Mathematical Foundations of Artificial Intelligence

STUDY GUIDE

SUBIN VASANTHAN

This guide is prepared by Subin Vasanthan as a personal knowledge reference and learning companion for understanding how core mathematical principles fuel the intelligence behind modern AI and machine learning systems.

Table of Contents

Linear Algebra

Definition: Study the determinants and matrices

Why is linear algebra important?

- Converting the data into matrices can ease out calculation or computations
-

Types of matrices

1. Column Matrix or vector
   Only 1 column but n number of rows

2. Row Matrix or vector
   Only 1 row but n number of columns

3. Rectangular matrix
   Where rows are not equal to columns

4. square matrix
   Where rows are equal to columns

5. Diagonal Matrix
   Where all the elements other than the principal diagonals are zero and at least 1 element in diagonal should be a non-zero.

6. Identity matrix
   Where diagonal elements are 1

7. Scalar matrix
   Where diagonal elements are same values that is 1 or 2 or 3 etc

8. Null matrix
   Where all elements are zero

9. Triangular matrix
   a. Upper Triangular
      Is a square matrix where all the values below principal diagonal is zero.
   b. Lower Triangular
      Is a square matrix where all the values above principal diagonal is zero.

a) Matrix Operations

i. Transpose of a Matrix

A= [1 2
3 4]

A^T= [1 3
2 4]

Properties of a transposed Matrix

1. (A+B)^T= A^T+B^T
2. (AB)^T = B^T A^T
3. (kA)^T= k A^T
4. (A^T)^T = A

ii. Multiplication

A= [1 2
3 4]

B = [1 1
2 1]

AxB= [1x1+2x3 1x1+2=1

3x1+4x2 3x1+4x1]

Probability

Sample space: Collection of all possible outcomes in an experiment

Formulae:

P(e) = Total Observations supporting the count

Total no. of observations in sample space

Ex: Tossing 2 coins and I want the probability that there is at least 1 head

Total obs in the sample space = { (H,H), (T,T), (H,T), (T,H)}

P€ = 3/4

- Probability of null event is 0
- Probability of an entire event is 1
- Probability of an event plus complimentary of probability of an event is 1

Important Formulae

- P(A U B) = P(A) + P(B) – P (A ^ B)

^ is noted as intersection

a) Independent Events

If the occurrence of A doesn't impact the probability of B and vice versa

If A and B are independent events then

- P(A ^ B) = P(A) x P(B)
- P(A U B) = P(A) + P(B) – P(A) x P(B)

P(A^B) = P(A) x P(B)

Eg:
Event A: Getting head on Toss 1
Event B: Getting Tail on Toss 2

P(A^B) = ½ x ½ = ¼

b) Mutually Exclusive

If the events cannot occur at the same time. If event A occurs event B will definitely not occur and vice versa

Mutually exclusive events are not independent events

c) Conditional Probability

Probability calculated on a certain condition

Question:
Probability of event B when we know Event A has already occurred

Eg: Find the probability of getting a spade card when we are told that the randomly picked card is a face card.

Sample space- J clavier, J Spade, J heart, J Diamond similarly for K and Q. So total 12 cards.
So picking spade from the above is 3/12

Important Formulae

P(B/A) = P(B^A) / P(A) Bayes Theorem

Where B is the event and A is the Event that has already occurred

Eg:
From the above example probability of getting a spade card P(A)= 12/52
P(B^A) = 3/52

P(B/A) = (3/52)/(12/52) = ¼

Eg:2
There is bag1 with 4 white and 6 black balls
There is bag2 with 4 white and 3 black balls
One ball is picked and it is found to be black, what is the probability that is it is drawn from bag2

A: Getting a black ball
B: Getting a ball from Bag 2

P(B/A)= P(Getting a ball from Bag 2) / P(Getting a black ball)

P(B/A) = (3/7) / (3/7+6/10)

Eg:2
A Man tells truth 2 out of 3 times. He throws a die and reports the number to be 4. Find the probability that the number is actually 4
=P(Telling the truth & no is 4) / P(reported number is 4)

= P(telling truth) x P(the no is 4) / P(report a number 4)

= (2/3)x (1/6) / P(reports a no.4)
=2/7

P (reports a no is 4) = P (Telling the truth and no is 4) + P(telling a lie and no is not 4)
= (2/3)x(1/6) + (1/3)x(5/6) = 7/18

Calculus - Differentiation

Basic Formulas

d/dx x= 1

d/dx x^2 = 2x

d/dx x^n = n (x^ n-1)

d/dx k = 0

d/dx e^x = e^x

d/dx log(x) = 1/x log will be noted as ln

d/dx kx^2 = k d/dx X^2
= k 2x

d/dx (ax+bx^2) = a (d/dx x) + b (d/dx X^2)
= a + b (2X)

d/dx f(x).g(x) = f(x). d/dx g(x) + g(x). d/dx (f(x)

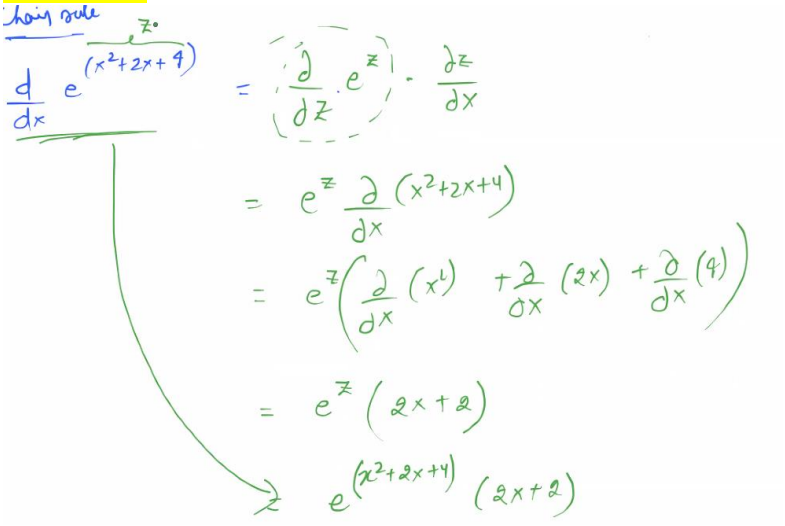
d/dx sinx = cos x
d/dx cos x = - sin x

a) Chain Rule

Formulae

d/dx = (d/dz) x (dz/dx)

EXAMPLE 1:



EXAMPLE 2:

$$\frac{\partial}{\partial x}\left(\sin(x^2) + (\cos(x))^2\right)$$

$$= \frac{d}{dx}\left(\sin \underbrace{x^2}_{z}\right) + \frac{d}{dx}(\cos x)^2$$

$$= \left(\frac{d}{dz}\sin(z) \cdot \frac{\partial z}{\partial x}\right) + \left(\frac{\partial}{\partial y} y^2 \cdot \frac{\partial y}{\partial x}\right)$$

$$= \left(\cos z \cdot (2x)\right) + \left(2y \cdot \frac{\partial(\cos x)}{\partial x}\right)$$

$$= \cos x^2 (2x) + 2(\cos x) \cdot (-\sin x)$$

$$= 2x\cos x^2 - 2\sin x \cos x$$

Q $4(2-x^2)^3 = ?$

$4x^3$

$x^3 \Rightarrow 3x^2$

$x^n = n x^{n-1}$

$4x3(2-x^2)^2 \times (0-2x)$

$-24x(2-x^2)^2$

Statistics

Inferential Statics Duke University Coursera

a) Descriptive statistics

Used to describe data. To understand the data we compute summary of the data.
To compute the summary there are 3 ways

i. Central Tendency

- Mean, Median or mode gives the central value of the data

  Here we will consider an age group for understanding mean. Median and mode

  - Mean

$$\text{Mean} = \frac{\sum_{i=1}^{n} age_i}{n}$$

  This is telling us what is the average age present in the dataset

  - Median
    - First we will sort the age values and we will take the age that is at the center point of the set
    - If the number of observation are odd median is calculated as below

$$\text{value at } \left(\frac{n+1}{2}\right)^{th} \text{ position}$$

    - If the number of observation are even median is calculated as below

$$\text{mean}\left(\frac{n}{2}^{th}, \frac{n}{2}+1^{th}\right) \text{ position}$$

    This shows that if the observed values are of even qnty we take the average o the middle 2 values.

  - Mode

  Mode is the value that has the highest frequency

ii. Dispersion

- Spread in the data
  - Range = (Max value – min value)
  - Variance = Average spread around the mean

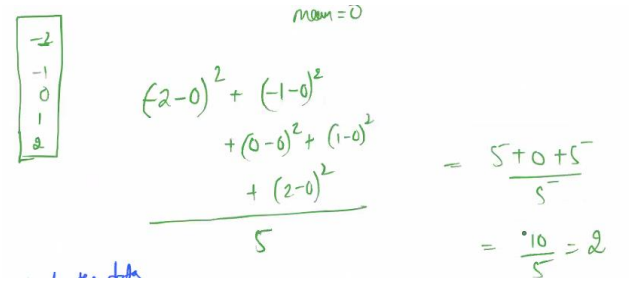
Formulae:

= Average spread around the mean

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

$\bar{x}$ = mean of the data

Example:



- ❖ Standard Deviation = Sqrt (Variance)

Formulae:

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

Suppose we get standard deviation as 10 then physical meaning of the standard deviation is, on an average every value is at distance of 10 units from the mean

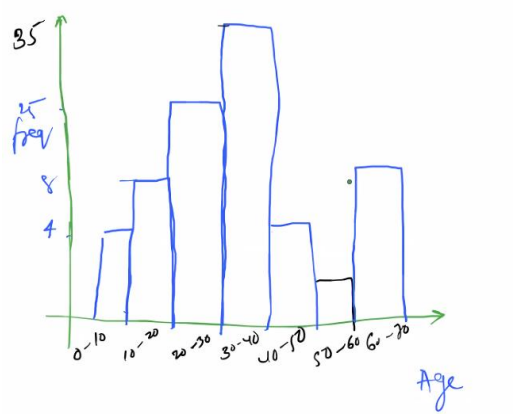
- ❖ Percentage & Percentile
  - If my Percentile in an exam is 75% percentile and there are 100 people, then it means that
    - I am in the top 25 people
    - And there are 75 people below my score

- ❖ IQR (Inter Quartile Range)

  - Is the difference between (75^{th} Percentile value – 25^{th} percentile value)
    - Higher IQR means Larger spread
    - Lower IQR means lower spread

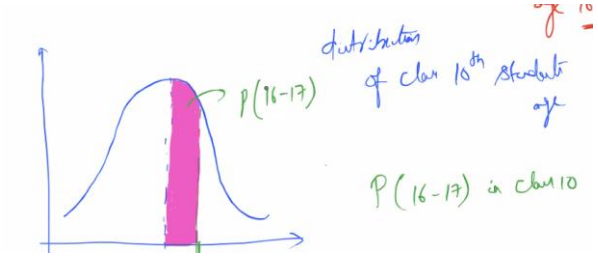
iii. Distribution

- ❖ Frequency plot



- ❖ Probability density function

  - The area under the probability is the probability of finding a student of age 16-17 from a given set of 100 entries.



Formulae:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Note:

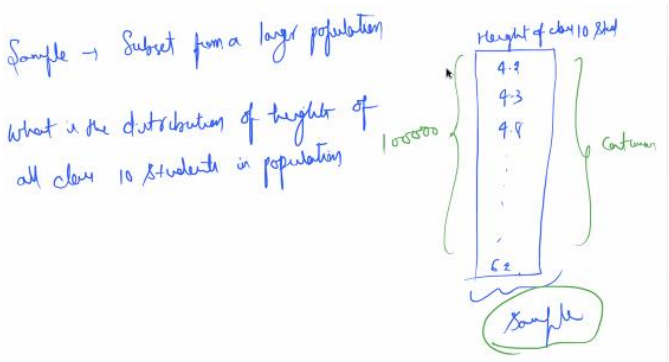
1. Continuous Values
   - That can be value like height of a person eg: 5.5, 5.3, 5.6 etc.

2. Discrete Values
   - It can be thought as a rating value. That is a value from a fixed set. Eg: rating 1,2,3,4,5 or game positions 1, 2 or 3
     a. Ordinal Data Values in the discrete that have an ordering is called categorical values. For example. Rating. Rating 5 is greater than rating 3
     b. Nominal Discrete value without ordering. For example, vaccination.

