

Statistics

⇒ Population & Sample

- ⇒ A population is the entire group of individuals or items that we want to study. " N "
- ⇒ A sample is a subset of population that we want to study. " n "
- ⇒ The sample's needs to be identically distributed.
- ⇒ In machine learning every dataset you work with is a sample not population.

Sample mean " $\bar{x}$ "

- ⇒ The mean of a large sample is more likely to be representative of the population mean.

Sample Proportion " $\hat{p}$ "

Ex ⇒ Population size = 10

car = 1

bus = 2

bike = 4

motorcycle = 3

Q $\Rightarrow$ What is the proportion of people owning a bike?

Solⁿ $\Rightarrow \frac{4}{10} \times 100 \Rightarrow 40\%$

$$\frac{4}{10} = 0.4 = 40\%$$

The metric above is called Population proportion denoted by "p".

$$P = \frac{\text{no of items with given characteristics (x)}}{\text{Population (N)}}$$

& the above proportion is of the population but if we want to take the proportion of a sample it's given by " $\hat{p}$ ".

Sample Variance "s"

$\Rightarrow$ Variance is a measure of how spread out the data is & how far points are from the mean.

$$\text{Var}(x) = \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

~~the~~ Since we have a sample we don't know what "~~N~~" ^{Population} ~~sample~~ size is & if what population mean is? So let's come up with a different way to find out about the sample variance,

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

Here, $\bar{x}$ is mean ~~the~~, $n \Rightarrow$ sample size

in sample variance
$$s^2 = \frac{\sum (x - \bar{x})^2}{(n-1)}$$

we divide, it by $(n-1)$ & not (n) because

- ⇒ $(n-1)$ fixes bias when all you have is a sample
- ⇒ As n gets big, The difference matters less
- ⇒ If it matters, then you have too little data & you should gather more data
- ⇒ Some accepted Statistical techniques use " n " in the denominator to estimate variance, for example maximum likelihood estimation.

Law of large numbers

As the sample size increases, The average of the sample will tend to get closer to the average of the entire population

it is under certain Condition.

- ⇒ Sample is randomly drawn
- ⇒ Sample size must be sufficiently large
- ⇒ each observation in the sample size is independent of each other.

Central limit theorem.

In Discrete random variable,
As n becomes sufficiently large we will get a normal distribution with parameters

$$\mu = np$$
$$\sigma^2 = np(1-p)$$

In Continuous random variables,

⇒ Let's say there is a population

⇒ You take " n " samples from population, & take the mean of the n samples " $\bar{x}$ ".

⇒ Now you do this many times & record the mean each time. " y_n "

⇒ Now if you plot the means of the samples then, you will get the almost normal or gaussian distribution

⇒ & The mean of the distribution will be equal to the ~~gaussian mean~~ Population mean

$$\text{mean } y_n \leftarrow (\bar{y}_n) = (\mu) \rightarrow \text{Population mean}$$

⇒ & The variance of y_n will be equal to the variance of population divided by " n "

$$\Rightarrow \sigma_{y_n}^2 = \frac{\sigma^2}{n} \text{ or } \text{Var}(y_n) = \frac{\text{Var}(\text{Pop})}{n}$$

Point Estimation.

⇒ It is the method that we use to estimate the Population mean from the sample & other population parameters. like population proportion (p)

Maximum likelihood estimation : — a motivation.

⇒ A type of point estimation.

⇒ Imagine you have evidence of something & you want to find the scenario that may have led to that evidence. So you pick ~~at~~ out of all the possible scenarios, the one that created the evidence with the highest probability,

⇒ Imagine popcorn lying on the floor, ~~what~~ there are 3 events that you know ~~off~~ that created the evidence,

- (1) Some one was watching movies
- (2) Some one was playing games
- (3) Someone taking a nap.

The answer is (1),
So,

The maximum likelihood that popcorn was on the floor is given by (1).

So,

$P(\text{Popcorn on floor} | 1)$ is the highest or max likely.

Maximum Likelihood: Bernoulli Example

⇒ Suppose you toss a coin 10 times it lands in heads eight times in tails twice.

H, H, H, H, H, H, H, H, T, T

& Now you have 3 possible coins that you could have flipped to get these results.

Coin 1	Coin 2	Coin 3
$P(H) = 0.7$	$P(H) = 0.5$	$P(H) = 0.3$
$P(T) = 0.3$	$P(T) = 0.5$	$P(T) = 0.7$

Which one do you think was the coin used for these 10 tosses.

The answer is ⇒ coin 1, has the biggest probability.

$$P(\text{outcome}) \text{ by coin 1} = (0.7)^8 \times (0.3)^2$$

$$= 0.0051$$

$$P(\text{outcome}) \text{ by coin 2} = (0.5)^8 \times (0.5)^2$$

$$\Rightarrow 0.0010$$

$$P(\text{outcome}) \text{ by coin 3} \Rightarrow (0.3)^8 \times (0.7)^2$$

$$\Rightarrow 0.0003$$

So, can we do better,

is there a better coin that maximise the probability

So, $p = P(H)$

for 10 tosses & getting 8 Heads prob ⇒

$$p^8 (1-p)^2$$

You want p that maximizes the chances of seeing 8H,

that is called likelihood

Likelihood

$$L(P; 8H) = P^8 (1-P)^2$$

↳ The probability of seeing this data based on the current model which is a coin flip experiment where you get with probability P .

