\alpha = 0.05$ $\quad t_{0.05/2} = t_{0.025}$

because it's t-test

$df = n - 1 = 25 - 1 = 24$

→ Column (24) Row

$t = 2.0639$

$520 \pm 2.0639 \left(\frac{80}{\sqrt{25}}\right) = 486.978 \;\; 553.022$

Lower Bound, Upper bound

95% confidat mean SAT score lie Betwee

Confidence Interval for Population proportion

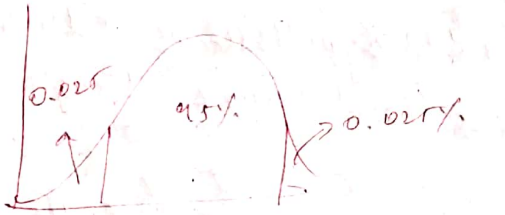
In a recent poll of 200 households, it was found that 152 households had at least one computer. Estimate the proportion of households in the population that have at least one computer

$$\mu_{\hat{p}} = \hat{P} = \frac{152}{200} = 0.76$$

$$\sigma_{\hat{p}} = np(1-p) = (200)(0.76)(1-0.76) \geq 10$$

$$36.48 \geq 10$$

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad z_{0.025} = 1.96$$



$$0.76 \pm 1.96 \sqrt{\frac{(0.76)(1-0.76)}{200}}$$

95% confidence: 0.701 (lower Bound) ; 0.819 (Upper Bound)

Type I & Type II :-

Type I : We reject Null hypothesis when in reality it is true Type I

Type II : We retain Null Hypothesis when in reality it is false
Type II

Effect Size :-

It a measure of the strength of an effect

After doing statistical analysis if we reject null hypothesis we calculate effect size to determine strength of the effect

Cohen's d = Mean difference / S.D

Different sample take in each analysis

Cohen's d = (120 - 100) / 30 = 0.66

If we have 2 cases

Ex: Medicine A → mean 120
Medicine B mean → 100

If 'd' lowers → 0.2 → small effect
0.5 → medium effect
0.8 → large effect

Power :

It the probability of correctly rejecting Null hypothesis

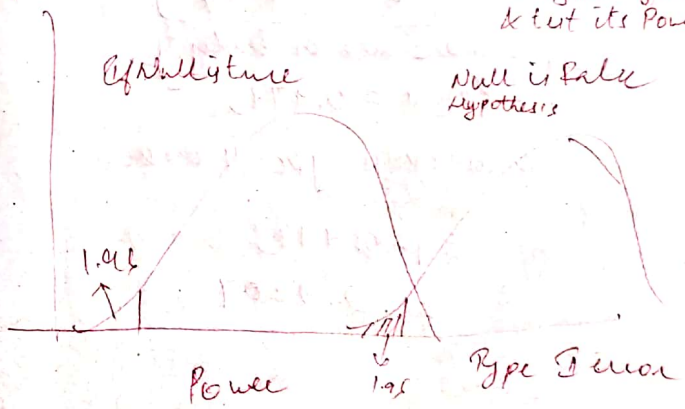
Beta : (Type II)

It the Probability of incorrectly retaining the null hypothesis

Power Ex :

A new medication that claims to improve typing ability is currently being tested the average person types at 30 (words per minute) with a standard deviation of 16. The medication is expected to increase average wpm to 46. A sample of 16 individuals is taken to determine if the medication improves typing ability use $\alpha = 0.05$

Here we assume we reject hypothesis & test its Power



We find Standard error

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{16}} = 4$$

$Z_\alpha = 0.05\%$
= Z score (σ_M)
$(1.96)(\sigma_M)$
$= (1.96)(4)$
$= 7.84$

Soil mean is at a distance of 7.84 from mean

$30 + 7.84 = 37.84 \rightarrow M$
(mean)

$$Z = \frac{M - \mu}{\sigma_M} = \frac{37.84 - 46}{4}$$

$$\boxed{Z = -2.04}$$

from Z score area in body
Power = 0.9793

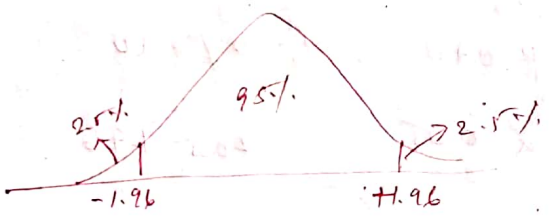
Remaining is Type II error

$= 1 - 0.9793$
$= 0.0207$

One Sample Z-test

In population, Avg IQ is 100 & with S.D of 15. A team of scientists wants to test a new medication to see if it has either a positive (or) negative effect on intelligence (or) no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence

$H_0: \mu = 100$; $H_1: \mu \neq 100$; $\alpha = 0.05$



$Z_{0.05} = 1.96$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

$Z = 14.60$

So we reject Null H_0

$Z = 14.60$; $p < 0.05$

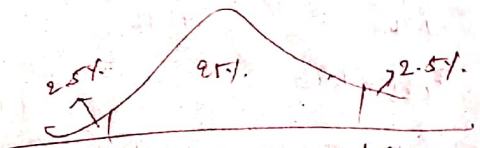
One Sample z-test for proportions

A survey claims that 9 out of 100 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is to be obtained. Of these 100 doctors, 82 indicate that they recommend aspirin., $\alpha = 0.05$

9 out of 100 so proportion $\hat{p} = 0.82$

$H_0: p = 0.90$; $H_1: p \neq 0.10$

$\alpha = 0.05$ $\quad z_{0.05} = 1.96$



$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.82 - 0.90}{\sqrt{\frac{0.90(1-0.90)}{100}}} = -2.667$$

So we reject null hypothesis

$z = -2.667$, $p < 0.05$

F-table

Note: For t-test of its unpaired we use F-test i.e $F = \frac{\text{Highest } \sigma^2}{\text{Study Lower } \sigma^2}$ & Find D.F ; If $F_{static} < F_{critical}$ (equal value)

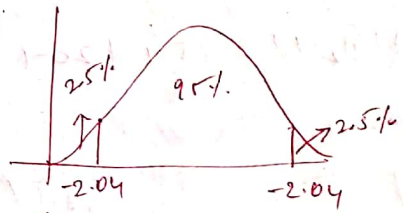
One Sample t-Test (when S.D is unknown)

In the population, Avg IQ is 100, a group scientist check if it has effect. A sample of 30 participants who have taken medication has a mean of 140, S.D of 20, $\alpha = 0.05$

$H_0: \mu = 100$; $H_1: \mu \neq 100$

$\alpha = 0.05$ $\quad df = n - 1 = 30 - 1 = 29$ → Row

↓ column



$t_{0.05} = \pm 2.04$

t-test

$\bar{x} = 140$

$\mu = 100$

$s = 20$

$n = 30$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{20}{\sqrt{30}}} = \frac{40}{3.65} = 10.96$$

So we reject Null Hypothesis

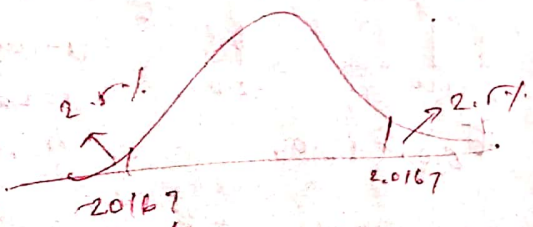
$t = 10.96$, $p < 0.05$

Independent sample t-test

A statistics teacher wants to compare his two classes to see if they performed any differently on the tests he gave that semester. Class A had 25 students with an average score of 70, S.D of 15, class B had 20 students with an average score of 74 standard deviation 25. Using $\alpha = 0.05$ did two classes perform differently on