3. Categorical Values
   - Are values that contain discrete string values For example Gender male or female.

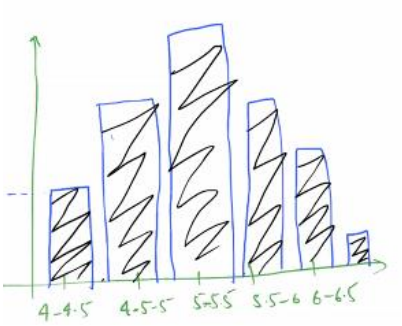
Continuous Data Example

Problem:

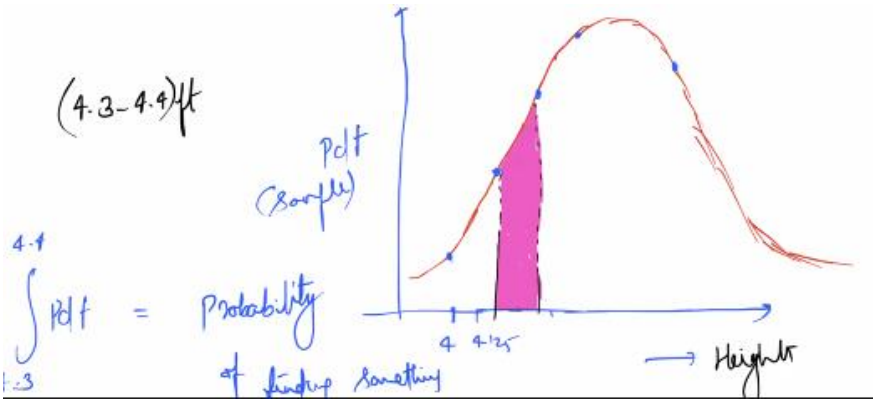


Solution:

Below histogram talks about the distribution

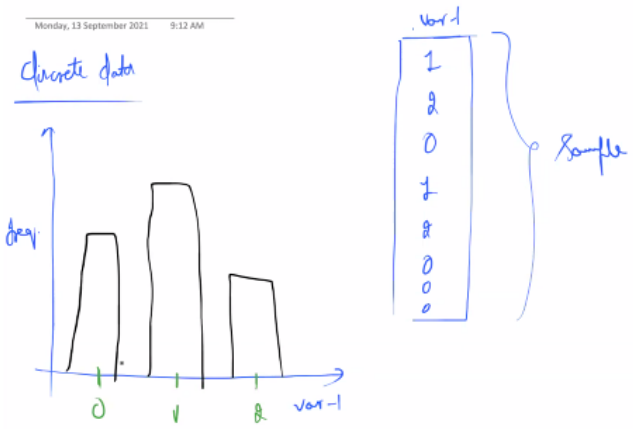


Now I can plot a continuous data curve to estimate the heights of the students of the entire populations.



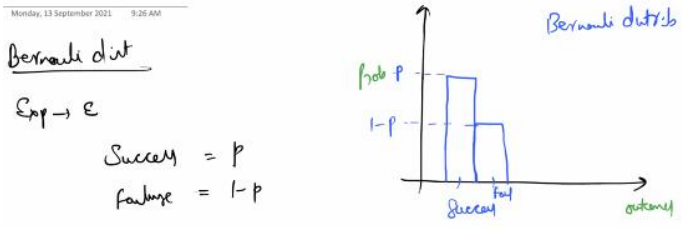
Discrete Data Example

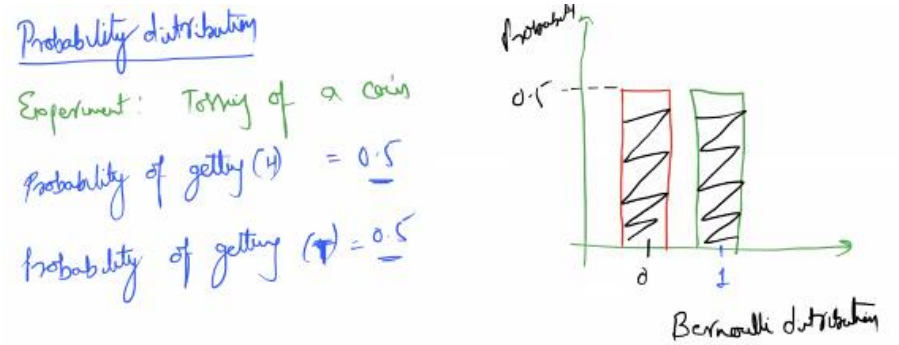
Below plot indicates the distribution of a discrete data

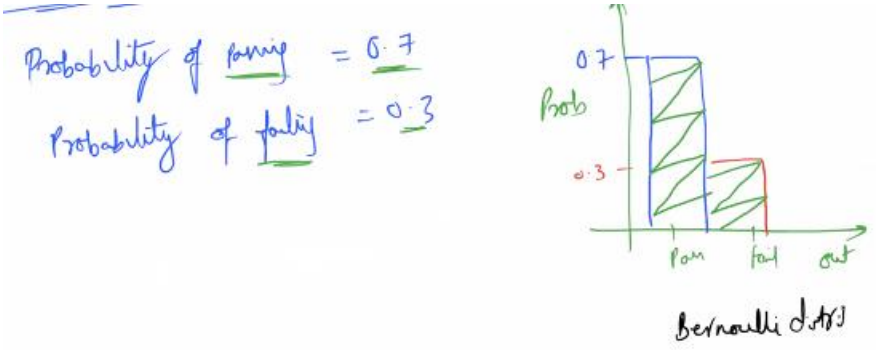


Bernoulli Distribution

It is representation of the probability outcomes of an experiment should have only 2 outcomes which is repeated once





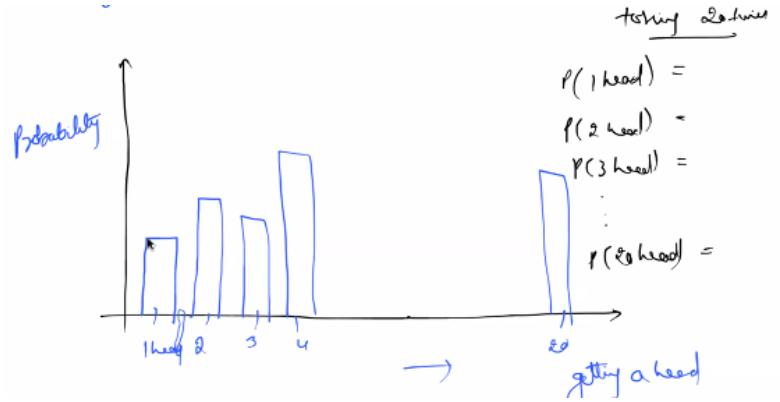


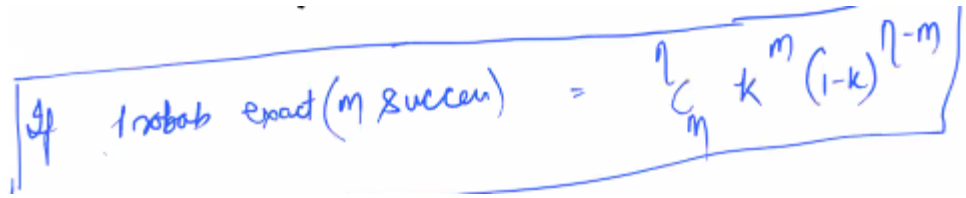
Binomial Distribution

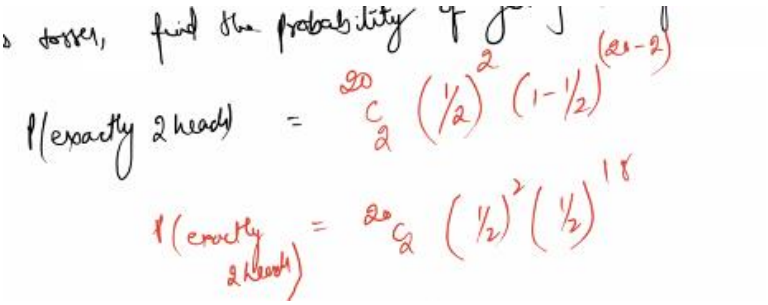
Is the distribution that we obtain when Bernoulli trials are conducted more than 1 times.

Tossing a coin 20 time

P(1 head)
P(2 head)
……
P(20 heads)







Poisson’s Distribution

Check yourself

b) Inferential statistics

We don't understand the data, but we try to get some insights (inferential) from the data.
We need to get sample from a given data and do averaging

i. Random Sampling

- Randomly pick observations from a population.
- Each observation has an equal chance of getting picked

ii. Convenient Sampling

- Picking up a sample from the population on the basis of convenience

iii. Representative Sampling

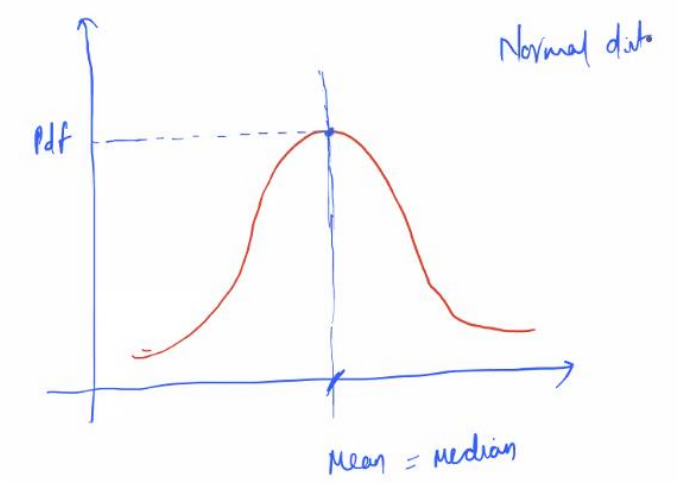
- A technique to create sample that has all possible components of a population

iv. Sampling Distribution

- Is a distribution of sample means.

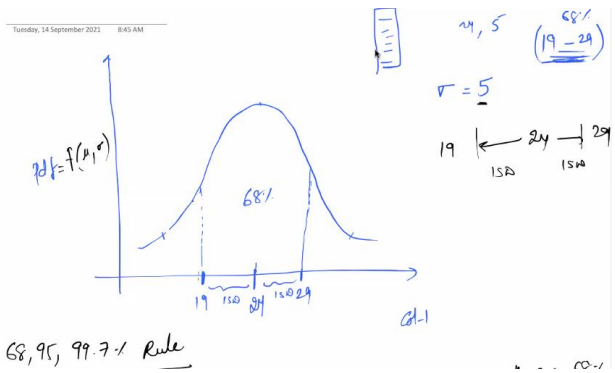
Normal Distribution or Gaussian Distribution

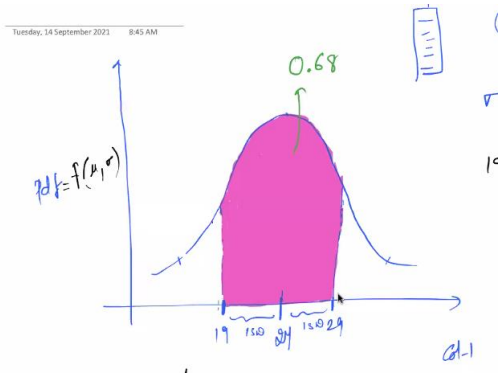
It is a continuous probability distribution which is symmetric in nature and the mean is equal to median in a normal distribution. Also the peak of the distribution of the mean or median point.



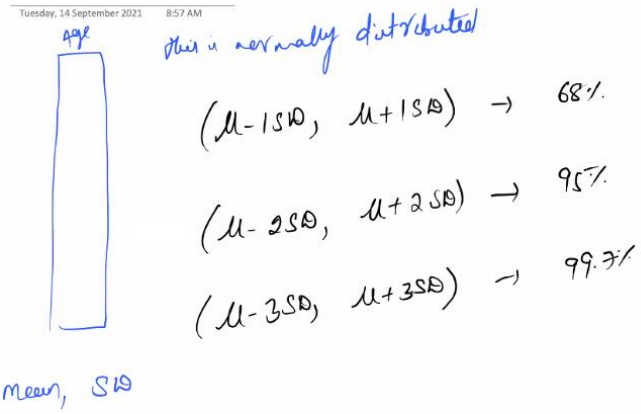
To check if the given distribution is normal distribution we can

1. Check if mean=median=mode this is only a first check
2. Apply the 68,95,99.7% rule In any normal distribution,
    a. within 1 standard deviation (1SD) from the mean there is 68% of the observations or the area would be 0.68





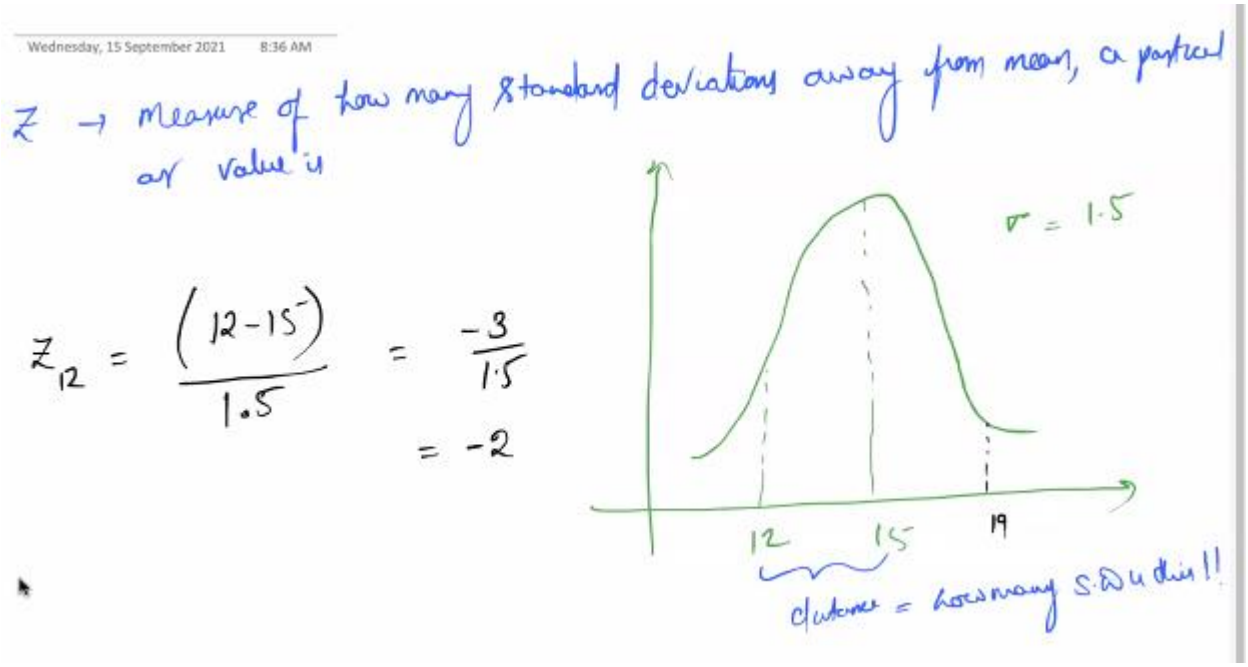
b. similarly, within 2 standard deviation (1SD) from the mean there is 95% of the observations or area would be 0.95
c. similarly, within 3 standard deviation (1SD) from the mean there is 99.7% of the observations or area would be 0.997