This is a function of P , so what you gotta do is maximize it,

so, to ^{avoid} dealing with the multiplication of many things we take log.

so,

$$\log(P^8 (1-P)^2) = 8 \log(P) + 2 \log(1-P)$$

& the above eqn is called log likelihood.

$$\text{log-likelihood, } l(P; 8H) = \log(P^8 (1-P)^2) = 8 \log(P) + 2 \log(1-P)$$

so, now we gotta maximize log likelihood.

$$\text{so, } \frac{d}{dP} (8 \log(P) + 2 \log(1-P)) = \frac{8}{P} + \frac{2}{1-P} (-1) = 0$$

Putting equal to zero,

$$\hat{P} = \frac{8}{10}$$

, so $P(H) = (\hat{P})$, should be $\left(\frac{8}{10}\right)$.

i.i.d = independent & identically distributed

for general case:- n_i i.i.d Bernoulli (p)

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

The value of $\hat{p}$ is precisely equal to the mean of the population,

So in other words, if you have k heads among the population of coins.

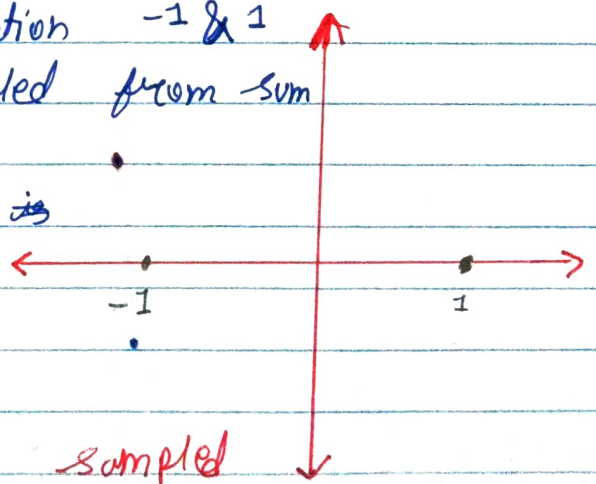
Then the optimal probability to obtain these k -heads would be $\left(\frac{k}{n}\right)$.

Nahh!! This shit was too 2EZ i don't know why i was stressing.

Maximum Likelihood : Gaussian Example

Let's say you have some observation -1 & 1 & these observations were sampled from some distribution.

& the question is what distribution is they could have been sampled from.



⇒ The distribution they were sampled from or the best distribution these samples belong to is the distribution where mean of the distribution is the mean of sample & same goes for variance.

total intial provided:

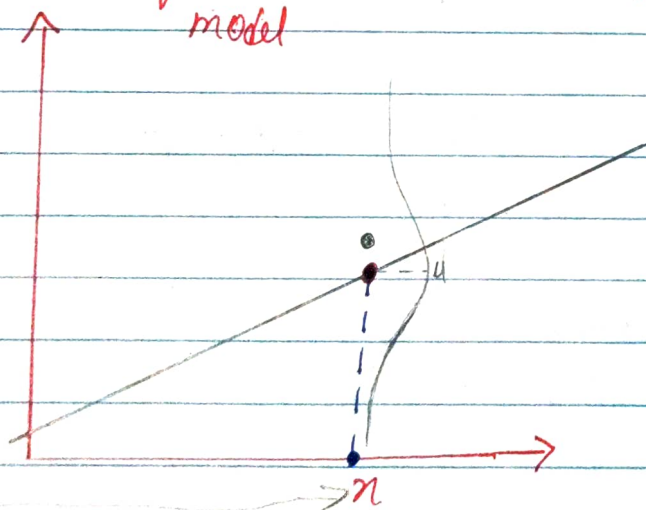
MLE: Linear Regression

⇒ A way that maximum likelihood is used in machine learning is the following.

→ we have a bunch of data & some possible models that could have generated that data. & for each one of the models, you calculate the probability that the data appears based on models & the one that gives the highest prob is the winner.

it's the same as finding best model that fits the data. & that's what we do in linear regression bigger.

How exactly Does a Line produce point's.

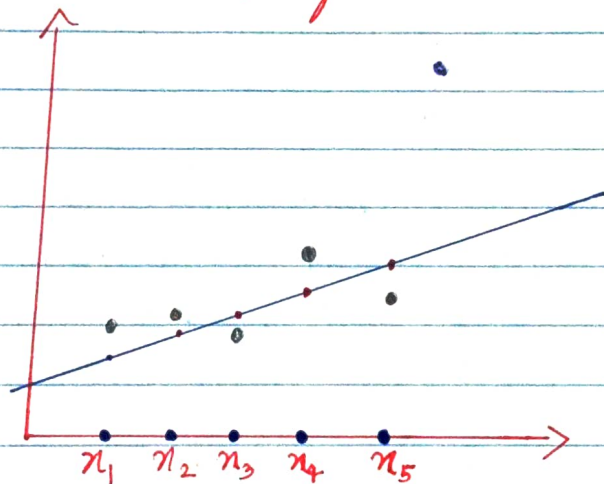


⇒ Let's say that this is your model & this is a point x , & that joins the point on the line above, so you're going to generate a point close to the line using a gaussian & we center it at the point.

where the line meets the horizontal line.