$H_0: \mu_A = \mu_B$; $H_1: \mu_A \neq \mu_B$ $\alpha = 0.05$

$df = (n_1 - 1) + (n_2 - 1) = 25 - 1 + 20 - 1 = 43$



$t_{0.05} = -2.0167$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(sp)^2}{n_1} + \frac{sp^2}{n_2}}} \qquad sp^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

$df_1 = n_1 - 1 = 25 - 1 = 24$

$df_2 = n_2 - 1 = 20 - 1 = 19$

$SS_1 = s_1^2(df_1) = (15)^2(24) = 5400$

$SS_2 = s_2^2(df_2) = (25)^2(19) = 11875$

$$sp^2 = \frac{5400 + 11875}{24 + 19} = 401.74$$

$$t = \frac{(70 - 74)}{\sqrt{\frac{401.74}{25} + \frac{401.74}{20}}} = \frac{-4}{\sqrt{36.14}} = -0.67$$

So we do not reject null hypothesis

$t = -0.67$, $p > 0.05$

Confidence Intervals for Independent Samples t-test:

After t-test, confidence intervals can then be constructed to estimate how large that mean difference is. We do this only when we reject null hypothesis

Ex: We use

$\bar{x}_1 = 28.00$ $s_1 = 2.00$ $n_1 = 30$

$\bar{x}_2 = 20.00$ $s_2 = 3.00$ $n_2 = 30$

Construct 95% confidence interval for the difference of these two means

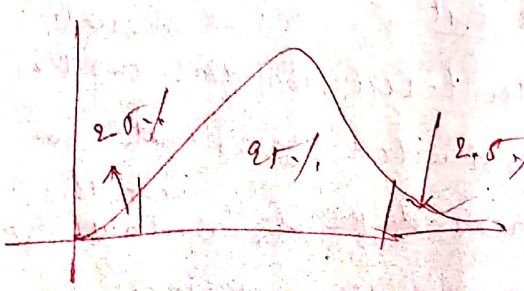
Confidence Interval = Point estimate ± Margin of error

Lower bound: $(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_{12}^2}{n_2}}$

Upper bound: $(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_{12}^2}{n_2}}$

UB: $(28-20) - t_{\alpha/2} * \sqrt{\frac{2^2}{30} + \frac{3^2}{30}} = 6.65$

LB: $(28-20) + t_{\alpha/2} * \sqrt{\frac{2^2}{30} + \frac{3^2}{30}} = 9.35$



$df_1 = 30-1 = 29$ $t_{0.05/2} = 2.0452$

$df_2 = 30-1 = 29$

We are 95% confident mean difference between sample 1 & sample 2 is between 6.65 & 9.35

Effect size: (After rejecting H_0)

$\bar{x}_1 = 28.00$, $s_1 = 2.00$, $n_1 = 30$

$\bar{x}_2 = 20.00$, $s_2 = 3.00$, $n_2 = 30$

$s_p^2 = 6.5$, $df = 58$, $t = 12.15$

Cohen's $d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2}} = \frac{28-20}{\sqrt{6.5}} = 3.14$

$d = 0.2 \rightarrow$ small effect
$d = 0.5 \rightarrow$ medium effect
$d = 0.8 \rightarrow$ large effect

3.14 indicates a very large. Our means are likely very different

other way $= r^2 = \frac{t^2}{t^2 + df} = \frac{(12.15)^2}{(12.15)^2 + 58} = .718$

$r^2 = 0.01$ → small effect

$r^2 = 0.09$ → medium effect

$r^2 = 0.25$ → large effect

0.718 indicates a very large effect

Our means are likely very different

Rsquared is the variance shared between two conditions but it is similar to cohen's 'd' [to measure effect size]

Dependent samples t-Test:-

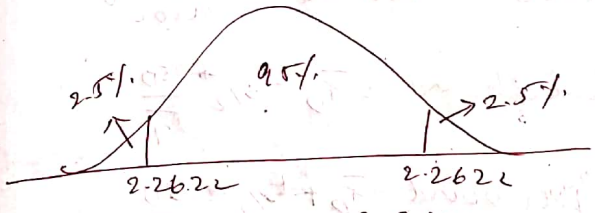
Researchers want to test a new ~~anti~~ hunger weight loss pill. They have 10 people rate their hunger both before & after taking the pill Does the pill do anything $\alpha = 0.05$

Before	After	Difference
9	7	2
10	6	4
7	5	2
5	4	1
7	4	3
8	6	-1
6	7	2
8	5	1
7	5	3
	7	0

$H_0: \mu_{before} = \mu_{after}$ $\alpha = 0.05$

$H_1: \mu_{before} \neq \mu_{after}$ $N = 9$

$df = N - 1 = 10 - 1 = 9$



$t_{0.05} = 2.2622$

$t = \frac{\bar{X}_D}{S_D/\sqrt{n}}$ $S_D = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$

$= \sqrt{\frac{49 - \frac{(17)^2}{10}}{9}} = 1.49$

$\bar{X}_D = \frac{2+4+2+1+3+(-1)+(2)+1+3+0}{10} = 1.7$

$t = \frac{1.7}{1.49/\sqrt{10}} = 3.61$

$t = 3.61$ so we reject null hypothesis.

So we rejected null hypothesis & find confidence interval for Dependent Sample t-test

$\bar{x}_D = 1.7$; $S_D = 1.49$ $n = 10$

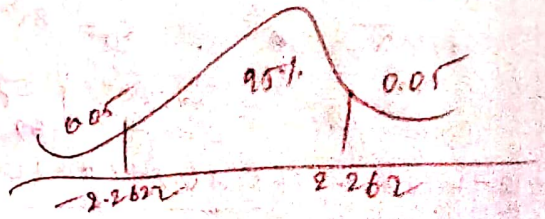
$t = 3.61$

95% confidence

Confidence Interval = Point estimate ± Margin of error

Lower bound $= \bar{x}_D - t_{\alpha/2} * \frac{S_D}{\sqrt{n}}$

Upper bound $= \bar{x}_D + t_{\alpha/2} * \frac{S_D}{\sqrt{n}}$



$df = N - 1 = 10 - 1 = 9$

$LB = 1.7 - 2.262 * \frac{1.49}{\sqrt{10}} = 0.634$

$UB = 1.7 + 2.262 * \frac{1.49}{\sqrt{10}} = 2.76$

Effect size : allows us to measure the magnitude of mean difference & we do this only after rejecting H_0

$\bar{x}_D = 1.7$, $S_D = 1.49$, $n = 10$, $t = 3.61$

Cohen's 'd' $= \frac{\bar{x}_D}{S_D} = \frac{1.7}{1.49} = 1.14$

$d = 0.2$ small effect

$d = 0.5$ medium effect

$d = 0.8$ large effect. i.e mean are very different

r^2 squared

Amount of variability shared between the groups

$r^2 = \frac{t^2}{t^2 + df} = \frac{(3.61)^2}{(3.61)^2 + 9} = 0.59$

r^2 0.01 small effect

$r^2 = 0.09$ medium effect

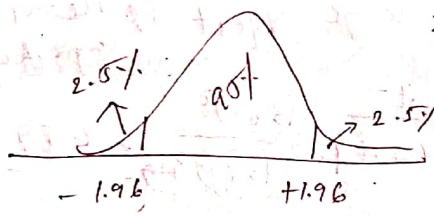
$r^2 = 0.25$ large effect

2-test for Proportions, Two samples

Researchers want to test the effectiveness of a new anti anxiety medication. In clinical testing, 64 out 200 people taking the medication report ~~[illegible]~~ symptoms of anxiety. Of the people receiving a placebo, 92 out of 200 report symptoms of anxiety. Is the medication any differently than the placebo? Test this claim using alpha $\alpha = 0.05$

$H_0: P_1 = P_2$; $H_1: P_1 \neq P_2$ $\alpha = 0.05$

$2\alpha_2 = 0.025$



$n_1 = 200$ $\hat{P}_1 = \frac{64}{200} = 0.32$ $x_1 = 64$

$n_2 = 200$ $\hat{P}_2 = \frac{92}{200} = 0.46$ $x_2 = 92$

$$z = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\hat{p}(1-\hat{p})} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{p} = \frac{64+92}{200+200} = 0.39$$

$$z = \frac{0.32 - 0.46}{\sqrt{(0.39)(1-0.39)} \cdot \sqrt{\frac{1}{200} + \frac{1}{200}}} = 2.869$$

$z = 2.869$

So we reject Null hypothesis.