- We know that the following
  - Area under the curve is .68 if the Standard deviation is 1
  - Area under the curve is .95 if the Standard deviation is 2
  - Area under the curve is .997 if the Standard deviation is 3

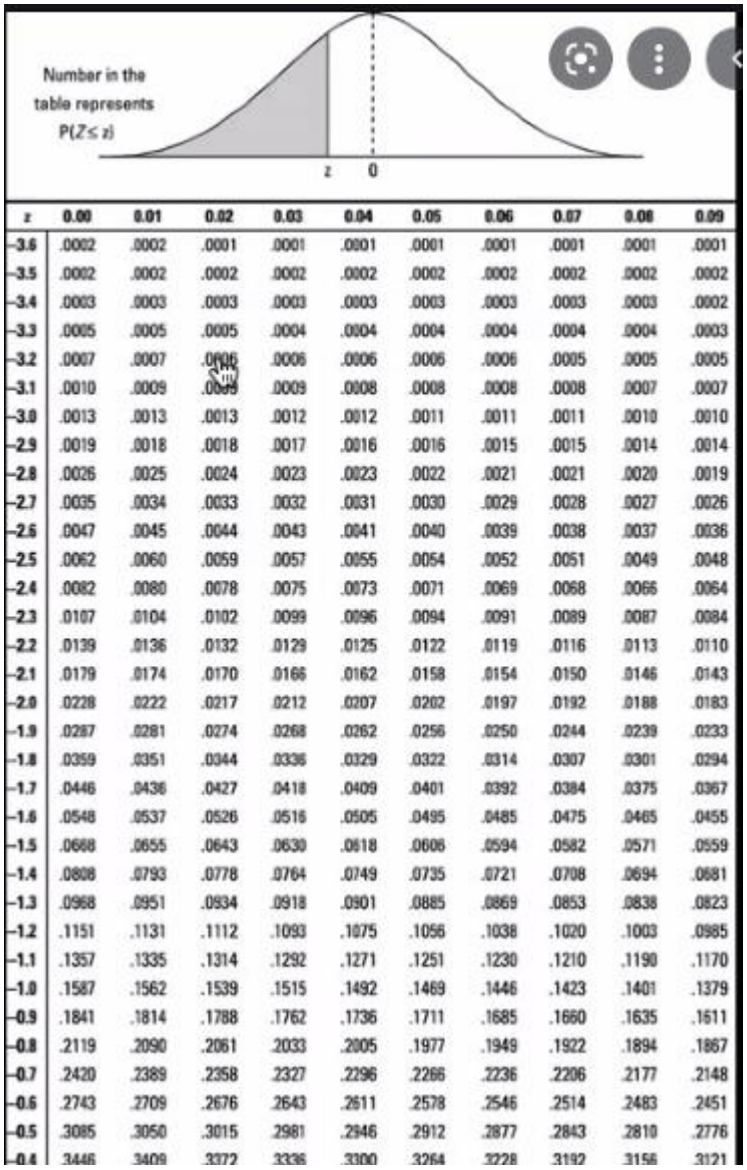
So there are 2 ways to calculate area for other values of S.D

- 1 way is by using python
- The other way is by using Z value this is explained below

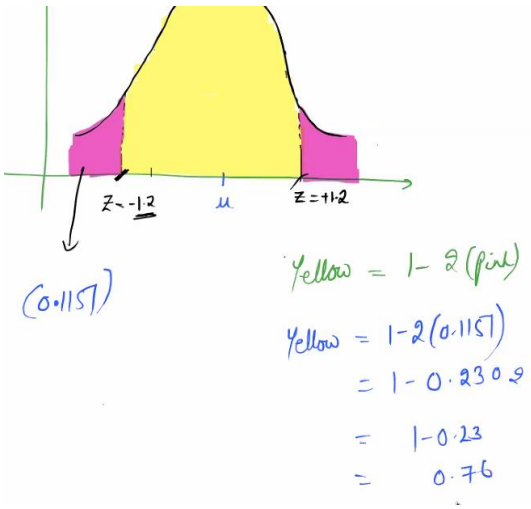


The below area is giving the area under the z value.
Here if they have asked if the SD is1.20 then the area below z is .1151 taken from the table.
Please refer the website for more details. http://www.z-table.com/

The second decimal place is taken as values in the columns and z is the row value.

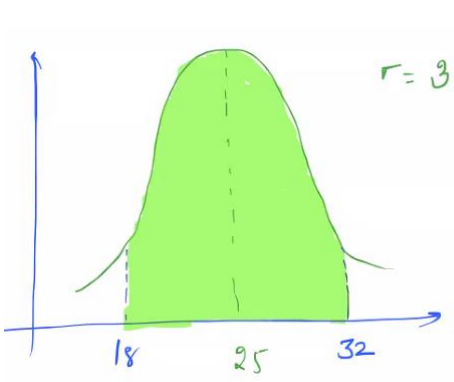


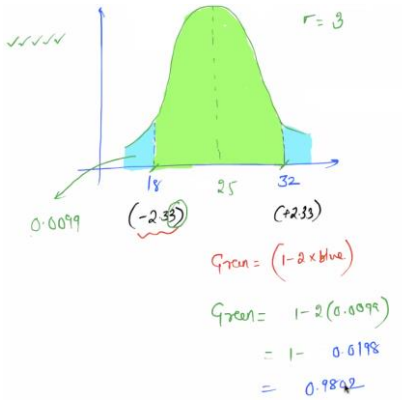
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121



Question:

Calculate the area under the below curve





Z= (18-25)/3

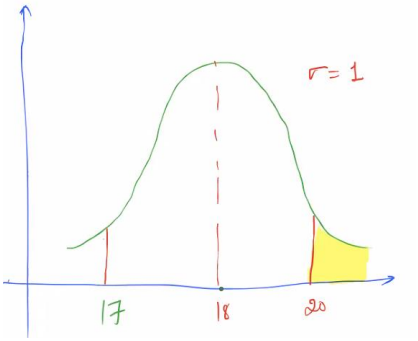
=-2.333333

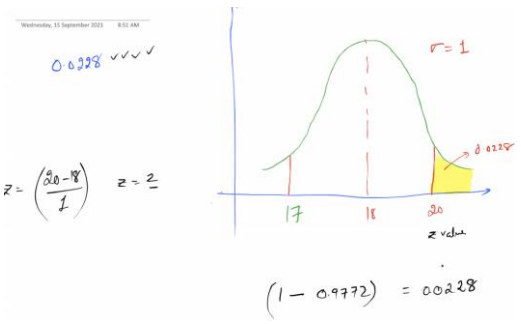
Finding the value from the Z table is .0099

So the area under the curve is (1-2*.0099)= .9802

Question:

Calculate the yellow area under the below curve

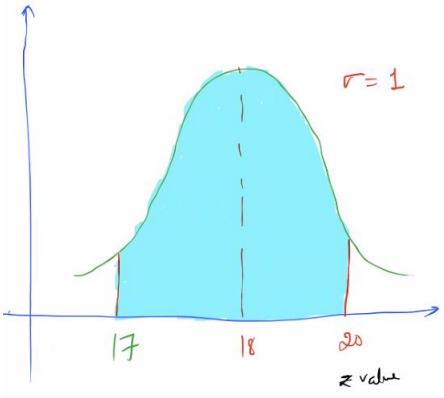


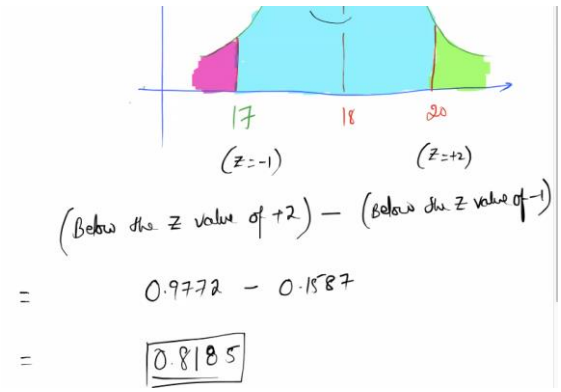


Z= 20-18/1= 2
Finding the value from the Z table is .9772
So area under yellow is 1-.9772= .0228

Question:

Calculate the yellow area under the below curve





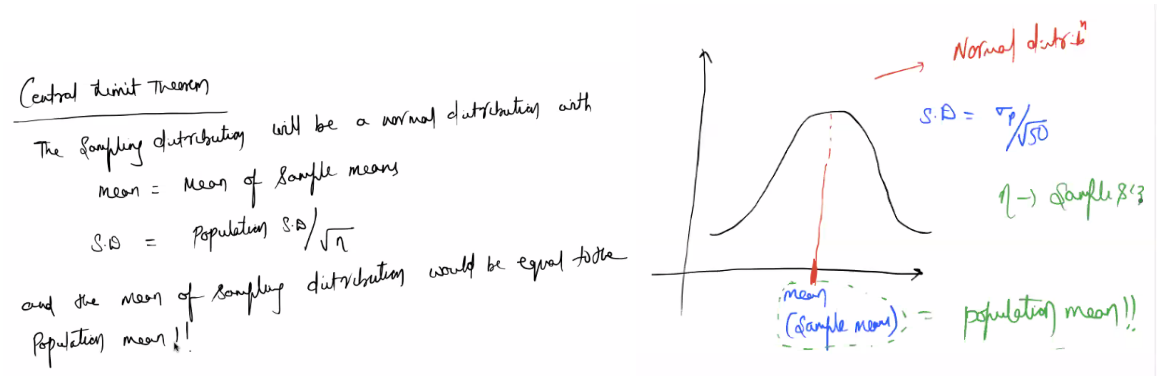
From the above we know the area after 20 is .0228
Z below 17 can be now calculated as 17-18/1= -1
Finding the value from the Z table is .1587
So area on either sides is = 0.228 + .1587=.1815
So area in blue is 1-.1815=.8185