& we are gonna sample out of that gaussian, so we sample a point (on the graph given by pencil)

Now if we have a bunch of points x_1 to x_5 & then you will get 5 gaussians all centered at the point on the blue line. & from each one of these we're going to sample some point. & These are the five points we are sampling based on this blue line.

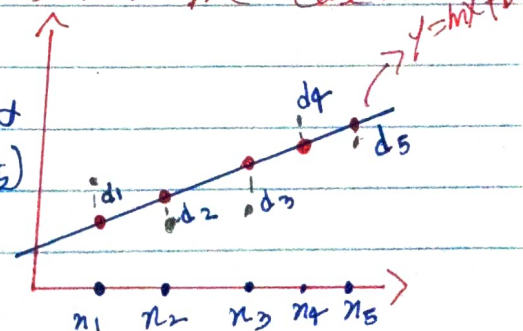


Now,

we have got the points & all we have to do is find line that best produced the points in another words "Line that best fits the data" is that is linear regression.

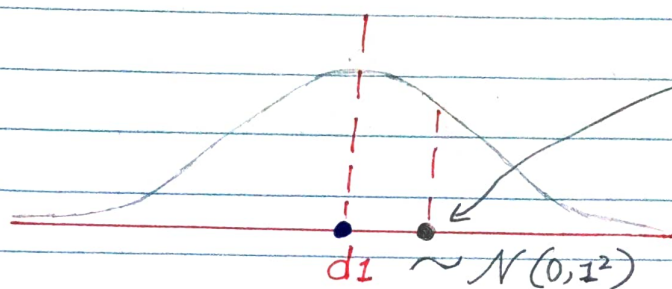
Let's do some maths, to find out why is this the case.

⇒ The distances between the generated point & point on line are called $(d_1, d_2, d_3, d_4, d_5)$ respectively.



Let's calculate the likelihood of generating point given by the point x_1 .

Since we have a gaussian, then what really happens is that the gaussian is centered at the point x_1 intersects that blue line, & were generating the other point.



Now, if we assume that it's a gaussian with standard deviation one & mean equal to 0, then the likelihood of generating these points can be given by the formula of the gaussian

$$\Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_1^2}$$

Now, for ^{the likelihood of} generating all the 5 "independent" points are

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} d_5^2}$$

& this is what we have to maximize above.

you can just remove these constants because we don't need them, & how

we are left with, $e^{-1/2 \cdot d_1^2} \cdot e^{-1/2 \cdot d_2^2} \cdot e^{-1/2 \cdot d_3^2} \cdot e^{-1/2 \cdot d_4^2} \cdot e^{-1/2 \cdot d_5^2}$

$$\Rightarrow e^{-1/2 (d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)}$$

Now taking log,

we get

$$-\frac{1}{2} (d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)$$

Now, we have a -ve no so insted of maximizing we will have to minimize.

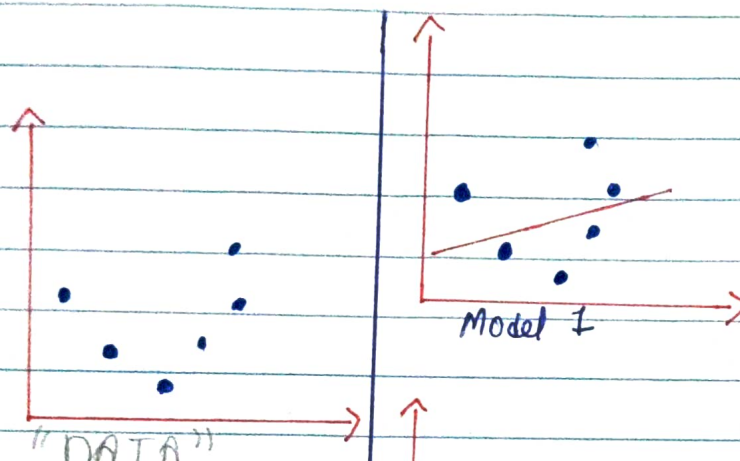
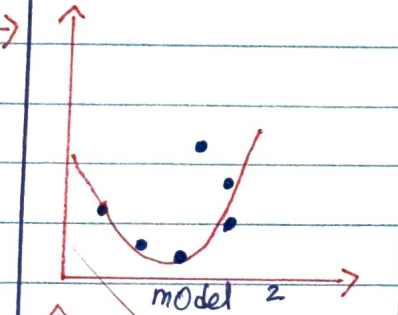
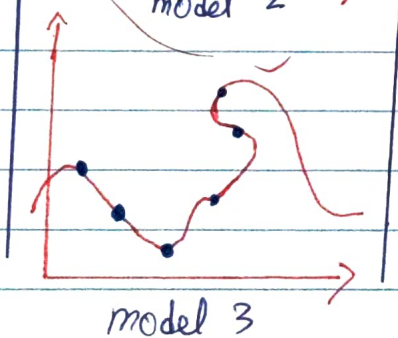
$$\Rightarrow (d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2) \quad \left\{ \begin{array}{l} 1/2 \text{ doesn't} \\ \text{matter cause it's} \\ \text{a constant} \end{array} \right\}$$

& what you are left with above is least square error & that is what linear regression is.