$z = -2.869$, $p < 0.05$

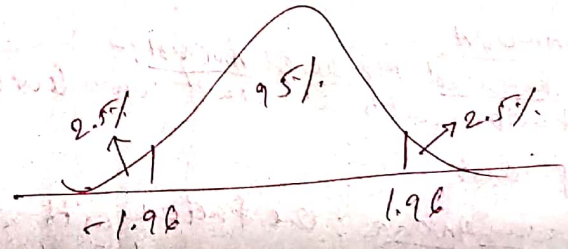
Confidence Interval → 95% confidence

$\alpha = 5\%$ $n_1 = 200$, $n_2 = 200$

$\hat{P}_1 = \frac{64}{200} = 0.32$; $\hat{P}_2 = \frac{92}{200} = 0.46$

$$L.B = (\hat{P}_1 - \hat{P}_2) - z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

$$U.B = (\hat{P}_1 - \hat{P}_2) + z_{\alpha/2} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$



$$L.B = (0.46 - 0.32) - 1.96\sqrt{\frac{(0.46)(1-0.46)}{200} + \frac{(0.32)(1-0.32)}{200}}$$

$$U.B = (0.46 - 0.32) + 1.96\sqrt{\frac{(0.46)(1-0.46)}{200} + \frac{(0.32)(1-0.32)}{200}}$$

$L.B = 0.045$; $U.B = 0.235$

We are 95% confident mean difference between two proportion is between 0.045 & 0.235.

ANOVA

It is a statistical method used to compare the means of two (or) more groups

* Factors (Variables) Ex: Gender
* Levels (Categories) Ex: Male & Female

Example: Factor - "Dosage"
Levels - "0mg, 5mg, 10mg"

Types of ANOVA

Repeated Measures ANOVA:-

* One factor with at least two levels dependent
* One way Anova: One factor with at least two levels are independent

One way ANOVA Example

0mg	50 mg	100mg
9	3	4
6	2	3
⋮	⋮	2
1	⋮	⋮
⋮	⋮	⋮

Repeated Measures ANOVA: Example

Day 1	Day 2	Day 3
9	7	4
8	6	3
⋮	⋮	2
⋮	⋮	⋮

Factorial ANOVA:- Two (or) more factors (each of which with at least two levels), levels can be either independent, dependent (or) both mixed

	Day 1	Day 2	Day 3
Men	9	7	4
	8	6	3
	7	6	2
	⋮	⋮	⋮
Women	⋮	⋮	⋮
	1	1	⋮

Assumptions in ANOVA

1. Normality of Sampling Distribution of Means
   → Distribution of sample means is normally distributed
2. Independence of Errors
   Errors between cases are independent of one another.
3. Absence of Outliers
   Outlying scores have been removed from the data set
4. Homogeneity of Variance
   Population variances in different levels of each independent variable are equal

Hypothesis in ANOVA

ANOVA with one factor ("A", three levels)

$H_0: \mu_{A1} = \mu_{A2} = \mu_{A3}$ ← Main effect

H_1: not all means are equal

ANOVA with two factors (A & B, each 3 levels)

$H_0: \mu_{A1} = \mu_{A2} = \mu_{A3}$

H_1: not all means are equal

$H_0: \mu_{B1} = \mu_{B2} = \mu_{B3}$

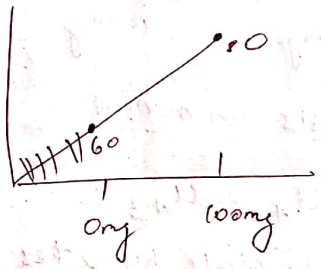
H_1: not all means are equal

H_0: Interaction absent

H_1: Interaction present ← Iteration effect

Ex: (ANOVA with One factor) → Main effect

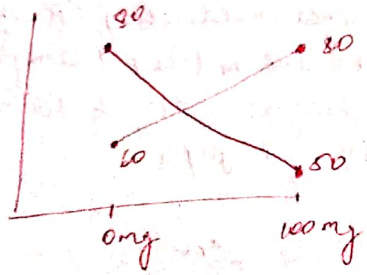
Pretend we are comparing test score of people who have received medication (100mg dosage) & people who have taken (0 mg dosage) 0mg condition has mean 60 & 100mg condition has mean of 80 in Graph



Interaction effects in ANOVA:-

$H_0: \mu_{A1} = \mu_{A2} = \mu_{A3}$	$H_0: \mu_{B1} = \mu_{B2} = \mu_{B3}$	H_0: Interaction absent
H_1: not all means are equal	H_1: not all means are equal	H_1: Interaction present

Here we have a factorial ANOVA with two factors: dosage (0mg & 100mg) & gender (Men & Women). In the 0mg dosage condition men have a mean of 80 while woman have a mean of 50. This could be represented in a graph like this:



60-10: Men
80-50: Women

Post Hoc Analysis in ANOVA

$H_0: \mu_{A1} = \mu_{A2} = \mu_{A3}$

If we reject the Null hypothesis all we know that there is a difference somewhere among the group

Additional tests called Post Hoc tests can be done to determine where differences lie

F distribution in ANOVA:-

When doing an ANOVA we calculate an 'F' statistic. It is similar to other statistics such as '2' & 't'

$$F = \frac{\text{Treatment Differences} + \text{Random Differences}}{\text{Random Differences}}$$

$$F = \frac{\sigma^2 \text{High}}{\sigma^2 \text{low}}$$

F → it always +ve

If there are no treatment differences (that is, there is no actual effect), we expect F to be 1. If there are treatment differences, we expect F to be greater than 1

F statistic has its own one tailed distribution much like how the "2" & "t" statistic have their own separate distributions

One way ANOVA

One factor with at least two levels, levels are independent

Ex: Researcher want to test a new anti anxiety medication. They split participants into 3 conditions (0mg, 50mg, 100mg), then ask them to rate their anxiety level on a scale of 1-10 Are there any differences between 3 conditions using alpha = 0.05

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
9	8	4
8	7	3
	6	2

$H_0: \mu_{0mg} = \mu_{50mg} = \mu_{100mg}$ $n = 7$

H_1: not all μ's are equal $N = 21$

$\alpha = 0.05$

$N = 21$, $n = 7$

$df_{Between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

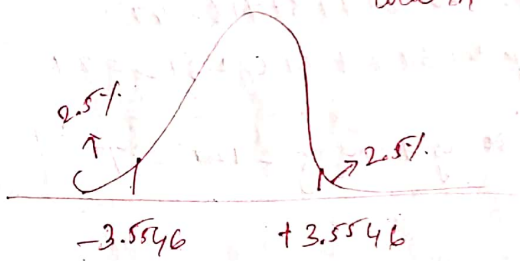
$df_{Total} = N - 1 = 21 - 1 = 20$

To look up critical value we take

$(df_{Between}, df_{within}) = (2, 18)$

↓ Row ↓ column

Column Row



If f is greater than 3.5546 we reject H_0

	SS	df	MS	F
Between	98.67	2	49.34	
within	10.29	18	0.57	
Total	108.95	20		

$SS_{between}$ SS_{within} SS_{Total}

$$SS_{between} = \frac{\Sigma (\Sigma a_i)^2}{n} - \frac{T^2}{N}$$

$$SS_{between} = \frac{(57)^2+(47)^2+(21)^2}{7} - \frac{T^2}{N}$$

0mg group: 9+8+7+8+8+9+8 = 57
50mg group: 7+6+6+7+8+7+6 = 47
100mg group: 4+3+2+3+4+3+2 = 21