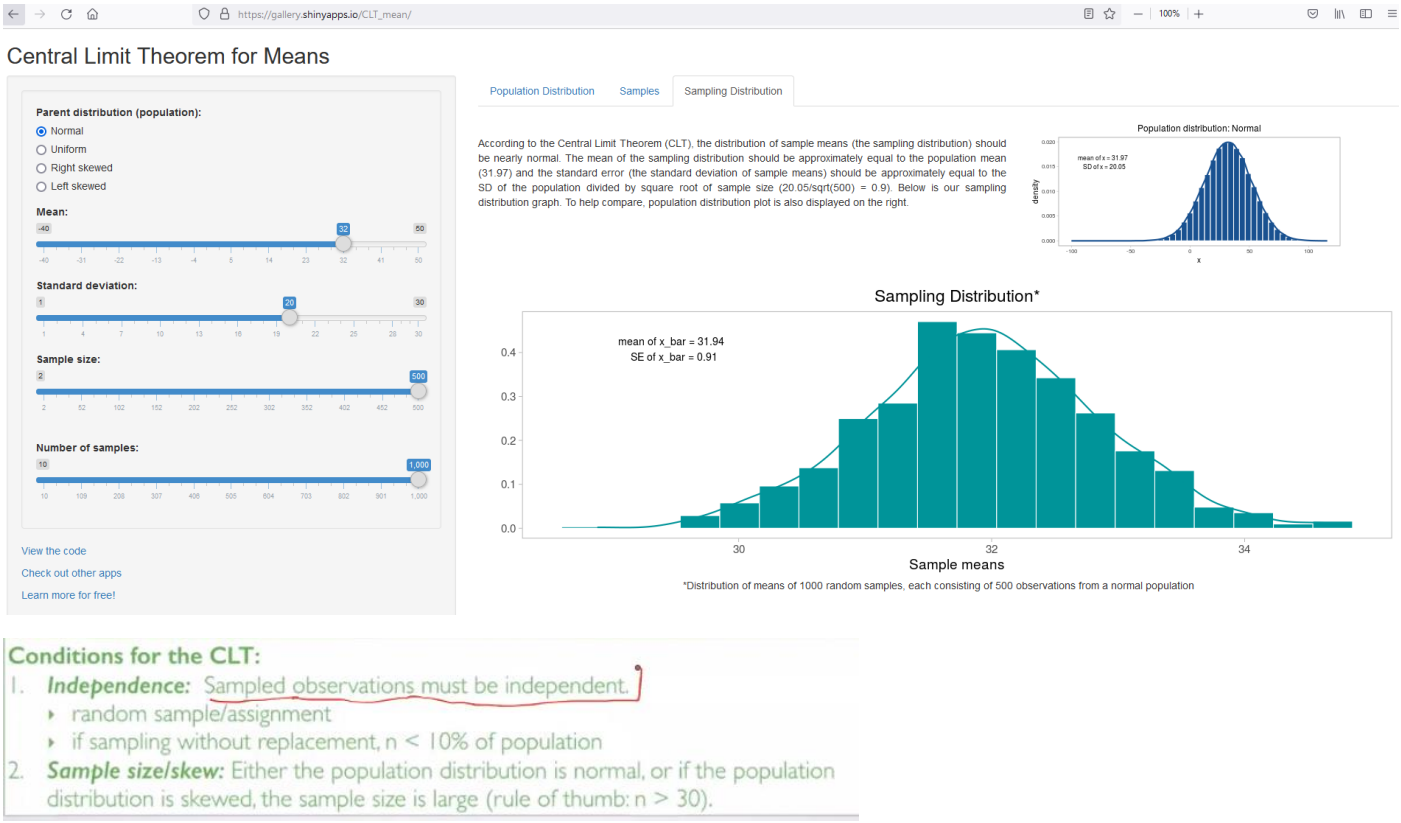
Central Limit Theorem

Check out this website

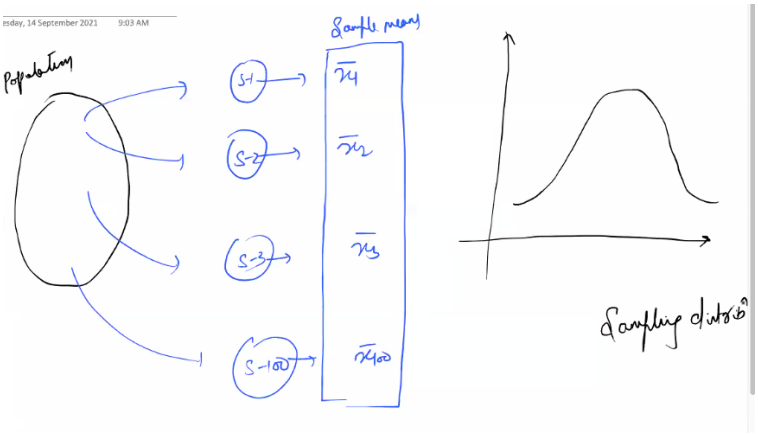
https://gallery.shinyapps.io/CLT_mean/

Assuming we know the Population S.D.



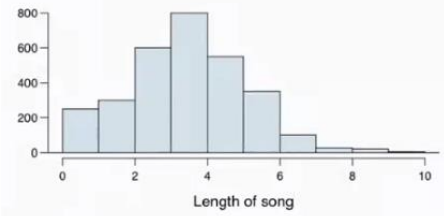


- Below given different samples taken say 50 counts for 1 sample . So there are 100 samples taken S-1, S-2 ….S-100.
- X1, X2, etc are the sample means.

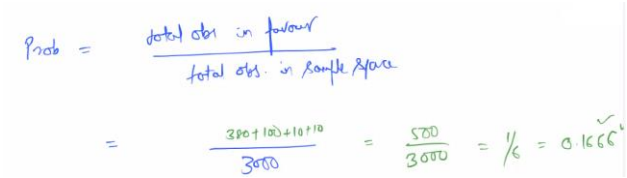


Question considering the below is a normal distribution

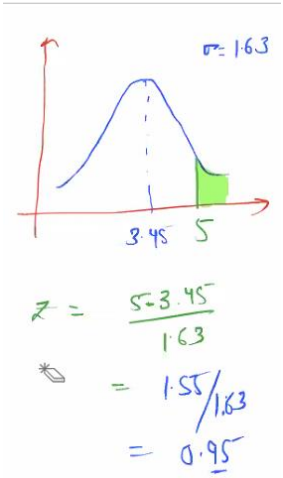
Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



For the above problem we can approximately calculate the probability by counting.



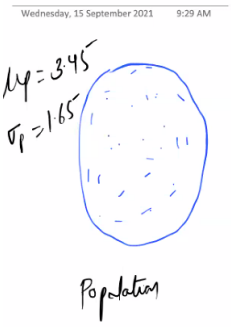
Now let us assume the above is normally distributed. then



Checking the value of z from the z table is 0.8289

So the area is $1 - 0.8289 = 0.1711$

Question considering central limit theorem



I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

$(x_1) + (x_2) + \cdots + (x_{100}) \geq 6 \text{ hours}$

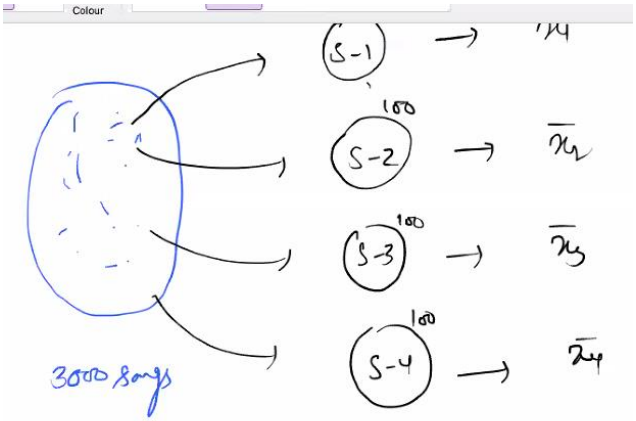
$\sum_{i=1}^{100} x_i \geq 360 \text{ min}$

$\frac{\sum_{i=1}^{100} x_i}{100} \geq \frac{360}{100} \text{ min}$

Mean Song length $\geq$ 3.6 min

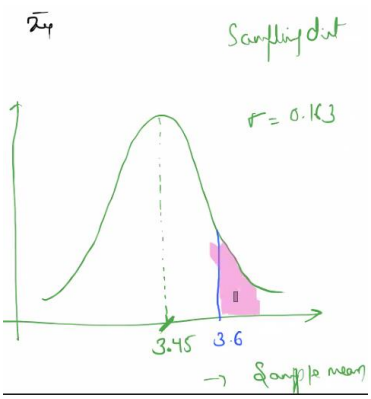
So now we have to calculate the

Probability of finding a playlist / sample of 100 songs where the mean song length is greater than 3.6 min ✓

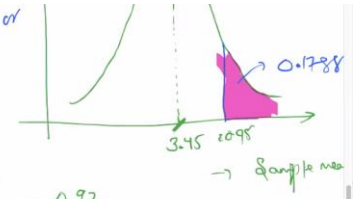


According to central limit theorem the mean will be normally distributed
And the mean is the same as the population mean

In the below standard deviation is calculated as = 1.63/Sqrt(100)=.163



having mean 3.6 mins or greater



$$Z = \frac{3.6 - 3.45}{0.163} = \frac{0.15}{0.163} = 0.92$$

Probability that a randomly created playlist of 100 songs will last the entire drive is 17.88%

Question

According to data, the average IQ of Indians across all age group is 124 with a standard deviation of 50. What is the probability of finding a sample of 50 Indians with mean IQ greater than 140?

Population mean= 124

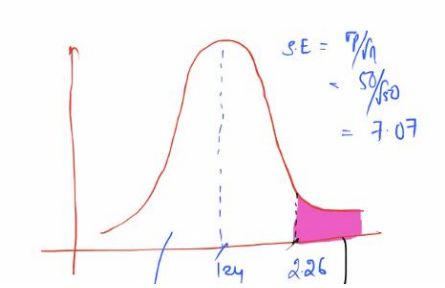
Population SD= 50

Standard deviation is = 50/sqrt(50)= 7.07

Z= (140-124)/7.07= 2.26

From z table z value for 2.26 is .9881

So the answer is 1-0.9881= .0119



Standard Error-- > Standard deviation of sampling distribution.

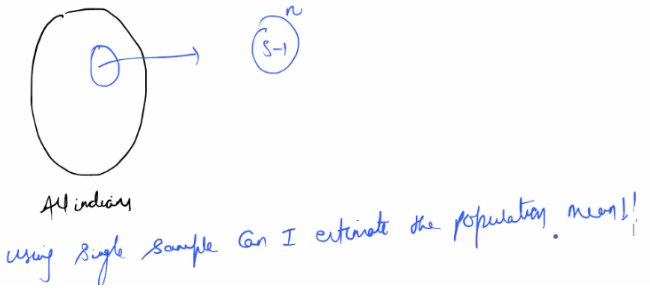
Confidence Intervals

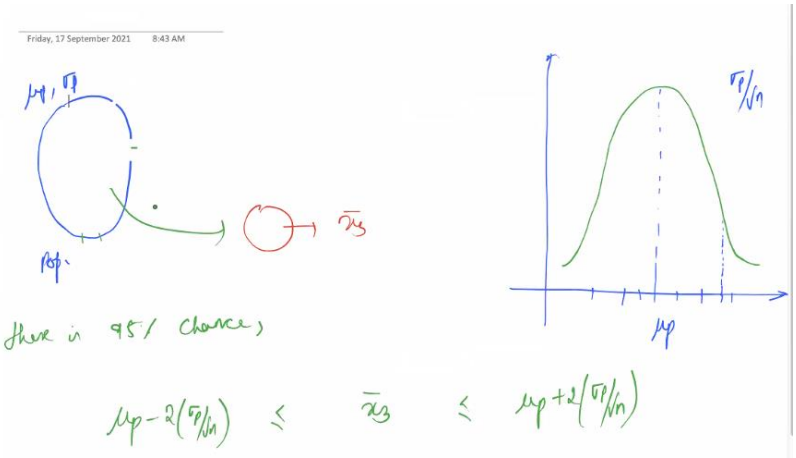
IF I have only 1 single sample. So here we cannot say which is the central mean. So we will have to estimate some interval for population mean.

So in simple terms when a person come to you and say that that

- He is 75% sure that the IQ of the population lies between 120 and 125 Here this is not a good estimate
- He is 95% sure that the IQ of the population lies between 110 and 130 Here this is a good estimate

A good estimation is when the interval is a narrow range and confidence level is high.





$$\mu_p - 2(\sigma_p/\sqrt{n}) \leq \bar{x}_s \leq \mu_p + 2(\sigma_p/\sqrt{n})$$

Subtracting X3 from both sides

$$\left(\mu_p - (2\sigma_p/\sqrt{n})\right) - \bar{x}_s \leq \bar{x}_s - \bar{x}_s \leq \mu_p + (2\sigma_p/\sqrt{n}) - \bar{x}_s$$

$$(\mu_p - \bar{x}_s) - (2\sigma_p/\sqrt{n}) \leq 0 \leq (\mu_p - \bar{x}_s) + 2(\sigma_p/\sqrt{n})$$

Subtracting up from both sides

$$(\mu_p - \bar{x}_s - \mu_p) - (2\sigma_p/\sqrt{n}) \leq \theta - \mu_p \leq (\mu_p - \bar{x}_s - \mu_p) + 2(\sigma_p/\sqrt{n})$$

$$\left(-\bar{x}_s - 2\sigma_p/\sqrt{n}\right) \leq -\mu_p \leq -\bar{x}_s + 2\sigma_p/\sqrt{n}$$

Now removing the minus sign

$$\left(\bar{x}_s + 2\sigma_p/\sqrt{n}\right) \geq \mu_p \geq \bar{x}_s - \left(2\sigma_p/\sqrt{n}\right)$$

$$\boxed{\left(\bar{x}_s - 2\sigma_p/\sqrt{n}\right) \leq \mu_p \leq \left(\bar{x}_s + 2\sigma_p/\sqrt{n}\right)}$$

So now we can say that there is 95% chance that my up is lying between

There is 95% chance

Confidence level

$$\mu_p = \left(\bar{x}_s - 2\cdot\sigma_p/\sqrt{n},\ \bar{x}_s + 2\sigma_p/\sqrt{n}\right)$$

Confidence interval

<u>General way of determining confidence interval</u>

We should already know the following

$\bar{x}$ → Sample mean

σ_p → population standard deviation

n → Sample size

If I have the above values, then I can calculate the

At 95% Confidence level,

$$\mu_p = \left(\bar{x} - 2\left(\sigma_p/\sqrt{n}\right),\ \bar{x} + 2\left(\sigma_p/\sqrt{n}\right)\right)$$

At 90% Confidence level,

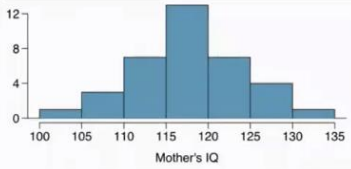
$$\mu_p = \left(\bar{x} - z_{90}\left(\sigma_p/\sqrt{n}\right),\ \bar{x} + z_{90}\left(\sigma_p/\sqrt{n}\right)\right)$$

Now to calculate the Z value for the 90% confidence interval, we have to check the value of Z value 5% (100%-90%)/2 i.e value of .05 from the Z table

So we get a Z value of 1.16 approx for

Question:

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores. The histogram shows the distribution of these data, and also provided are some sample statistics.