In ~~go~~ linear regression you wanna find the blue line that minimizes the sum of squared distances from the point.

Regularization-:

It prevents ~~trainings~~ model's from overfitting the training data.

	LOSS	Equation	Penalty	New Loss
	10	$y = 4x + 3$	$L_2 = 4^2$ $\Rightarrow 16$	26
	2	$y = 2x^2 - 9x + 5$	$L_2 = 2^2 + (-9)^2$ $\Rightarrow 20$	22
	0.1	$y = 4x^{10} - 9x^8 - 2x^6 + 3x^5 - 6x^4 - 10x + 4$	$L_2 = 4^2 + (-9)^2 + (-2)^2 + (3)^2 + (-6)^2 + (-10)^2$ $\Rightarrow 246$	246.1

$\Rightarrow$ Let's say you have a data & you want to find out which model best fits the data, now above you are given 3 models, model 1 loosely or doesn't fit the data, model 2 generalises the data & fits it & model 3 fits the data perfectly, But

we know that model 3 overfits the data & we don't want that, notice the equation for model 3 is very complex & is raised to 10 has the power of 10 as a polynomial.

We just want the model to generalize the data not to perfectly fit it & hence we introduce "penalty".

We penalize the ~~data~~ ^{model} that overfits the data,

we calculate the penalty for every model by "L2 regularization", here, it essentially tells us that ~~what~~ what is the complexity of the model the higher "L2" value the more complex a model becomes.

in L2 regularization, we calculate the sum of squared of every coefficient of x ,

$$\text{ex} \Rightarrow y = 4x^2 + 3x + 4$$

$$\text{Loss} = (4)^2 + (3)^2 \Rightarrow \boxed{25}$$

In general,
Now, for new ^{loss} ~~penalty~~, we just add the value of previous loss & Penalty.

In general,

"Regularized" or New model $\Rightarrow$ $\text{Loss} = \text{Original Loss} + \lambda \sum_{i=1}^n w_i^2$

$$y = w_n x^n + w_{n-1} x^{n-1} + \dots + w_1 x + a_0$$

λ is a regularization parameter, how much we want to simplify our model, if " $\lambda=0$ ", then model may overfit, " λ " too high. model becomes too simple & underfits

to add more information

L1 regularization $\rightarrow$ It makes the large weights zero

L2 " $\rightarrow$ It makes the large weights small.

L2 regularization is commonly used in standard machine learning tasks like,

linear regression, logistic regression, Neural networks.

There is also a combo called Elastic Net, which mixes both L1 & L2 regularization,

used in real world problems with lot's of features. It is commonly

Maximum likelihood estimation :- a motivation part 2.

⇒ Going Back to the previous example of pop-corn lying on the floor, but let's say the ~~even~~ Hypothesis that could have caused the evidence of pop-corn lying on the floor are different.

(H₁) ⇒ Someone ~~was~~ watching a movie

(H₂) ⇒ There was a popcorn throwing contest.

So, the probability, that There can be pop corn on the floor # After a popcorn throwing contest is very very high than the probability, that there can be Popcorn on the floor after a movie.

$$P(\text{Popcorn} | \text{Movies}) < P(\text{Popcorn} | \text{contest}),$$

However the probability of watching a movie is higher than that of popcorn throwing contest.

& we have to take that in account, & taking that into account we get

$$P(\text{Popcorn} | \text{movies}) P(\text{Movies}) > P(\text{Popcorn} | \text{contest}) P(\text{contest})$$

which can be also written as

$$P(\text{Popcorn} \cap \text{movies}) > P(\text{Popcorn} \cap \text{contest})$$

Frequentist vs. Bayesians.

⇒ **Frequentist** $\therefore$ Believes probability is the long-run frequency of events. They don't assign probabilities to beliefs or parameters.

⇒ **Bayesian** $\therefore$ Believes Probability is a degree of belief based on prior knowledge, & updates it with new data.

⇒ Imagine 2 guys are sitting with a coin now to find the probability of Heads. They have to flip the coin 10 times & coin lands on heads 8 times & coin lands on tails 2 times, & here are their perspectives

Frequentist

Bayesian

~~The Bay's~~

He says the coin landed on Heads 8 out of 10 times meaning the probability of heads is $\frac{8}{10}$ or 0.8.

He says I know the Probability of heads is 50%, but the coin landed 8 out of 10 times on head, so I will update my prior belief by increasing my Probability for heads to let's say 60%.