0mg + 50mg + 100mg = 125 = T.

$$SS_{between} = \frac{(57)^2+(47)^2+(21)^2}{7} - \frac{(125)^2}{21} = 98.67$$

$$SS_{within} = \Sigma Y^2 - \frac{\Sigma (\Sigma a_i)^2}{n}$$

$$\Sigma Y^2 = \frac{(57)^2+(47)^2+(21)^2}{7}$$

$\Sigma Y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2+7^2+6^2+6^2+7^2+8^2+\ldots+2^2 = 853$

$$SS_{within} = 853 - \frac{(57)^2+(47)^2+(21)^2}{7}$$

$$= 10.29$$

$$SS_{Total} = \Sigma Y^2 - \frac{T^2}{N} = 853 - \frac{(125)^2}{21}$$

$$= 108.95$$

$$F = \frac{MS_{between}}{MS_{within}} \quad \left| \quad MS_{between} = \frac{98.67}{2} = 49.34 \right.$$

$$MS_{within} = \frac{10.29}{18} = 0.57$$

$$F = \frac{49.34}{0.57} = 86.56$$

As $3.55 < 86.56$ so we reject Null Hypothesis

Conclusion:

3 conditions differed significantly on anxiety level $F(2, 18) = 86.56$, $p < 0.05$

Effect Size for One way ANOVA

ANOVA measures means are different from one another if it's large (or) small so we use effect size it helps us to measure

Continue to previous problem

The most common measure of effect size for a One way ANOVA is Eta Squared

$$\eta^2 = \frac{SS_{between}}{SS_{Total}} = \frac{98.67}{108.95} = 0.91$$

Eta Squared

91% of the total variance is accounted for by the treatment effect.

So we conclude there is a large meaningful difference between 3 groups.

Post Hoc Tests for One way ANOVA :-

So after doing One way ANOVA all we know that groups are different in some way

But Post Hoc Test tell us how they are different μ_0 is different (or) μ_{50} is different (or) μ_{100} is different vice versa. & where they are different

Two different tests

* Tukey HSD
* Scheffe

In Tukey test is more liberal than Scheffe test. In Tuky we have greater chance of rejecting Ho & making Type I error.

In Scheffe we have less chance of rejecting Null hypothesis & less chance of making error.

In tuckey we might find what we want but make a mistake as the chance of

making mistake is high

In scheffe we have less chance of finding what we want & also less chance of making mistake

Tukey HSD

With this test we are interested in examining mean differences

0mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$\bar{x}_{0mg} = 8.14$; $\bar{x}_{50mg} = 6.71$;

$\bar{x}_{100mg} = 3.00$

$$HSD = q\sqrt{\frac{MS_{within}}{n}}$$

$$= q\sqrt{\frac{0.57}{7}} \qquad n=7$$

q-table

Here we q table & only df_{within} & 3 groups we are comparing & $\alpha = 0.05$

$$\boxed{q = 3.61}$$

$$= 3.61\sqrt{\frac{0.57}{7}} = 1.03$$

$\bar{x}_{0mg} = 8.14$; $\bar{x}_{50mg} = 6.71$; $\bar{x}_{100mg} = 3.00$

$\bar{x}_{0mg} - \bar{x}_{50mg} = 8.14 - 6.71 = 1.43$

$\bar{x}_{0mg} - \bar{x}_{100mg} = 8.14 - 3.00 = 5.14$

$\bar{x}_{50mg} - \bar{x}_{100mg} = 6.71 - 3.00 = 3.71$

Any mean above 1.03 are apart i.e are different.

Scheffe

$$F_{comparison} = \frac{MS_{between}}{MS_{within}}$$

MS_{within} is same as in One way ANOVA

$MS_{Between}$ is different, Now scheffe is a pair wise test so we to calculate 3 $SS_{between's}$ (0,50) (0,100) (50,100)

(3)

$$SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$$

For (0,50)

$$SS_{between} = \frac{57^2+47^2}{7} - \frac{104^2}{14} = 7.14$$

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{7.14}{2} = 3.57$$

$$F_{comparison} = \frac{3.57}{0.57} = 6.26$$

We are going use same Fvalue we used in One way ANOVA

(0,50) for → $F_{comparison} > F_{value}$ so we reject Null Hypothesis. I say Omg I song groups different from one another.

Similarly for (0,100) & (50,100)

Repeated Measures ANOVA

One Factor with atleast two levels, Levels are dependent.

By saying that levels are dependent, it means that they share variability in some way. It's almost same expect one additional calculation we must perform to account for this shared variability.

Same example of One way anova but Medication is given once a week & another time once ~~two~~ weeks.

$\alpha = 0.05$

Before	week 1	week 2
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$H_0: \mu_{before} = \mu_{week1} = \mu_{week2}$

H_1: Not all μ's are equal

$\alpha = 0.05$

$N = 21$ $S = 7$

$df_{Between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

$df_{subjects} = S - 1 = 7 - 1 = 6$

$df_{error} = df_{within} - df_{subjects} = 18 - 2 = 16$

$df_{Total} = N - 1 = 21 - 1 = 20$

To look up critical value

$df_{between} = a - 1 = 3 - 1 = 2$

$df_{error} = df_{within} - df_{subjects}$

$= 18 - 6 = 12$

$(df_{between}, df_{error}) = (2, 12)$

(Column, Row)

F value = 3.8853

Aff is greater than 3.8853 we reject Null Hypothesis

	SS	d.f	MS	F
Between	98.67	2	49.34	224.27
within	10.29	18		
subjects	7.62	6		
Error	2.67	12	0.22	
Total	108.96	20		

$$SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{S} - \frac{T^2}{N}$$

Before group = $9+8+7+8+8+9+8 = 57$

week1 group = $7+6+6+7+8+7+6 = 47$

week2 group = $4+3+2+3+4+3+2 = 21$

$$SS_{between} = \frac{(57)^2 + (47)^2 + (21)^2}{S} - \frac{T^2}{N}$$

$$= \frac{(57)^2 + (47)^2 + (21)^2}{7} - \frac{(125)^2}{21}$$

$$= 98.67$$

$$SS_{within} = \Sigma Y^2 - \frac{\Sigma(\Sigma a_i)^2}{S}$$

$\Sigma Y^2 = 9^2 + 8^2 + \dots + 2^2 = 855$

$$SS_{within} = 853 - \frac{(57)^2 + (47)^2 + (21)^2}{7}$$

$$= 10.29$$

$$SS_{subjects} = \frac{\Sigma(\Sigma S_i)^2}{a} - \frac{T^2}{N}$$

Subject one: 9+7+4 = 20

Subject Two: 8+6+3 = 17

⋮

Subject seven = 8+6+2 = 16

$$SS_{subjects} = \frac{(20)^2 + (17)^2 + (15)^2 + (18)^2 + (20)^2 + (19)^2 + (16)^2}{3} - \frac{(125)^2}{21}$$

$$= 7.62$$

Error = Within − Subjects = 10.29 − 7.62

= 2.67

Total = Between + Within

= 98.67 + 10.29 = 108.96

$$F = \frac{MS_{between}}{MS_{error}}$$

$$MS_{between} = \frac{98.67}{2} = 49.34$$

$$MS_{error} = \frac{2.67}{12} = 0.22$$

$$F = \frac{49.34}{0.22} = 224.27$$

F value = 3.8853

F = 224.27

So we reject Null Hypothesis

Conclusion:

3 conditions differed significantly on anxiety level, $F(2,12) = 224.27$, $p < 0.05$

Factorial ANOVA, Two independent Factors:

Two factors with at least two levels each, levels are Independent.

Factorial ANOVA (with independent factors) is kind of like the One way ANOVA, except now you're dealing with more than one independent variable.