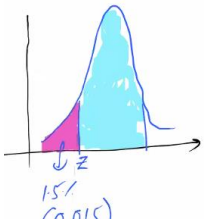
n	36
min	101
mean	118.2
~~sd~~	~~6.5~~ $\sigma_p = 6.5$
max	131

Question: Estimate the Confidence interval for the average IQ of mothers of gifted children at 97% Confidence level

Soln

Confidence Interval. =

$$\left(\bar{x} - z_{97}\left(\sigma_p/\sqrt{n}\right),\ \bar{x} + z_{97}\left(\sigma_p/\sqrt{n}\right)\right)$$



Z value of .015 is 2.17 from the Z table.

$$\left(118.2 - z_{97}\left(1.083\right),\ 118.2 + z_{97}\left(1.083\right)\right)$$

$$= \left(118.2 - 2.17\left(1.083\right),\ 118.2 + 2.17\left(1.083\right)\right)$$

$$= \left(118.2 - 2.35,\ 118.2 + 2.35\right)$$

$$= \left(115.85,\ 120.55\right)$$

Question:

A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show (an American late-night TV show). The standard error of this estimate is 0.014. Estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

$\bar{x}$ → Sample mean
σ_p → population standard deviation
n → Sample size

Sample size = 1099

Standard Error=0.014

Sample mean= 33%= .33

$$\text{Standard Error} = \left(\frac{\sigma_p}{\sqrt{n}}\right) = 0.014$$

Confidence Interval,

$$\left(\bar{x} - z_{95}(S.E.),\ \bar{x} + z_{95}(S.E.)\right)$$

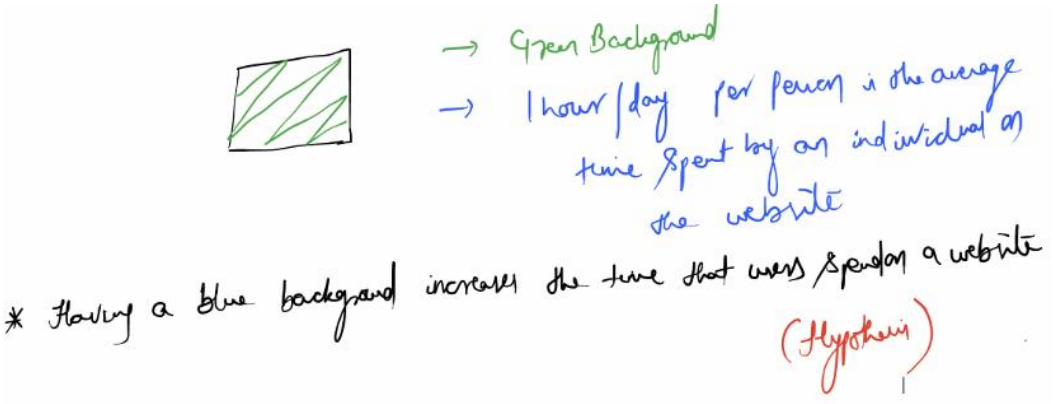
$$\left(\bar{x} - 2(0.014),\ \bar{x} + 2(0.014)\right)$$

$$(0.33 - 0.028,\ 0.33 + 0.028)$$

$$(0.302,\ 0.358)$$

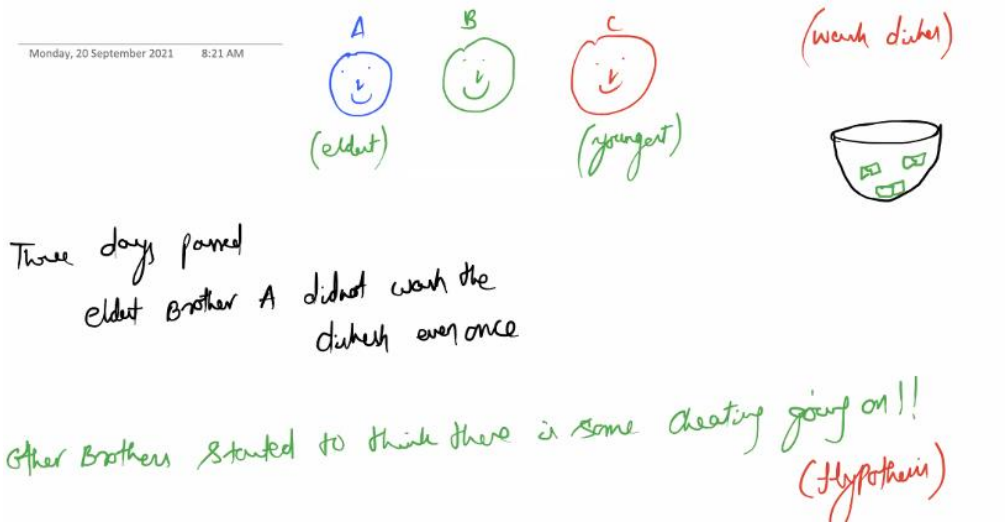
Hypothesis Testing

Testing of an hypothesis and concluding if the hypothesis is True / False

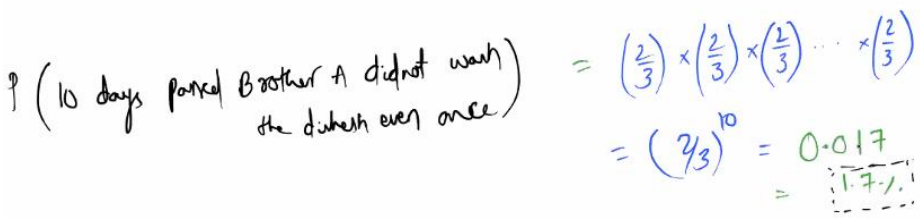


A study or survey for the above Hypothesis will be conducted

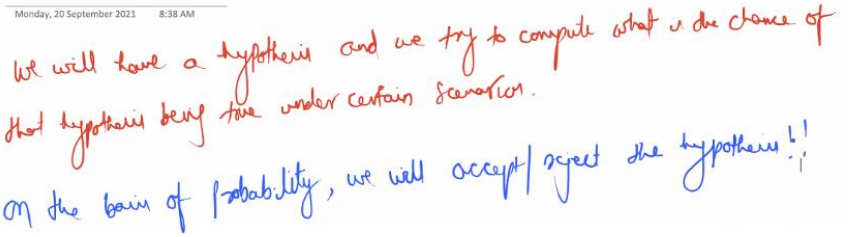
Scenario 2



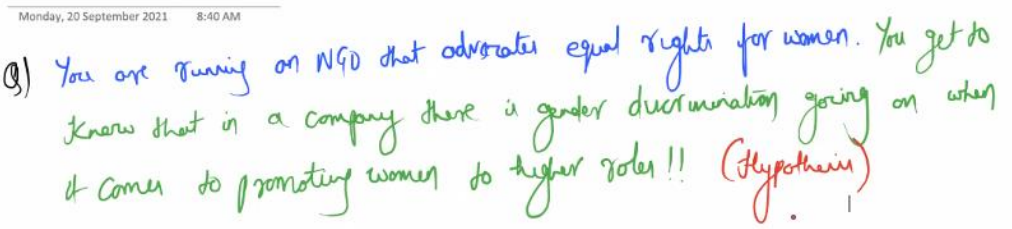
Here in 3 days there is only 30% chance that brother would be correct. 30% cannot be taken a goof probability. Now let's see the probability for 10 days

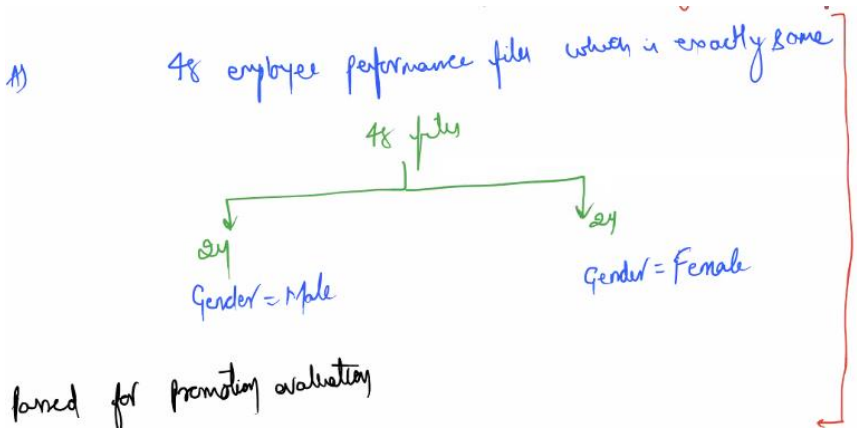


Here we have convincing reasons to believe that there is cheating going on with 1.7%



Scenario 3





The result what we got is the below

data

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

Does the above data provide evidence of gender discrimination???

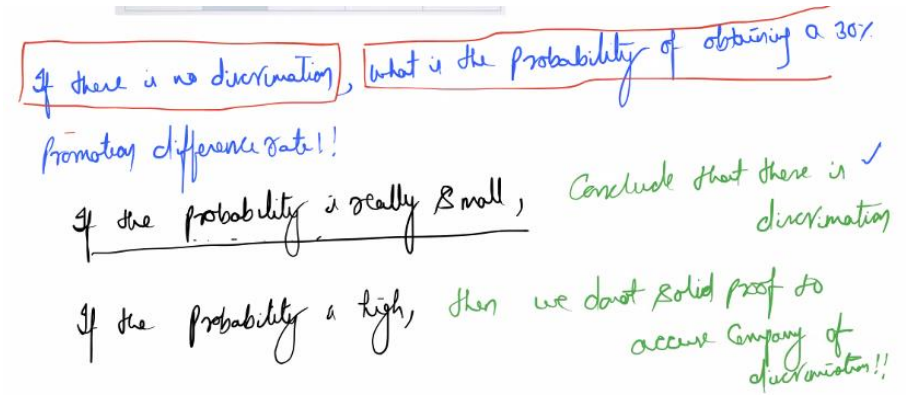
Solution

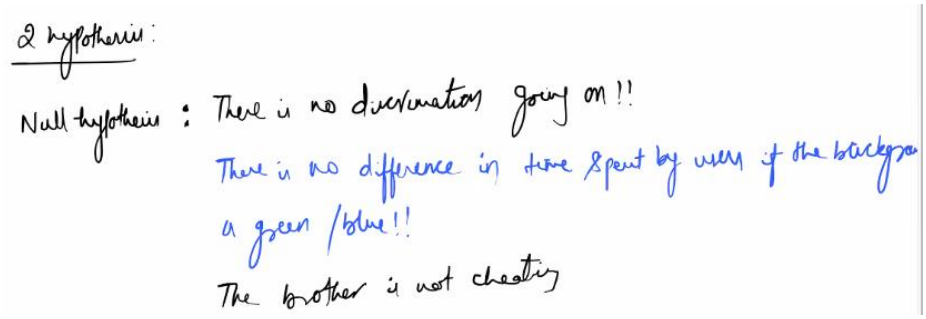
Males got promoted % = 21/24 == 87.5%

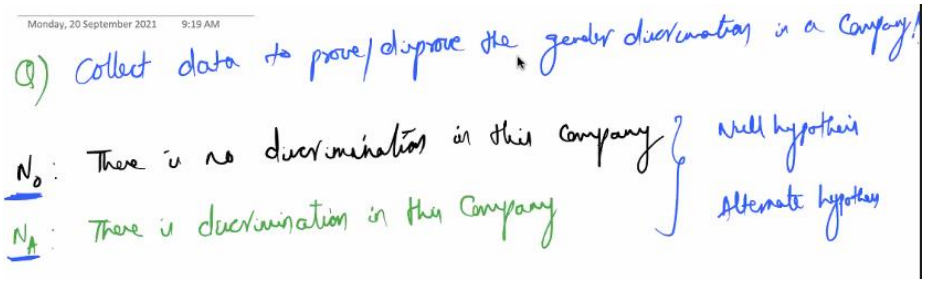
Females got promoted % = 14/24 = 58.33%

So there is a difference of roughly 30%

So if there is no discrimination what is the probability of obtaining a 30% promotion difference.

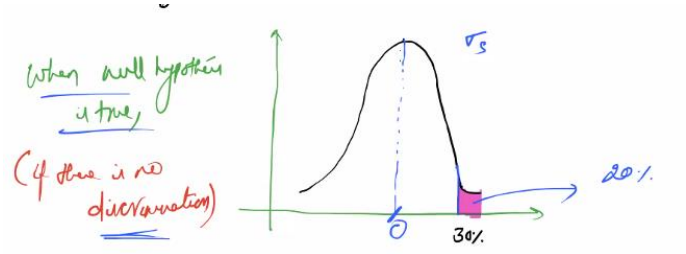






If null hypothesis was true,

If this survey was conducted multiple times



Tuesday, 21 September 2021 8:07 AM

observed data

N_0: There is no gender discrimination

N_A: There is gender discrimination going on

Assuming null hypothesis is true, find out the probability of the observed event happening

If P(observed event) < Significance level (α) → Reject null hypothesis or accept alternate hypothesis

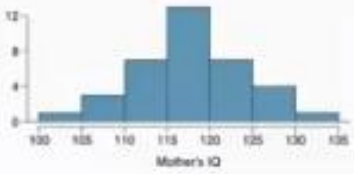
If P(observed event) > Significance level (α) → fail to reject null hypothesis or accept null hypothesis

Most commonly used Significance level 5%

Question:

Perform a hypothesis test to evaluate if these data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large, which is 100. Use a significance level of 0.01.