Frequentist

- ⇒ No prior belief
"concept of likelihood"
- ⇒ focuses on data alone & repeated trials

⇒ Goal :- Find the model that most likely generated the observed data.

ex ⇒ MLE

Bayesian

- ⇒ Starts with a prior belief.
- ⇒ Updates it's Bayesian theorem as new data comes in

⇒ Goal: Update prior belief based on observations.

ex ⇒ MAP

Bayesian Statistics - Maximum a Posterior ("MAP")

- ⇒ It's a method of updating your Prior beliefs with the new given evidence & getting a Posterior
- ⇒ MAP gives a compromise between :-

- ⇒ what you believed before (Prior), &
- ⇒ what the new data says (Likelihood)

⇒ MAP estimation gives the most likely value of a parameter after updating our beliefs with data

⇒ Any time you use MAP estimation with an uninformative prior that weights all possible values equally, the result is identical to what you would get from a frequentist approach.

Bayesian Statistics : Updating Priors

It is just bayes theorem by heart.

let's see the example of determining whether you have a fair or biased coin & let's see how the expression

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \text{ is used}$$

let's say, you know there are 2 possible types of coins, the first one is fair & the other is biased.

$P(H) = 0.5$ "fair" $P(H) = 0.8$ "biased", you will be given a mystery coin, it can be either fair or biased, but you don't know which one it is.

The only way to figure this out is to flip it.

Event to predict

$A \rightarrow Y$ takes some value

Y : odds of H for your coin

$$Y = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.8 & \text{if coin is biased} \end{cases}$$

let's say you don't know the probability of coin being fair or biased so, you can assume anything for example, for probability ($\gamma = 0.5$)

it's "0.75" & for Probability ($y=0.80$) it's "0.25"

$$\begin{array}{l} P(Y=0.5) \Rightarrow 0.75 \\ P(Y=0.80) \Rightarrow 0.25 \end{array} \rightarrow \text{Priors}$$

Evidence

$B \rightarrow X$ takes some value
 X : result of coin flip.

$$X = \begin{cases} 0 & \text{if Tails} \\ 1 & \text{if Heads} \end{cases}$$

we are replacing event B with the outcome of a new random variable x .
 x will be the result of flipping your mystery coin.

Now, we are ready to update our priors based on the outcome of flipping your coin.

& flipping the coin you get result as "Heads"

So $x=1$

Now we are going to use Bayes ^{→ Prior} theorem.

$$\underbrace{P(Y=0.5|X=1)}_{\text{Posterior}} = \frac{P(X=1|Y=0.5) \cdot \overbrace{P(Y=0.5)}^{\text{Prior}}}{P(X=1|Y=0.5) \cdot P(Y=0.5) + P(X=1|Y=0.8) \cdot \overbrace{P(Y=0.8)}^{\text{Prior}}}$$

$$P(X=1|Y=0.5) = 0.5 \Rightarrow \text{A fair coin will land on heads half of the time}$$

$$P(Y=0.5) \Rightarrow \text{This comes from your priors, you assumed these values} = 0.75$$

$$P(Y=0.8) \Rightarrow \text{You assumed it} = 0.25$$

$$P(X=1|Y=0.8) \Rightarrow \text{it is given that } P(\text{heads}) \text{ with biased coin is} = 0.80$$

So, putting up all the values in the eqn you get,

$$\Rightarrow \frac{0.5 \times 0.75}{0.5 \times 0.75 + 0.8 \times 0.25}$$

$$P(Y=0.5|X=1) \Rightarrow 0.652 \rightarrow \text{Posterior}$$

Compare this value to your earlier belief & you can see that previously (prior) you had the chance of fair coin as "0.75" but now you have the chance of fair coin (Posterior) as "0.625",

This is how you update beliefs using Bayes theorem.

new, Probabilities

$$\begin{aligned} P(Y=0.5|X=1) &= 0.652 \\ P(Y=0.8|X=1) &= 0.348 \end{aligned} \rightarrow \text{Posterior}$$

The above bayesian equation can be written in Probability Mass function "PMF" as.

PE

$$P_{Y|X=1}(0.5) = \frac{P_{X+Y=0} P_{X|Y=0.5}(1) \cdot P_Y(0.5)}{P_X(1)}$$

→ PMF of Y given that X=1 evaluated in (0.5).

$\ominus \rightarrow$ Capital theta
 $\theta \rightarrow$ Small theta

$P_{X|Y=0.5}(1) \Rightarrow$ This is the conditional PMF of X given that $Y=0.5$, evaluated at (1)

$P_Y(0.5) \Rightarrow$ The prior is simply the PMF of Y evaluated at (0.5).

$P_X(1) \Rightarrow$ It's also just the probability that $X=1$, which is PMF of " X " evaluated over "1"

So, the general expression for bayes theorem when you have 2 discrete random variable,

$$P_{Y|X=x}(y) = \frac{P_{X|Y=y}(x) \cdot P_Y(y)}{P_X(x)}$$

for, 2 Continuous random variable, you just replace probability mass fn with prob density function.

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x) \cdot f_Y(y)}{f_X(x)}$$

for 1 variable being discrete & one variable being Continuous,

let's say " X " is Continuous, " Y " is Discrete.

$$P_{Y|X=x}(y) = \frac{f_{X|Y=y}(x) P_Y(y)}{f_X(x)}$$

$\Rightarrow$ In an machine learning models, you are solving for the probability that a parameter in a model takes a certain value, in these instances you can write replace
 $Y \rightarrow \ominus, y \rightarrow \theta$

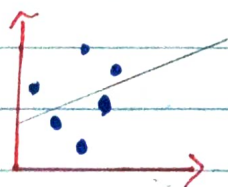
Bayesian Statistics: Final summary.