Researches want to test a new anti anxiety medication. They measure anxiety of 36 participants on different dosages of the medication: 0mg, 50mg, 100mg. Participants are also divided based on what school they are attending which researcher hypothesize will also effect anxiety levels. Anxiety is rated on a scale of 1-10, with 10 being 'high anxiety & 1 being "low anxiety". Use alpha = 0.05 to conduct your analysis.

	0mg	50mg	100mg
High school students	3	5	7
	4	4	8
	5	3	7
	3	5	8
	4	5	7
	3	5	7
College students	1	5	9
	2	4	8
	1	3	9
	1	5	8
	1	5	7
	2	4	9

H_0: $\mu_{highschool} = \mu_{college}$

H_1: $\mu_{highschool} \neq \mu_{college}$

H_0: $\mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

H_1: not all dosage means.

H_0: An interaction is absent

H_1: An interaction is present

$\alpha = 0.05$

$df_{school(A)} = a - 1 = 2 - 1 = 1$

$df_{dosage(B)} = b - 1 = 3 - 1 = 2$

$df_{school \times dosage(A \times B)} = (a-1)(b-1) = (2-1)(3-1) = 2$

$df_{error} = N - ab = (36) - (2)(3) = 30$

$df_{Total} = N - 1 = 36 - 1 = 35$

in F table — column ROW

School $(df_{school(A)}, df_{error})$: $(1, 30)$ C.V = 4.17

Dosage $(df_{dosage(B)}, df_{error})$: $(2, 30)$ C.V = 3.32

school * Dosage $(df_{school*dosage(A*B)}, df_{error})$

$= (2, 30)$

C.V = 3.32

3 hypothesis, so we have 3 decision rules

[school] → F is greater than 4.17, reject Null hypothesis

[Dosage] → F is greater than 3.32, reject Null hypothesis

[Interaction] → F is greater than 3.32, reject null hypothesis.

	SS	df	MS	F
School	2.25	1	2.25	4.166
Dosage	175.17	2	87.59	162.20
Interaction	17.16	2	8.58	15.89
Error		30	0.54	
Total	210.75	35		

SS_{school} SS_{dosage} $SS_{school*dosage}$ $S.S_{error}$

S_{total}

$$SS_{school} = \frac{\sum (\sum a_i)^2}{(b)(n)} - \frac{T^2}{N}$$

High school sum = 3+4+5+3+4+3+5+4+3+5+5+5+7+8+7+8+7+7 = 93

College sum = 1+2+1+1+1+2+5+4+3+5+5+4+9+8+9+8+7+9 = 84

$$SS_{school} = \frac{(93)^2 + (84)^2}{(3)(6)} - \frac{(177)^2}{36} = 2.25$$

$$SS_{dosage} = \frac{\sum (\sum b_j)^2}{(a)(n)} - \frac{T^2}{N} = \frac{(30)^2 + (53)^2 + (94)^2}{(2)(6)} - \frac{(177)^2}{36}$$

$= 175.17$

0 mg sum = 3+4+5+3+4+3+1+2+1+1+1+2 = 30

50 mg sum = 5+4+3+5+5+5+4+3+5+5+4 = 53

100 mg sum = 7+8+7+8+7+7+9+8+9+8+7+9 = 94

Test statistic

$SS_{school \times dosage}$ (Interaction)

$$= \frac{\Sigma(\Sigma a_i b_i)^2}{n} - \frac{\Sigma(\Sigma a_i)^2}{(b)(n)} - \frac{\Sigma(\Sigma b_i)^2}{(a)(n)} + \frac{T^2}{N}$$

$$= \frac{(22)^2+(8)^2+(27)^2+(26)^2+(44)^2+(50)^2}{6} - \frac{(93)^2+(84)^2}{(3)(6)} - \frac{(80)^2+(53)^2+(94)^2}{(2)(6)} + \frac{(177)^2}{36} = 17.16$$

Cell 1 = 3+4+5+3+4+3 = 22, Cell 4 = 5+4+3+5+5+4 = 26

Cell 2 = 1+2+1+1+1+2 = 8, Cell 5 = 7+8+7+8+7+7 = 44

Cell 3 = 5+4+3+5+5+5 = 27, Cell 6 = 9+8+9+8+7+9 = 50

$$SS_{total} = \Sigma Y^2 - \frac{T^2}{N} = 1081 - \frac{(177)^2}{36} = 210.75$$

$\Sigma Y^2 = 3^2+4^2+5^2+3^2+\dots+1^2+2^2+\dots$

$+5^2+4^2+\dots 5^2+4^2+\dots+9^2$

All scores

$= 1081$

$SS_{error} = SS_{total} - SS_{school} - SS_{dosage} - SS_{Interaction}$

$= 16.17$

$MS_{school} = \frac{2.25}{1}$

$MS_{dosage} = \frac{175.17}{2}$

$MS = \frac{SS}{df}$ $MS_{Interaction} = \frac{17.16}{2}$

$F = \frac{MS_{effect}}{MS_{error}}$

$F_A = \frac{2.25}{0.54} = 4.17$

$F_B = \frac{87.59}{0.54} = 162.20$, $F_{A \times B} = \frac{5.58}{0.54} = 15.89$

[School] If F is greater than 4.17 reject Null hypothesis, Our F = 4.166 Retain Null hypothesis

[dosage] If F is greater than 3.32, reject the Null hypothesis, Our F = 162.20, Reject Null hypothesis

[Interaction] If F is greater than 3.32, reject the Null hypothesis, Our F = 15.89 Reject Null hypothesis.

Conclusion

High school students & college students did not have significantly different anxiety levels $F(1,30) = 4.166, p > 0.05$. There was a significant difference between 3 different levels of dosage, $F(2,30) = 162.20, P < 0.05$. An interaction effect was also present, $F(2,30) = 15.59, p < 0.05$

Factorial ANOVA, Two Dependent Factors

Two factors with at least two levels each, Both factors are Dependent

Factorial ANOVA (with two dependent factors) is most like an extension of a Related Measures ANOVA

Researches want to compare the anxiety levels of six individuals at two marital states after they have been divorced & then again after they have been gotten married. Anxiety is measured at 3 times: week1, week2, week3. Anxiety is rated on a scale of 1-10 with 10 being "high anxiety" & 1 being "low anxiety"

$\alpha = 0.05$ to conduct Analysis

	subjut	week1	week2	week3
Divorced	1	3	5	7
	2	4	4	8
	3	5	3	7
	4	3	5	8
	5	4	5	7
	6	3	5	7
Married	1	1	5	9
	2	2	4	8
	3	1	3	9
	4	1	5	8
	5	1	5	7
	6	2	4	4

"Same person" across different weeks

$H_0: \mu_{divorced} = \mu_{married}$

$H_1: \mu_{divorced} \neq \mu_{married}$

$H_0: \mu_{week1} = \mu_{week2} = \mu_{week3}$

H_1: not all week means are equal

H_0: an interaction is absent

H_1: an interaction is present

$\alpha = 0.05$

Degree of Freedom

Source	SS	df	MS	F
Marital Status (A)	2.25	1	2.25	9.00
Error (AxS)	1.25	5	0.25	
week (B)	175.17	2	87.59	93.18
Error (BxS)	9.4	10	0.94	
Interaction (A x B)	17.16	2	8.58	17.51
Error (AxBxS)	4.94	10	0.49	
Error (S)	0.58	5		
Total	210.75	35		

DF:-

$df_{\text{marital status (A)}} = a - 1 = 2 - 1 = 1$

$df_{\text{error (A x S)}} = (a-1)(n-1) = (1)(5) = 5$

$df_{\text{week (B)}} = b - 1 = 3 - 1 = 2$

$df_{\text{error (B x S)}} = (b-1)(n-1) = (2)(5) = 10$

$df_{\text{error (A x B x S)}} = (a-1)(b-1)(n-1)$
$= (1)(2)(5) = 10$

$df_{\text{marital status x week (A x B)}}$
$= (a-1)(b-1) = (1)(2) = 2$

$df_{\text{error (S)}} = n - 1 = 6 - 1 = 5$

$df_{\text{Total}} = N - 1 = 36 - 1 = 35$

We have 3 Hypothesis so 3 decision rules

Marital Status $(df_{\text{marital status (A)}}, df_{\text{error (A×S)}})$

$= (1, 5)$ → Row C.V = 6.61

Week $(df_{\text{week (B)}}, df_{\text{error (B×S)}})$

$= (2, 10)$ → Column C.V = 4.10

Marital Status × Week $(df_{\text{school × week (A×B)}}, df_{\text{error (A×B×S)}})$

$= (2, 10)$

C.V = 4.10

[Marital Status] If F is greater than 6.61 reject the Null hypothesis

[Week] If F is greater than 4.10 reject the Null hypothesis.