Population S.D = 6.5



n	36
min	101
mean	118.2
max	131

Mean IQ of the mothers of the gifted children in a sample of 36 is 118.2

Average IQ score of Population = 100

Standard deviation = 6.5

Solution

N_0 : There is no difference in the average IQ scores of general people in the population and the average IQ score of mother of gifted children

i.e. $\mu_{\text{IQ score mothers of gifted child}} = \mu_{\text{IQ score of general population}} = 100$

N_A : There is difference in the average IQ score of general people and the avg. IQ score of mothers of gifted children

i.e. $\mu_{GCM\,IQ} \neq \mu_{GP} \neq 100$

Tuesday, 21 September 2021 8:26 AM

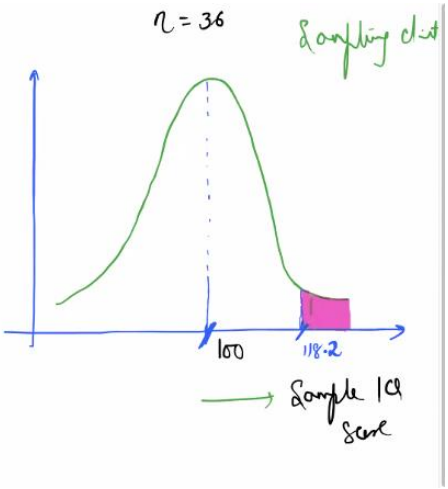
Assuming null hypothesis is true

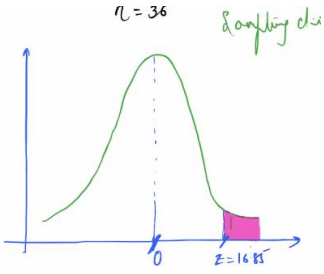
$\mu_{GM\,IQ} = 100$

$Z\ score = \frac{118.2 - 100}{(\sigma_p/\sqrt{n})}$

$= \frac{118.2 - 100}{(6.5/\sqrt{36})}$

$= 16.85$





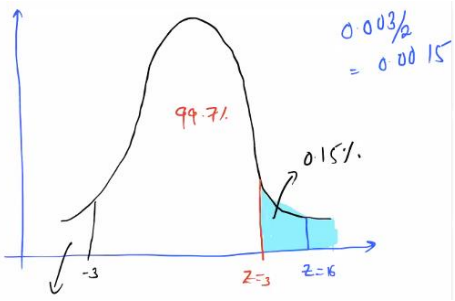
If

$P(Z\ value \geq 16.85) < 0.01$ Reject null hypothesis

$P(Z\ value \geq 16.85) > 0.01$ Fail to reject null hypothesis

Tuesday, 21 September 2021 8:34 AM

Above z = 3 there is 0.15% area



So even at Z=3 (no need to even calculate z=16 as even at Z=3 we get such a lesser %) we have only .15% which is less than .01 or 1%

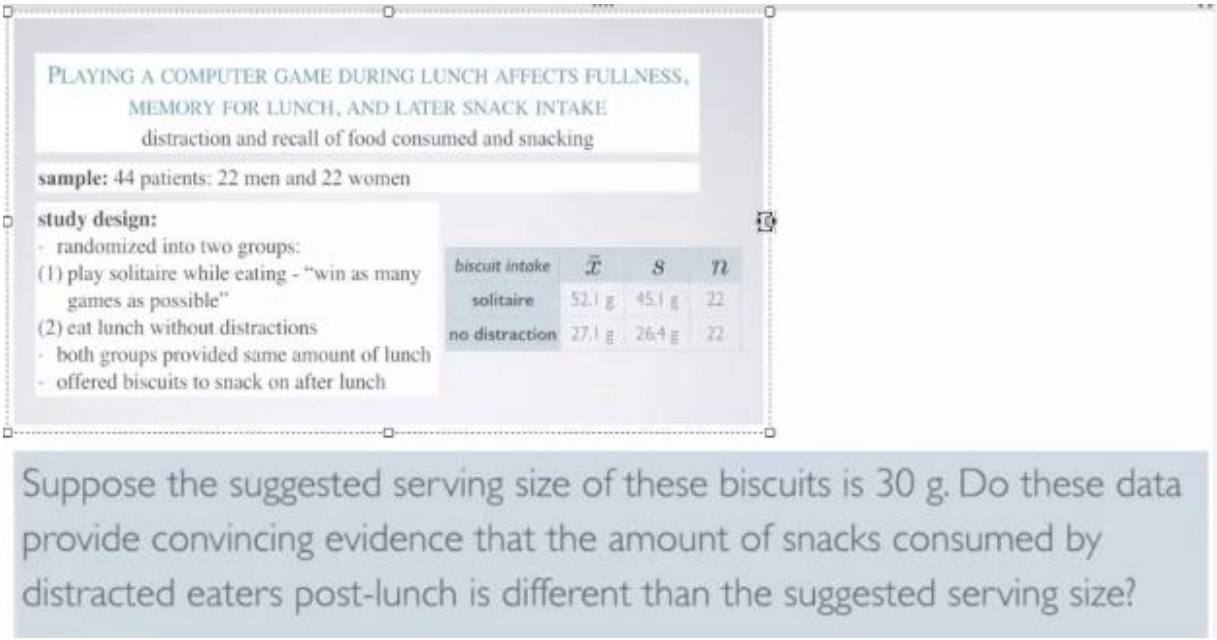
So we will reject null hypothesis

Conclusion

Reject null hypothesis in favour of alternative.

We have convincing evidence to prove that the mothers of gifted children exhibit different average population IQ score than the rest of the population

Question:



PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

Solution

We write down the two scenarios

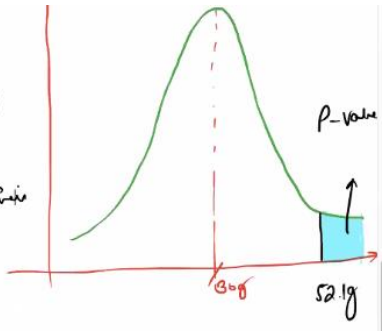
Null Hypothesis There is no difference in the Snack consumed by the people post lunch with playing solitaire during lunch and the people having lunch without any distraction

Alternate hypothesis ⍰ There is a difference in the Snack consumed by the people post lunch with playing solitaire during lunch and the people having lunch without any distraction

Assuming null hypothesis true,

P-value < 2.5% → Reject null hypo

P-value > 2.5% → fail to reject null hypothesis



$$= \frac{52.1 - 30g}{(45.1/\sqrt{22})}$$

$$= \frac{22.1}{9.6} = 2.3$$

P(observed data) = 1% < 2.5% → Reject null hypothesis

Conclude

There is convincing evidence to suggest that the distracted eaters consume a different Snack amount than the suggested serving size!!

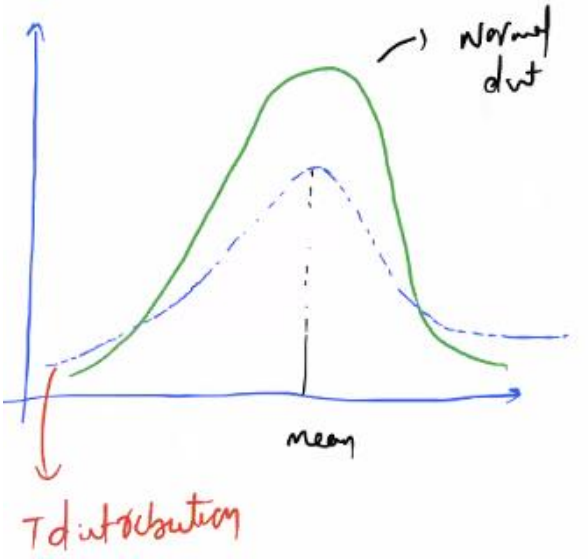
How to Solve this when ∇_p is Unknown

(T- distribution) → (∇_p is unknown)

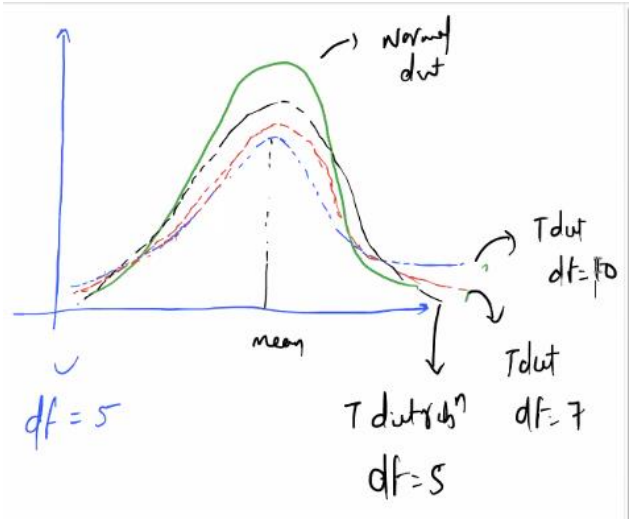
T-Distribution

https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

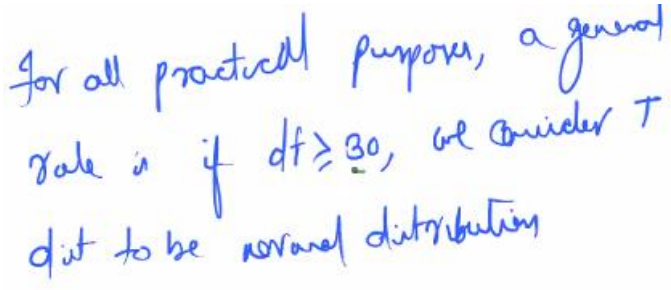
Is like Normal distribution where the tails are boarder and squeezed down

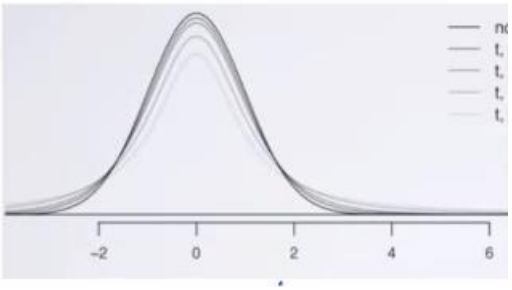


Every T-distribution has a factor called degrees of freedom (df)

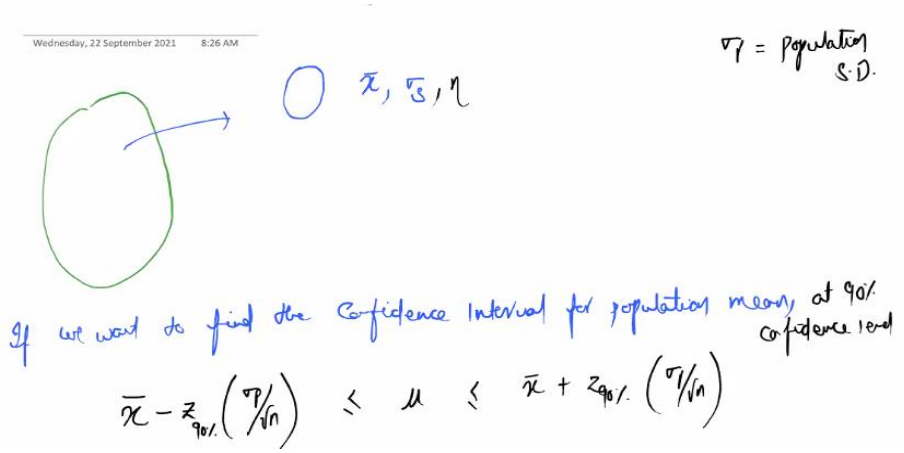


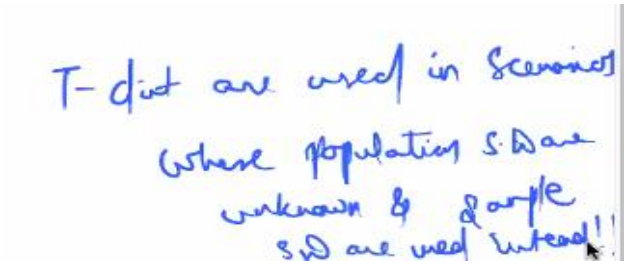
If the degree is a positive number between 1 and infinity. If the df tends to infinity, then the T-distribution becomes normal distribution.



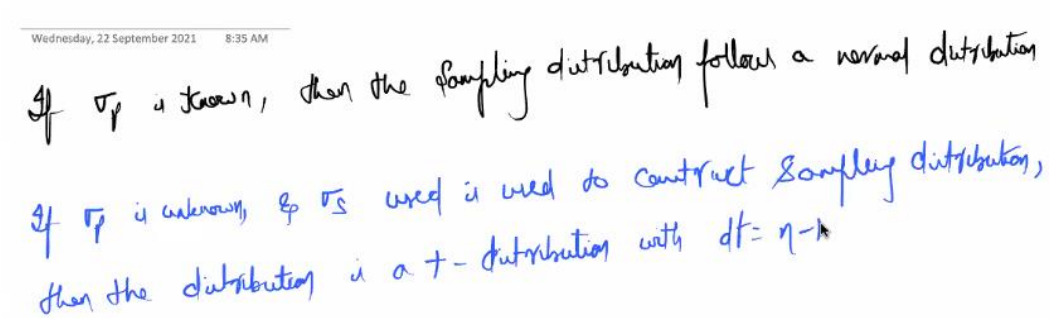


Formulae





Conclusion



Question:

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Estimate the average after-lunch snack consumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval.