- ⇒ Bayesians update prior beliefs
- ⇒ MAP with uninformative priors is just MLE.
- ⇒ With enough data, MLE & MAP estimators estimates usually converge.
- ⇒ Good for instances when you have limited data or strong prior beliefs.
- ⇒ Choesed often by statisticians
- ⇒ Down side is when priors are wrong, wrong conclusions will be generated

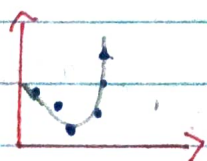
Relationship between MAP, MLE & Regularization.

Let's say you have the data, & you have 3 models that could have fit that data, & each one of them generates data with some probability, a model can generate the data as we saw before by simply creating a gaussian at every point on the line & generating points close to the line or to the curve

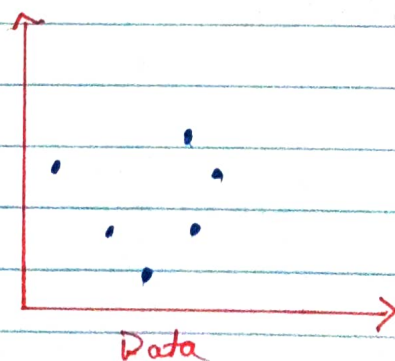
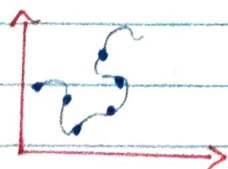
model ("1")



model ("2")



model ("3")



The first one generates the data with some probability,
 The second one generates the data with higher probability,
 The third one generates the data with the highest probability,

But just like before, with popcorn throwing contest,
 we have to take account for probability of the model.
 The lesser the complexity, the higher the probability.

and after taking in account these factor's the winner
 is the 2nd one.

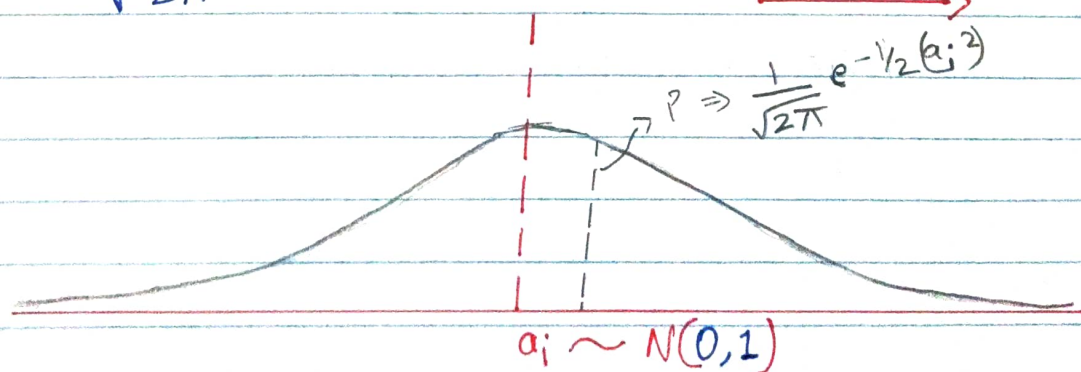
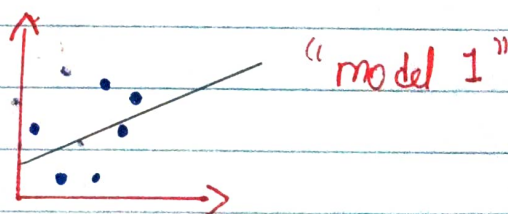
So, step ① what we did was find the "MLE"
 in this case $P(\text{Data}|\text{Model})$ & it's clear that
 model 3 was best

step ② Here we did the "MAP" & we took into
 account for $P(\text{model})$, & then we did it with
 Bayesian approach " $P(\text{Data}|\text{Model}) \cdot P(\text{model})$ ". This helped
 us to pick the model that is less complex &
 doesn't overfit the data

So, what is $P(\text{model})$?

model 1 $= a_1x + b$

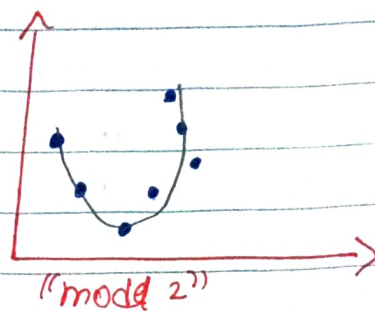
$$P(\text{model 1}) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}a_1^2}$$



$$\boxed{\text{model 2}} = a_1 x^2 + a_2 x + b$$

$$P(\text{model 2}) \Rightarrow$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_1)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_2)^2}$$



$$\boxed{\text{model 3}} \Rightarrow a_1 x^{10} + a_2 x^9 + \dots + a_{10} x + b \Rightarrow \boxed{\prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_i)^2}} \Rightarrow P(\text{model 3})$$