[Interaction] If F is greater than 4.10 reject the Null hypothesis

Test statistic

$SS_{\text{marital status (A)}}$, $SS_{\text{week (B)}}$, $SS_{\text{school×week (A×B)}}$

$SS_{\text{error (S)}}$, $SS_{\text{error (A×S)}}$, $SS_{\text{error (B×S)}}$

$SS_{\text{error (A×B×S)}}$, SS_{Total}

$$SS_{\text{marital status}} = \frac{\Sigma(\Sigma a_i)^2}{bn} - \frac{T^2}{N}$$

Divorced sum = 3+4+5+ ... +8+7+7 = 93

Married sum = 1+2+1+ ... +8+7+9 = 84

$$SS_{\text{marital status}} = \frac{(93)^2 + (84)^2}{(3)(6)} - \frac{(177)^2}{36} = 2.25$$

$$SS_{\text{week}} = \frac{\Sigma(\Sigma b_i)^2}{(a)(n)} - \frac{T^2}{N} = \frac{(30)^2 + (53)^2 + (94)^2}{(2)(6)} - \frac{(177)^2}{36} = 115.1$$

Week 1 Sum = 3+4+5+3+ ... +1+2 = 30

Week 2 Sum = 5+4+3+ ... +5+5+4 = 53

Week 3 Sum = 7+8+7+ ... +7+9 = 94

$$SS_{\text{marital status × week}} = \frac{\Sigma(\Sigma a_i b_i)^2}{n} - \frac{\Sigma(\Sigma a_i)^2}{bn} - \frac{\Sigma(\Sigma b_i)^2}{an} + \frac{T^2}{N}$$

$$= \frac{\Sigma(\Sigma a_i b_i)^2}{n} - \frac{(93)^2+(84)^2}{(3)(6)} - \frac{(30)^2+(53)^2+(94)^2}{(2)(6)} + \frac{(177)^2}{36}$$

$$= \frac{(29)^2+(16)^2+(27)^2+(26)^2+(44)^2+(60)^2}{6} \quad " \quad " \quad " = 17.16$$

Cell 1 = 3+4+5+3+4+3 = 22, Cell 4 = 5+4+3+5+5+4 = 26

Cell 2 = 1+2+1+1+1+2 = 8, Cell 5 = 7+8+7+8+7+7 = 44

Cell 3 = 5+4+3+5+5+5 = 27, Cell 6 = 9+8+9+8+7+9 = 50

$$SS_{Total} = \Sigma Y^2 - \frac{T^2}{N} = 1081 - \frac{T^2}{36} = 1081 - \frac{(177)^2}{36} = 210.75$$

$$\Sigma Y^2 = 3^2+4^2+\ldots+7^2+9^2 = 1081$$

$$SS_{error(A\times S)} = \frac{\Sigma(\Sigma a_i n_i)^2}{(b)} - \frac{\Sigma(\Sigma a_i)^2}{(b)(n)} - \frac{\Sigma(\Sigma n_i)^2}{(a)(b)} + \frac{T^2}{N}$$

$$= \frac{\Sigma(\Sigma a_i n_i)^2}{b} - \frac{(93)^2+(84)^2}{(3)(6)} - \frac{\Sigma(\Sigma n_i)^2}{(a)(b)} + \frac{(177)^2}{36}$$

Subject 1 = 3+5+7+4+5+9 = 30, Subject 4 = 3+5+8+4+5+8 = 30

Subject 2 = 4+4+8+2+4+8 = 30, Subject 5 = 4+5+7+4+5+7 = 21

Subject 3 = 5+3+7+4+3+9 = 28, Subject 6 = 8+5+7+2+4+4 = 30

$a_1 n_1 = 3+5+7 = 15$

$a_2 n_2 = 4+4+8 = 16$

$a_3 n_3 = 5+3+7 = 15$

$\vdots$

$a_{12} n_{12} = 2+4+9 = 15$

$$= \frac{(15)^2+(16)^2+(15)^2+(16)^2+(16)^2+(15)^2+(15)^2+(14)^2+(13)^2+(14)^2+(15)^2+(15)^2}{3}$$

$$- \frac{93^2+(84)^2}{(3)(6)} - \frac{(30)^2+(30)^2+(28)^2+(30)^2+(29)^2+(30)^2}{(2)(3)} + \frac{(177)^2}{36} = 1.25$$

$$SS_{error(B\times S)} = \frac{\Sigma(\Sigma b_i n_i)^2}{a} - \frac{\Sigma(\Sigma b_i)^2}{(a)(n)} - \frac{\Sigma(\Sigma n_i)^2}{(a)(b)} + \frac{T^2}{N}$$

$$= \frac{\Sigma(\Sigma b_i n_i)^2}{a} - \frac{(30)^2+(53)^2+(94)^2}{(2)(6)} - \frac{(30)^2+(30)^2+(28)^2+(30)^2+(29)^2+(30)^2}{(2)(3)}$$

$$+ \frac{(177)^2}{36}$$

$SS_{error(B\times S)} = (4^2+(0)^2+(16)^2+(6)^2+8^2+(16)^2+6^2+6^2 + 16^2+4^2+10^2+16^2+5^2+10^2 + 16^2+5^2+9^2+16^2$

$- \frac{(30)^2+(53)^2+(44)^2}{(2)\ (6)} - \frac{(30)^2+(30)^2+(25)^2+(30)^2+(29)^2+(30)^2}{(2)\ (3)} + \frac{(177)^2}{36} = 9.4$

Subject 1 = 3+1 = 4 ; subject 2 = 4+2 = 6

Subject 3 = 5+5 = 10 ; subject 2 = 4+4 = 8

Subject 4 = 7+9 = 16 ; subject 2 = 8+8 = 16

Subject 6 = 5

Subject 6 = 9

Subject 6 = 16

$SS_{error(S)} = \frac{\Sigma(\Sigma n_i)^2}{(a)\ (b)} - \frac{T^2}{N}$

$= \frac{(30)^2+(30)^2+(28)^2+(30)^2+(29)^2+(30)^2}{(2)\ (3)} - \frac{(177)^2}{36} = 0.58$

Error (A×B×S) = Total − sum of the found (SS)

= 4.94

$MS = \frac{SS}{df}$; $F = \frac{MS_{effect}}{MS_{error}}$

$f_A = \frac{2.25}{0.25} = 9.00$; $f_B = \frac{87.59}{0.94} = 93.19$

$f_{A\times B} = \frac{8.58}{0.49} = 17.51$

[Marital status] : Fi greater than 6.61 Our F = 9.00, reject Ho

[Week] : F is greater than 4.10, reject Our F = 93.18, null

[Interaction] : F is greater than 4.10 reject Our F = 17.5ϕ, Reject Ho

Conclusion: Anxiety levels differed significantly for divorced & then remarried individuals, $F(1,5) = 9.00$, $p < 0.05$. There was a significant difference between three different weeks, $F(2,10) = 93.18$, $p < 0.05$. An interaction effect was also present, $F(3,10) = 17.51$, $p < 0.05$

Factorial ANOVA, Two Mixed Factors:-

Two factors with atleast two levels such One factor is independent, while the other factor is dependant. This Factorial ANOVA is combination of One way ANOVA & Repeated Measures ANOVA.