Solution

$$\bar{x} - z_{95\%}\left(\sigma/\sqrt{n}\right) \leq \mu \leq \bar{x} + z_{95\%}\left(\sigma/\sqrt{n}\right)$$

$$\bar{x} - t^{df=21}_{95\%}\left(s/\sqrt{n}\right) \leq \mu \leq \bar{x} + t^{df=21}_{95\%}\left(s/\sqrt{n}\right)$$

$$52.1 - t^*\left(\frac{45.1}{\sqrt{22}}\right) \leq \mu \leq 52.1 + t^*\left(\frac{45.1}{\sqrt{22}}\right)$$

$$52.1 - t^*(9.6) \leq \mu \leq 52.1 + t^*(9.6)$$

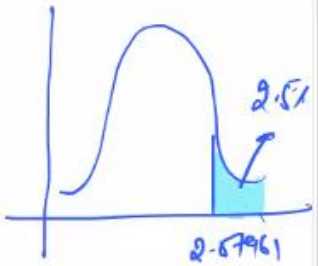
$(-t^*, t^*)$ → In a t dist with df=21, $-t^*$ & t^* are the value b/w there is 95% area

y, 22 September 2021 8:47 AM

→ Areas

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676

→ +value



$$52.1 - (2.07)(9.6) \leq \mu \leq 52.1 + (2.07)(9.6)$$

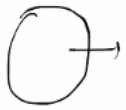
$$32.2 \leq \mu \leq 71.9$$

At 95% Confidence interval

Question:

A sample of 50 college students were asked how many exclusive relationships they've been in so far. The students in the sample had an average of 3.2 exclusive relationships, with a standard deviation of 1.74. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average number of exclusive relationships based on this sample using a 95% confidence interval.

Solution



$\bar{x} = 3.2$

$\sigma_s = 1.74$

$n = 50$

$$\bar{x} - z_{95\%}\left(\sigma/\sqrt{n}\right) \leq \mu \leq \bar{x} + z_{95\%}\left(\sigma/\sqrt{n}\right)$$

$$\bar{x} - t^{df=49}_{95\%}\left(\sigma_s/\sqrt{n}\right) \leq \mu \leq \bar{x} + t^{df=49}_{95\%}\left(\sigma_s/\sqrt{n}\right)$$

Here since we cannot df for 49 and it can be used only used for values less than 30, so we will use the Z table instead

$$\bar{x} - z_{95\%}\left(\sigma_s/\sqrt{n}\right) \leq \mu \leq \bar{x} + z_{95\%}\left(\sigma_s/\sqrt{n}\right)$$

$$3.2 - z^*\left(\frac{1.74}{\sqrt{50}}\right) \leq \mu \leq 3.2 + z^*\left(\frac{1.74}{\sqrt{50}}\right)$$

$$3.2 - 1.96\left(\frac{1.74}{\sqrt{50}}\right) \leq \mu \leq 3.2 + 1.96\left(\frac{1.74}{\sqrt{50}}\right)$$

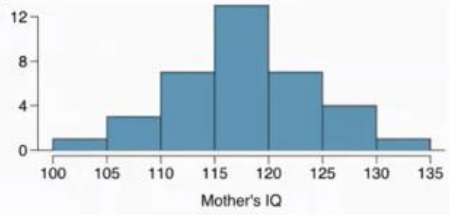
$$3.2 - 0.494 \leq \mu \leq 3.2 + 0.494$$

$$2.706 \leq \mu \leq 3.694$$

Question:

Wednesday, 22 September 2021 9:12 AM

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores. The histogram shows the distribution of these data, and also provided are some sample statistics.



n	20
min	101
mean	118.2
sd	6.5
max	131

Q) Find out Confidence Interval for the average IQ score of mothers of gifted children at 90% Confidence level

Solution

$$\bar{x} - t^{df=19}_{90\%}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{x} + t^{df=19}_{90\%}\left(\frac{s}{\sqrt{n}}\right)$$ (110.46, 125.932) ✓

$$118.2 - t^*\left(\frac{6.5}{\sqrt{20}}\right) \le \mu \le 118.2 + t^*\left(\frac{6.5}{\sqrt{20}}\right)$$ (106.884, 129.57)

$$118.2 - t^*(1.453) \le \mu < 118.2 + t^*(1.453)$$ (113.85, 122.54)

$$118.2 - 1.7291(1.453) \le \mu < 118.2 + 1.7291(1.453)$$

$$115.68 \le \mu < 120.71$$

Question:

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

Solution

If no significance level is given, we will use 5% as default

distracted eaters post-lunch is different than the suggested serving size?

N_0: There is no difference b/w the after lunch snack intake of distracted eaters and the suggested serving size

$\mu_{DE} = \mu_{SS} = 30\,g$

N_a: There is some difference b/w the after snack intake of distracted eaters when compared to suggested serving size

$\mu_{DE} \neq (\mu_{SS} = 30\,g)$

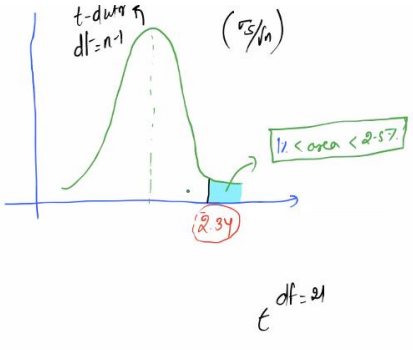
Wednesday, 22 September 2021 9:42 AM

If null hypothesis is true,

$$t \text{ Score} = \frac{52.1 - 30\,g}{\left(s/\sqrt{n}\right)}$$

$$= \left(\frac{22.1}{45.1/\sqrt{22}}\right)$$

$$= 2.34$$



$= 2.34$

Reject null hypothesis & say that we have convincing evidence that suggests distracted eaters on an average consume more than suggested serving size!

Question:

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:

(1) play solitaire while eating - "win as many games as possible"

(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size? NO

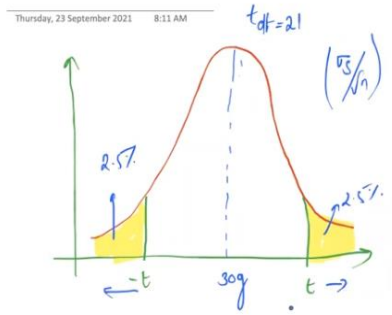
Solution

N_0 : There is no difference in the avg. Snack intake of non distracted eaters and the Suggested serving Size.

$\mu_{nd} = \mu_{ss} = 30g$

N_A :

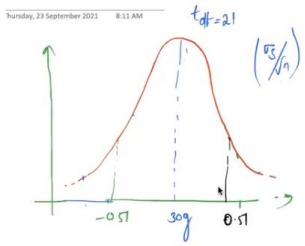
$\mu_{nd} \neq (\mu_{ss} = 30g)$



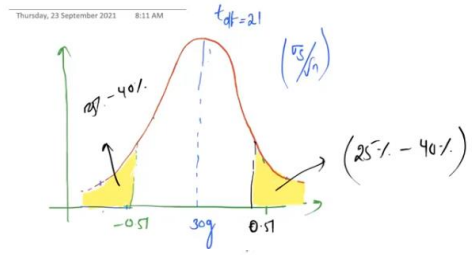
t value $= \frac{27.1 - 30}{26.4/\sqrt{22}}$

$= -0.51$

If area below $t = -0.51 <$ 2.5% Reject null hypothesis



From T Table the area when looked lies between .256580 and .685954 for the degrees of freedom 21. So the area values are between 40% to 25%



2.5%

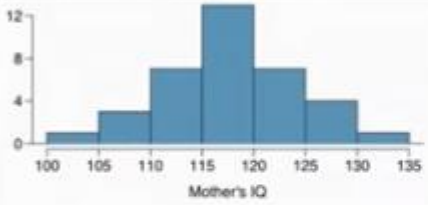
25% < area < 40%

Area below $t = -0.51$

Since the area is greater than 2.5% we fail to reject null hypothesis!

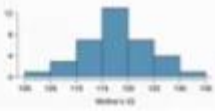
Question:

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores. The histogram shows the distribution of these data, and also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

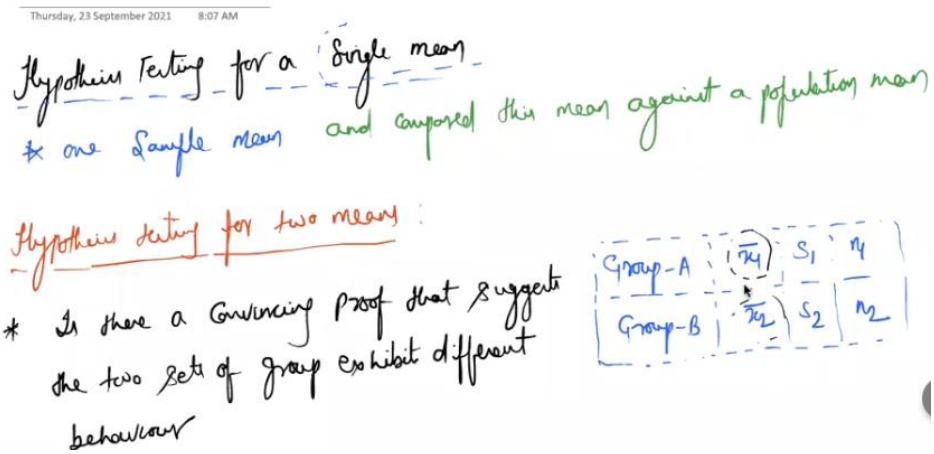
Perform a hypothesis test to evaluate if these data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large, which is 100. Use a significance level of 0.01.



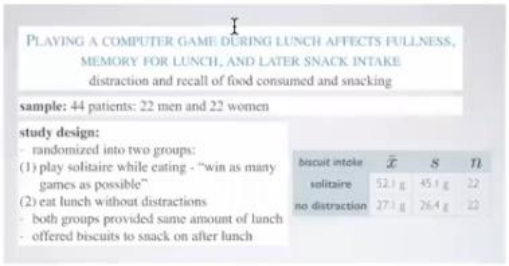
n	36
min	101
mean	118.2
sd	6.5
max	131

Solution

Same as the above solution



Question:



PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:
(1) play solitaire while eating - "win as many games as possible"
(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Q) Does this data provide convincing evidence of different average snack intakes of people who eat lunch distracted and people who eat lunch focused

$$\mu_{dist} = / \neq \mu_{non\text{-}distr}$$

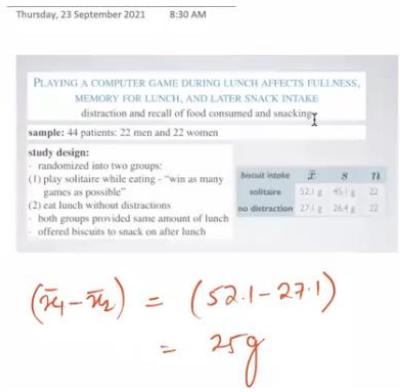
Solution

N_0 : There is no difference in population avg. Snack intake of people who eat lunch focused, and people who eat lunch distracted

i.e. $\mu_f = \mu_d \quad (\mu_f - \mu_d = 0)$

N_A : There is difference

i.e. $\mu_f \neq \mu_d \quad (\mu_f - \mu_d \neq 0)$



Thursday, 23 September 2021 8:30 AM

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

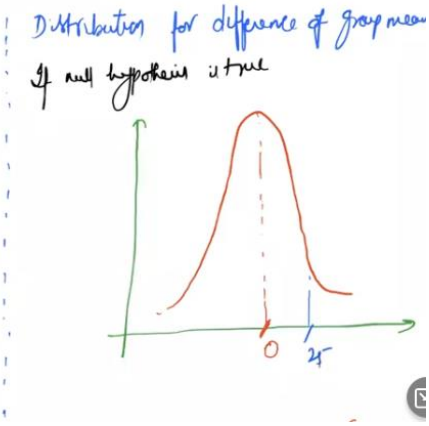
sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:
(1) play solitaire while eating - "win as many games as possible"
(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

$(\bar{x}_1 - \bar{x}_2) = (52.1 - 27.1)$

$= 25g$



observed diff = $(\bar{x}_1 - \bar{x}_2)$

$\sigma_{\bar{x}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

df = min $(n_1 - 1, n_2 - 1)$

$\left(\frac{\sigma_s}{\sqrt{n}}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{45.1^2}{22} + \frac{26^2}{22}}$

$= \sqrt{92.45 + 31.68}$

$= \sqrt{124.13} \approx 11.14$

dof = $\min(n_1 - 1, n_2 - 1)$

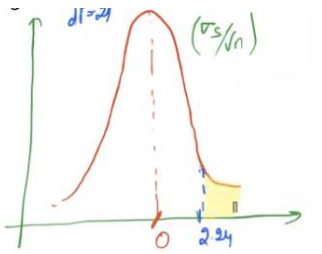
$= \min(22 - 1, 22 - 1)$

$= \min(21, 21) = 21$

t score = $\left(\frac{\bar{x} - 0}{\sigma_s/\sqrt{n}}\right)$

$= \frac{\bar{x}}{(\sigma_s/\sqrt{n})}$

$= \frac{25}{11.14} = 2.24$



Now if the area above the t value of 2.24 is above 2.5 %, I will fail to relate null hypothesis and if the area is less than 2.5% I will reject null hypothesis