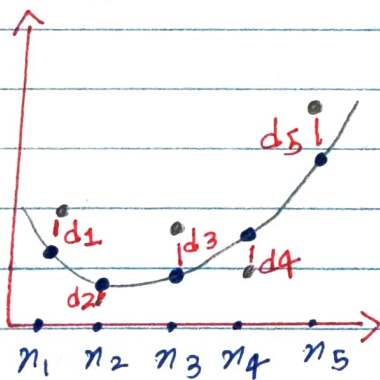
$\hookrightarrow P(\text{model 3})$

Now, according to step the best model will be given by,
(we using model 2 for example)

$$P(\text{Data}|\text{Model})$$

$$P(\text{model})$$

$$\Rightarrow \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_5^2} \right] \cdot \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_1)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_2)^2} \right]$$



$d \Rightarrow$ distance

$$\boxed{a_1 x^2 + a_2 x + b} \quad \text{model 2}$$

Now, removing all the constants & taking log we get,

$$\Rightarrow \left[-\frac{1}{2} (d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2) \right] + \left[-\frac{1}{2} (a_1^2 + a_2^2) \right]$$

Now, you have to maximize the eqn above,

But $\Rightarrow$ multiplying above eqn by $-\frac{1}{2}$

$$\Rightarrow \left[d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 \right] + \left[a_1^2 + a_2^2 \right]$$

&, now since the ~~minus~~ minus symbol is gone you need to minimize this eqn.

above $[d_1^2 + \dots + d_5^2] \Rightarrow$ is the squared loss
 $\rightarrow P(\text{Data}|\text{model})$

&, $[a_1^2 + a_2^2] \Rightarrow$ is regularization term,
 $\rightarrow P(\text{model})$

Maximizing the probability of a model is the same as minimizing the sum of squares of the coefficients.
& ,

Maximizing the conditional probability of the data given the model is the same as minimizing the Square loss.

Suppose, You have coefficients $a_1, a_2, \dots, a_n$
& Your data fit error (distances) is $d_1, d_2, \dots, d_n$

Then objective becomes,

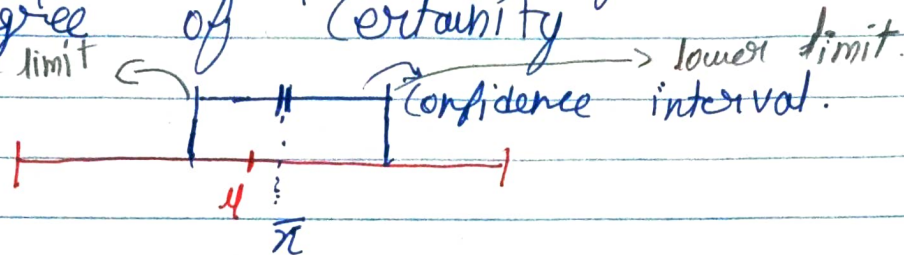
$$\text{Loss} = \sum_{i=1}^n (d_i)^2 + \lambda \sum_{i=1}^n (a_i)^2$$

"Data fit error" "Regularization penalty"

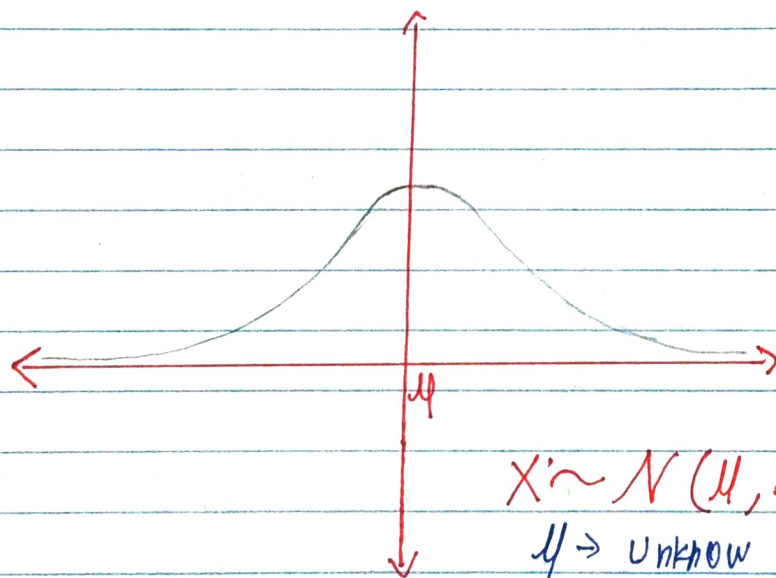
$\lambda =$ regularization parameter.

Confidence Interval

Insto A confidence interval is an interval of values that is a lower & upper limit which contains the probability population parameter for ex $\Rightarrow \mu$ with some degree of certainty



$$\text{Confidence interval} = \bar{\pi} \pm \text{margin error}$$



$$X \sim N(\mu, \sigma^2)$$

$\mu \rightarrow \text{unknown}$, $\sigma^2 \rightarrow \text{known}$

Generating a Confidence interval to estimate where μ is located, To do that you will take a random sample from population.