	Week1	Week2	Week3
High school students	3	5	7
	4	4	8
	5	3	7
	3	5	8
	4	5	7
	3	5	7
College students	1	5	9
	2	4	8
	1	3	9
	1	5	8
	1	5	7
	2	4	9

$H_0: \mu_{highschool} = \mu_{college}$; $H_0: \mu_{week_1} = \mu_{week_2} = \mu_{week_3}$

$H_1: \mu_{highschool} \neq \mu_{college}$; H_1 not all week means are equal

H_0: an interaction is absent

H_1: an interaction is present

$\alpha = 0.05$

	SS	df	MS	F
School(A)	2.25	1	2.25	12.5
-Error(S/A)	1.83	10	0.18	
Week(B)	175.17	2	87.59	121.65
Interaction (AxB)	17.16	2	8.58	11.92
-Error(BxS/A)	14.34	20	0.72	
Total	210.75	35		

$df_{school(A)} = a - 1 = 2 - 1 = 1$

$df_{error(S/A)} = (a)(n-1) = (2)(6-1) = 10$

$df_{week(B)} = b - 1 = 3 - 1 = 2$

$df_{school \times week(AxB)} = (a-1)(b-1) = (1)(2) = 2$

$df_{error(BxS/A)} = (a)(b-1)(n-1) = (2)(2)(5) = 20$

$df_{Total} = N - 1 = 36 - 1 = 35$

F Table

3 hypothesis, 3 Decision rules

$S_{chool}\left(df_{school(A)}, df_{error\left(\frac{S}{A}\right)}\right)$

$= (1,10)$ (1 → column, 10 → Row) C.V = 4.35

$week\left(df_{week(B)}, df_{error\left(B \times \frac{S}{A}\right)}\right) : (2,20)$

C.V = 3.49

$school \times week\left(df_{school \times week(A \times B)}, df_{error\left(B \times \frac{S}{A}\right)}\right) : (2,20)$ C.V = 3.49

[school]: If F is greater than 4.35, reject Null hypothesis

[week]: If F is greater than 3.49, reject H_0

[Interaction]: If F is greater than 3.49 reject H_0

$SS_{school(A)}$, $SS_{error\left(\frac{S}{A}\right)}$, $SS_{week(B)}$,

$SS_{school \times week(A \times B)}$, $SS_{error\left(\frac{B \times S}{A}\right)}$

SS_{Total}

$$SS_{school} = \frac{\Sigma(\Sigma a_i)^2}{(b)(n)} - \frac{T^2}{N} = \frac{(93)^2+(84)^2}{(3)(6)} - \frac{(177)^2}{36} = 2.25$$

High school sum $= 3+4+5+\cdots+7+7 = 93$

College sum $= 1+2+1+1+1+\cdots+7+9 = 84$

$$SS_{week} = \frac{\Sigma(\Sigma b_i)^2}{(a)(n)} - \frac{T^2}{N} = \frac{(30)^2+(53)^2+(94)^2}{(2)(6)} - \frac{(177)^2}{36} = 175.17$$

Week 1 Sum $= 3+4+5+3+\cdots+1+2+\cdots+2 = 30$

Week 2 Sum $= 5+4+3+5+\cdots+5+5+4 = 53$

Week 3 sum $= 7+8+7+8+\cdots+8+7+9 = 94$

$$SS_{school \times week} = \frac{\Sigma(\Sigma a_i b_i)^2}{(n)} - \frac{\Sigma(\Sigma a_i)^2}{(b)(n)} - \frac{\Sigma(\Sigma b_i)^2}{(a)(n)} + \frac{T^2}{N}$$

$$= \frac{\Sigma(\Sigma a_i b_i)^2}{n} - \frac{(93)^2+(84)^2}{(3)(6)} - \frac{(30)^2+(53)^2+(94)^2}{(2)(6)} + \frac{(177)^2}{36}$$

$$\frac{(22)^2+(8)^2+(27)^2+(26)^2+(44)^2+(50)^2}{(6)} - \frac{(93)^2+(84)^2}{(3)(6)} - \frac{(30)^2+(53)^2+(94)^2}{(2)(6)} + \frac{(177)^2}{36} = 17.16$$

Cell 1 = 3+4+5+3+4+3 = 22

Cell 2 = 1+2+1+1+2+1 = 8

Cell 3 = 27

Cell 4 = 26

Cell 5 = 44

Cell 6 = 50

$$SS_{total} = \Sigma Y^2 - \frac{T^2}{N}$$

$$\Sigma Y^2 = 3^2+4^2+\dots+7^2+9^2 = 1081$$

$$= 1081 - \frac{(177)^2}{36} = 210.75$$

$$SS_{error(S/A)} = \frac{\Sigma(\Sigma a_i n_i)^2}{(b)} - \frac{\Sigma(\Sigma a_i)^2}{(b)(n)}$$

$$\frac{\Sigma(\Sigma a_i n_i)^2}{(b)} - \frac{(93)^2+(84)^2}{(3)(6)}$$

$a_1 n_1 = 3+5+7 = 15$

$a_2 n_2 = 4+4+8 = 16$

$a_3 n_3 = 15$

$\vdots$

$a_{12} n_{12} = 18$

$$\frac{(15)^2+(16)^2+(15)^2+(16)^2+(16)^2+(15)^2+(15)^2+(14)^2+(13)^2+(14)^2+(13)^2+(15)^2}{3} - \frac{(93)^2+(84)^2}{(3)(6)} = 1.83$$

$$SS_{total} - SS_{school} - SS_{error(S/A)} - \dots = SS_{error(B \times S/A)}$$

$$SS_{error(B \times S/A)} = 14.34$$

$$MS = \frac{SS}{df} \; ; \; F = \frac{MS_{effect}}{MS_{error}}$$

$$F_A = \frac{2.25}{0.18} = 12.5 \; ; \; F_B = \frac{87.59}{0.72} = 121.65$$

$$F_{A \times B} = \frac{8.58}{0.72} = 11.92$$

[school] F is greater than 4.35, reject H_0
Our F = 12.5

[week] F is greater than 3.49, reject H_0
Our F = 121.65

[Interaction] F is greater than 3.49, reject H_0
Our F = 11.92

Conclusion:

High school students & college students had significantly different anxiety levels $F(1,10) = 12.5$, $p < 0.05$. There was a significant difference between 3 different weeks, $F(2,20) = 121.65$, $p < 0.05$. An interaction effect was also present, $F(2,20) = 11.92$, $p < 0.05$

Chisquare Test for Goodness of fit

χ^2 Test for Goodness of fit tests claims about population proportion.

It is non parametric test that is performed on categorial (nominal (or) ordinal) data.

Ex: In the 2000 U.S. Census, the ages of individuals in a small town were found to be following

Less than 18	18-35	Greater than 35
20%	30%	50%

In 2010 ages of $n = 500$ individuals were sampled Below are the results:

Less than 18	18-35	Greater than 35
121	288	91

Use $\alpha = 0.05$, would you conclude that the population distribution of ages has changed in the last 10 years?

	Less than 18	18-35	Greater than 35
Expected	20%	30%	50%

If $n = 500$

	Less than 18	18-35	greater than 35
Observed	121	288	91
Expected	$500 * 20 = 100$	$500 * 0.30 = 150$	$500 * 0.50 = 250$

H_0: Data Meets expected distribution

H_1: Data does not meet expected distribution.

k = no of groups

$df = k - 1 = 3 - 1 = 2$

In χ^2 table $\alpha = 0.05$ & $df = 2$

If χ^2 is greater than 5.99, reject H_0

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(121-100)^2}{100} + \frac{(288-300)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.494$$

So we reject Null hypothesis.