Now from the T table we get the area lies between 2.5 % and 1 %

$1\% < \text{area} < 2.5\%$

The area is below 2.5% hence we reject null hypothesis

Formulae for the difference

Thursday, 23 September 2021 8:41 AM

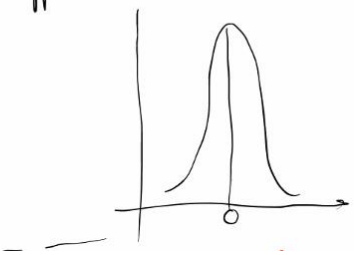
two groups, perform hypothesis testing on these two groups

* we do hypothesis testing on the mean difference

i.e. H_0 : $\mu_d = \mu_f$

$diff = 0$

H_A : $diff \neq 0$



observed diff $= (\bar{x}_1 - \bar{x}_2)$

$\sigma_s/\sqrt{n} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$df = \min(n_1 - 1, n_2 - 1)$

Question

Thursday, 23 September 2021 8:30 AM

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

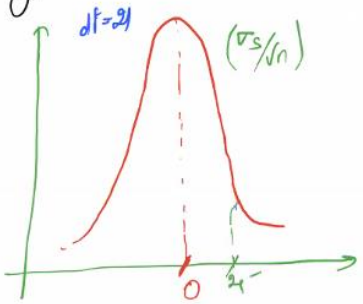
study design:
- randomized into two groups:
(1) play solitaire while eating - "win as many games as possible"
(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

$(\bar{x}_1 - \bar{x}_2) = (52.1 - 27.1)$

$= 25g$

Distribution for difference of group means if null hypothesis is true



$\left(\frac{\sigma_s}{\sqrt{n}}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}$

$= \sqrt{92.45 + 31.68}$

$= \sqrt{124.13} \approx 11.14$

t score $= \left(\frac{25 - 0}{\sigma_s/\sqrt{n}}\right)$

$= \frac{25}{(\sigma_s/\sqrt{n})}$

$= \frac{25}{11.14} = 2.24$

$df = \min(n_1 - 1, n_2 - 1)$

$= \min(22 - 1, 22 - 1)$

$= \min(21, 21) = 21$

$1\% < \text{area} < 2.5\%$

Reject null hypothesis

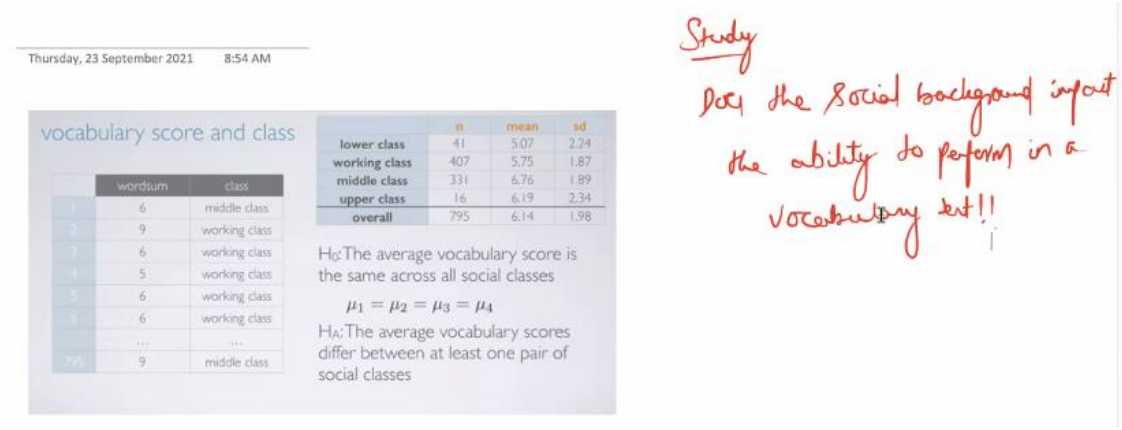
At a significance level of 5%, we believe there is convincing evidence that suggest different after lunch intake amongst people who lunch focused, and people who eat lunch distracted !!!

ANOVA

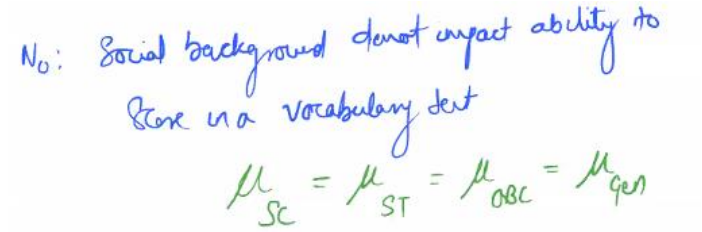
Analysis of Variance

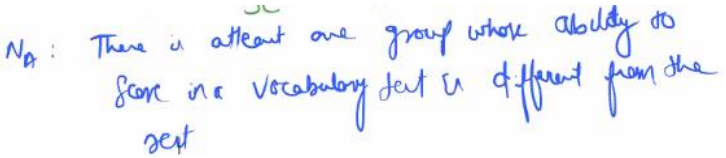
Hypothesis testing for more than 2 groups

Question



Solution





Formula for sum of squared errors

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

n_j : number of observations in group j
$\bar{y}_j$: mean of the response variable for group j
$\bar{y}$: grand mean of the response variable

$SSE_{(Social\ beach)} = \sum_{i=1}^{J} n_j (\bar{x}_j - \bar{x})^2$ $\quad J = no.\ of\ Socials\ beach$

$41 \times (5.67 - 6.14)^2$
$+ 407 \times (5.75 - 6.14)^2$
$+ 331 \times (6.76 - 6.14)^2$
$+ 16 \times (6.19 - 6.14)^2$

sday, 23 September 2021 9:12 AM

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

	df	SSE	SSE/df
Because of S.B. ✓	3	236.576	78.855
Due to all other	791	2869.80	3.628
Total ✓	794	3106.36 ✓	

$Ratio = \frac{78.855}{3.628} = 21$ ✓

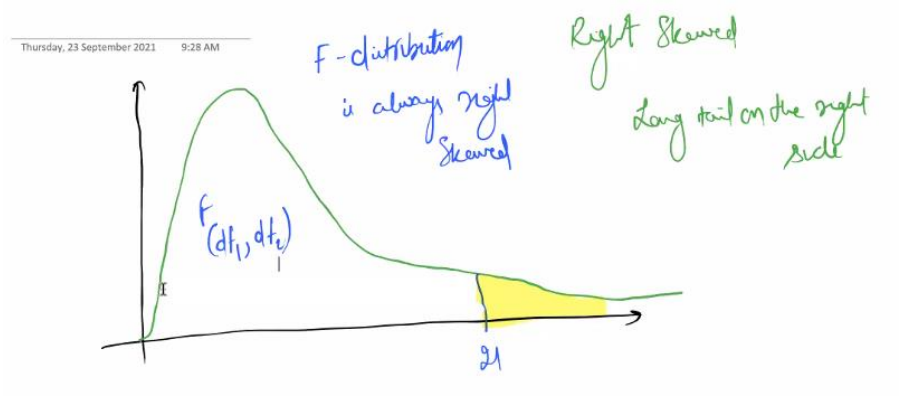
F-distribution
2 dof

F-dist ✓
(3, 791)

area under the curve
above a F score of 21.33

If this area < Sign we reject null hypothesis
else we accept null hypothesis

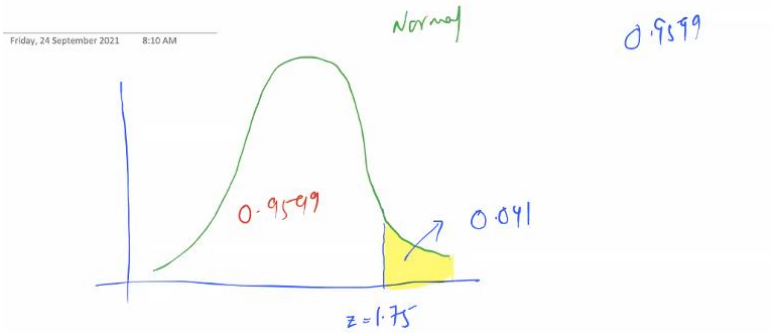
F Distribution is a right skewed distribution



We will use python to compute this value

c) P and Z values from Python

i. Z values

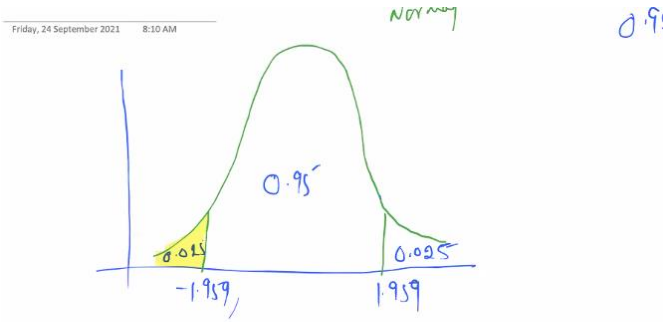



```
In [5]: 1 z_value=1.75
        2 x= st.norm.cdf(z_value)
        3 y= 1-x
        4 print(y)

0.040059156863817114
```

```
##### calculating probability below a certain z value in normal distribution ######

z_value = 1.75

st.norm.cdf(z_value) ### Area below particular z value

0.9599408431361829
```

ii. Z values for confidence interval

Friday, 24 September 2021 8:10 AM



$\bar{x} = 50, \quad \sigma = 10, \quad$ 95% Confidence

$\bar{x} - \left(\sigma/\sqrt{n}\right) \leq \mu \leq \bar{x} + z\left(\sigma/\sqrt{n}\right)$

```
[4] ###### calculating z value for a particular probability in normal distribution #######

    Area = 0.025

    st.norm.ppf(Area) ## z value below which there is a particular area

    -1.9599639845400545
```

iii. T- Distribution

T- Distribution

```
[7] #### Calculating the area below a given t value for a particular df #######

    degrees = 30

    t_value = 2

    st.t.cdf(t_value,df = degrees)

    0.9726874775185085
```

iv. T- Distribution where area is given

```
##### Calculating the t-value for a particular probability #####

df = 21

area = 0.975019

st.t.ppf(area, df = 21)

2.079992072119354
```

v. F Score- Anova

```
####### Calculating the area below an F-value in a F distribution #####

f_value = 21

1 - st.f.cdf(21,3,791)

4.2588155224621005e-13
```

vi. F-Score when area is given

```
area = 0.95

st.f.ppf(0.95,3,791)

2.616159809544512
```

vii. Function

```
#### confidence Interval #####

sample_mean = 30

sample_sd = 13 ## unbiased Sd

n_obs = 50

confidence_level = 0.95

## Can I find a confidence interval for population mean with 95% confidence level

def confidence_interval(sample_mean,sd,n_obs, confidence_level, type_sd):
  df = n_obs-1

  standard_error = sd/math.sqrt(n_obs) ### This gives sd/sqrt(n)

  if (type_sd == 'sample'):
    score = st.t.ppf(confidence_level + ((1-confidence_level)/2),df) ##### st.t.ppf(0.95 + (1-0.95)/2) ### st.t.ppf(0.95 + 0.25) ### st.t.ppf(0.975)

  if (type_sd == 'population'):
    score = st.norm.ppf(confidence_level + ((1-confidence_level)/2))

  confidence_interval = (sample_mean - score*standard_error, sample_mean + score*standard_error)

    return(confidence_interval)
```

Solution for a problem from above

```
confidence_interval(sample_mean=3.2,sd = 1.74, n_obs = 50,confidence_level = 0.95, type_sd = 'sample')

(2.705497472087071, 3.6945025279129293)
```

viii. Hypothesis testing function for a single mean

```
[19] ### Hypothesis Testing for single mean (one tail test) ####

     sample_mean = 30

     sample_sd = 13

     n_obs = 50

     population_mean = 40

     significance_level = 0.05

     def hypothesis_testing_1(sample_mean,sd,n_obs,population_mean,significance_level):

         df = n_obs-1

         standard_error = sd/math.sqrt(n_obs)

         score = abs((sample_mean - population_mean)/standard_error)

         area = 1 - st.t.cdf(score,df)
```

```
    if (area < (significance_level/2)):
      return ('The area above the calculated t value is {}, Hence we Reject null hypothesis'.format(area))
    else:
      return ('The area above the calculated t value is {}, Hence we Fail to reject null hypothesis'.format(area))
```

```
hypothesis_testing_1(sample_mean=52.1,sd = 45.1, n_obs = 22,population_mean=30,significance_level=0.05)
```

```
'The area above the calculated t value is 0.015954243787198 34, Hence we Reject null hypothesis'
```

```
[23] hypothesis_testing_1(sample_mean=118.2,sd = 6.5, n_obs = 36, population_mean=100,significance_level=0.01)
```

```
'The area above the calculated t value is 0.0, Hence we Reject null hypothesis'
```

Book: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow by O'Reilly

Resources

- Towards Data science
- Analytics Vidya
- Kaggle solutions- Machine learning
- Machine learning mastery