Conclusion:

Ages of 2010 population are different than those expected based on 2000 population.

Chi-Square Test for Independence

* It evaluates relationship between two variables
* It is a non parametric test that is performed on categorical (nominal or ordinal) data.

Ex: 500 elementary school boys & girls are asked which is their favourite color: Blue, green or pink. Results are shown Below.

Observed	Blue	Green	Pink	Column Total
Boys	100	150	20	300
Girls	20	30	180	200
Row Total	120	180	200	N = 500

Use $\alpha = 0.05$, would you conclude that there is relationship between gender & colour favourite.

H_0: Gender & colour not related

H_1: Gender & colour are related.

$\alpha = 0.05$

$df = (\text{rows} - 1)(\text{columns} - 1)$

$= (2-1)(3-1)$

$= 2$

$\alpha = 0.05$ & $df = 2$

If χ^2 is greater than 5.99 reject H_0

using χ^2 table

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad ; \quad f_e = \frac{f_c f_r}{n}$$

Expected	Blue	Green	Pink	
Boys	72	108	120	300
Girls	48	72	80	200
	120	180	200	N=500

$$(Boys, Blue) = \frac{(120 * 300)}{500} = 72$$

Similarly all values

$$\chi^2 = \frac{(100-72)^2}{72} + \frac{(20-48)^2}{48} + \frac{(150-108)^2}{108} + \frac{(30-72)^2}{72} + \frac{(20-120)^2}{120} + \frac{(180-80)^2}{80} = 226.389$$

If χ^2 is greater than 5.99, reject H_0

$\chi^2 = 276.39$

so we reject Null Hypothesis.

Conclusion : In population there is a relationship between gender & favorite colour.

Man Whitney U Test:

It is a version of the independent samples t-Test that can be performed on ordinal (ranked) data

Ordinal data in table, $\alpha = 0.05$, Is there a difference between treatment A & treatment B

Treatment - A	28	31	36	35	32	33
Treatment - B	12	18	19	14	20	19

H_0 : No difference between ranks of two treatment

H_1 : There is difference " " " "

$\alpha = 0.05$

It follows 'z' distribution if sample size is greater than 20

Two tailed Test

If z is less than -1.96 (or) > 1.96 reject H_0

Treatment A	Treatment B
28	12
31	18
36	19
35	14
32	20
33	19

Rank	Score	Sample	Points
1	12	(B)	6
2	14	(B)	6
3	18	(B)	6
4.5	19	(B)	6
4.5	19	(B)	6
6	20	(B)	6
7	28	(A)	0
8	31	(A)	0
9	32	(A)	0
10	33	(A)	0
11	35	(A)	0
12	36	(A)	0

$U_A = 0+0+0+0+0+0 = 0$

$U_B = 6+6+6+6+6+6 = 36$

$U = 0$ (Smaller value among U_A & U_B is U)

Usually lower the more different groups are

$U = 0$

$$z = \frac{U - \frac{n_A n_B}{2}}{\sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}} = \frac{0 - \frac{(6)(6)}{2}}{\sqrt{\frac{(6)(6)(6+6+1)}{12}}}$$

$$z = -2.88$$

So we reject H_0

Conclusion

There is difference between ranks of the two treatments $z = -2.88$, $P < 0.05$

Wilcoxon Signed-Ranks Test:- (U-Test)

It is a version of dependent sample t-test performed on ordinal (ranked) data.

$\alpha = 0.05$

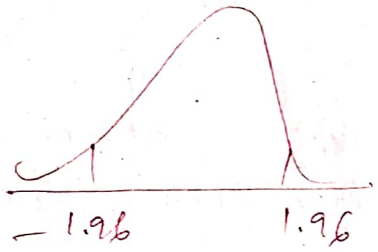
Before	28	17	36	35	32	33
After	12	31	19	14	20	19

H_0: No difference between treatments

H_1: Difference " "

$\alpha = 0.05$

Atleast 20 samples when applying Z distribution.



Before	After	Difference	Rank
28	12	16	4
17	31	-14	1
36	19	17	5
35	14	21	6
32	20	12	2
33	19	14	3

Lower → Lower number rank, High → Higher number rank

(Positive) Rank addition

$\Sigma R_+ = 4+5+6+2+3 = 20$

$\Sigma R_- = 1$

T is smaller among ΣR_+ & ΣR_-

$T = 1$

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{1 - \frac{(6)(6+1)}{4}}{\sqrt{\frac{(6)(6+1)(2(6)+1)}{24}}}$$

$Z = -1.99$

so reject H_0

Conclusion: There is difference between before & after groups $Z = -1.99$, $p < 0.05$

Kruskal Wallis Test (H-Test)

It a version of Independent measure (One way) ANOVA, performed on Ordinal (Ranked) Data.

	Group 1	Group 2	Group 3
	27	20	34
	2	8	31
	4	14	3
	18	36	23
	7	21	30
	9	22	6
T	39	65	67
n	6	6	6

H_0: No difference between treatments

H_1: Different

$\alpha = 0.05$

k = no of groups

$df = k - 1 = 3 - 1 = 2$

we use Chi square Table

If χ^2 is greater than 5.99, reject H_0

Original score	Rank
2	1
3	2
4	3
6	4
7	5
8	6
9	7
14	8
18	9
20	10
21	11
22	12
23	13
27	14
30	15
31	16
34	17
36	18

In this small number lower rank, higher large rank

Test Statistic (H)

$$H = \frac{12}{N(N+1)} \left(\sum \frac{T_i^2}{n} \right) - 3(N+1)$$

$$= \frac{12}{(18)(18+1)} \left(\frac{(39)^2}{6} + \frac{(65)^2}{6} + \frac{(67)^2}{6} \right) - 3(18+1) = 2.854$$

So we reject H_1

Conclusion: significant

There is no difference among 3 groups

$H = 2.854 (2, N = 18)$, $P < 0.05$

The Friedman Test

(It a version of Repeated Measures ANOVA, performed on ordinal (Ranked) data.

$\alpha = 0.05$

Week1	Week2	Week3
27	20	34
2	8	31
14	14	3
18	36	23
7	21	30
9	22	6

week1	week2	week3
2	1	3
1	2	3
2	3	1
1	3	2
1	2	3
2	3	1
R 9	14	13

H_0: No difference in 3 conditions

H_1: Is a difference in 3 conditions

$df = k-1$ (k: no of groups we are comparing)

$= 3-1 = 2$

using χ^2 square table

of χ^2 is grater 5.99

Rank the values in Rows

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum R^2 - 3n(k+1)$$

$$= \frac{12}{(6)(3)(3+1)} (9^2+13^2+(14)^2) - (3)(6)(3+1)$$

$$\chi_r^2 = 2.33$$

As critical value $\chi^2_{critical} < \chi^2_{statics}$

we reject H_1

Conclusion

There is no difference in 3 groups, $\chi_r^2 = 2.33$ $(2, n=6)$, $P > 0.05$

Calculating Required Sample size to Estimate Population Proportion:

In household of 200 ex: previous we found 95% confidence, at least one computer is between 0.701 & 0.819

Confidence = Pointestimate ± Margin of error

$= 0.76 \pm 0.059$

Now what sample size would I need to change margin of error from 0.059 to 0.030 in 95% confidence interval

If we have prior estimate of population proportion exists:

Required Margin of Error

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.76)(1-0.76) \times \left(\frac{1.96}{0.030}\right)^2 = 778.56$$

$$\approx 779$$

If prior estimate of population proportion does not exist:

$$n = 0.25\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.25)\left(\frac{1.96}{0.030}\right)^2$$

$$= 1067.111$$

$$\approx 1068$$

Alpha levels:

If H_0 is rejected when $\alpha = 0.03$

* What could you say if alpha = 0.05?
* " " " " " " = 0.01?