Sample $\Rightarrow n=1$, so mean $\bar{\pi} = 1$

let's create a random variable $\bar{X}$, to describe the probability of selecting different sample means. it's actually going to be identical to x ,

$$\bar{X} \sim N(\mu, \sigma^2)$$

you know suddenly know the true value of μ , but you do know x & $\bar{X}$ have the same mean.

when you take a large number of means they gather around each other pretty closely signifying the true mean is there somewhere.

it will be in the range of those points that gather around closely. & if your sample is typical

So, how far are most sample means from μ ?

To do this you need two related concepts. You will need a margin of error which is distance on either side of μ & confidence level which is the probability that your sample mean is within that margin of error.

To set these 2 values you start with " α " significance level, it is the probability that your sample mean falls outside of your margin of error.

let's set $\alpha = 0.05$, it's a very common choice,

now confidence level = $1 - \alpha$, This is the probability that a randomly generated sample mean is

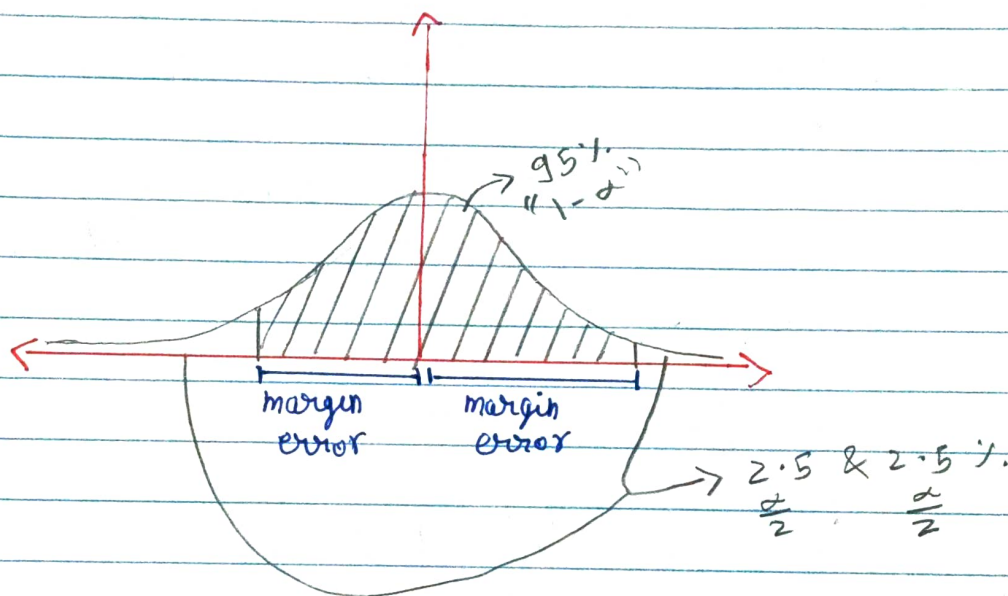
within the margin of error,

So, Confidence level $\Rightarrow 1 - \alpha$

$$\Rightarrow 1 - 0.05$$

$$\Rightarrow 0.95$$

Now you can increase the size of your margin of error's until 95% of all randomly generated sample mean ($\bar{x}$) will fall within the shaded portion of the distribution curve



& you can also say that 2.5% of the time your sample mean will lie outside the margin because it's too small & 2.5% because it's too big.

So, confidence interval = $\bar{x} \pm \text{margin error}$

Summary :-

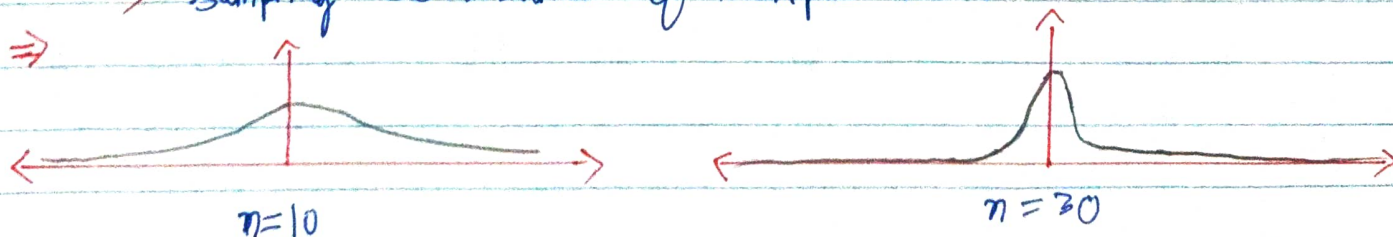
- Confidence intervals are a sample mean with a margin of error added to each side.
- **Confidence level** :- The probability of a confidence interval contains "u" (for example 95%).
- Ideally you have both high confidence & a narrow interval.
- Larger samples (more data) will give a narrower interval.
- Decreasing confidence level will also shrink the interval.

Margin of error

⇒ Margin of error is determined by the sample size & confidence level you are aiming for

⇒ As you increase the value of n the sample distribution of sampling mean gets thinner & hence the margin of error also decreases.

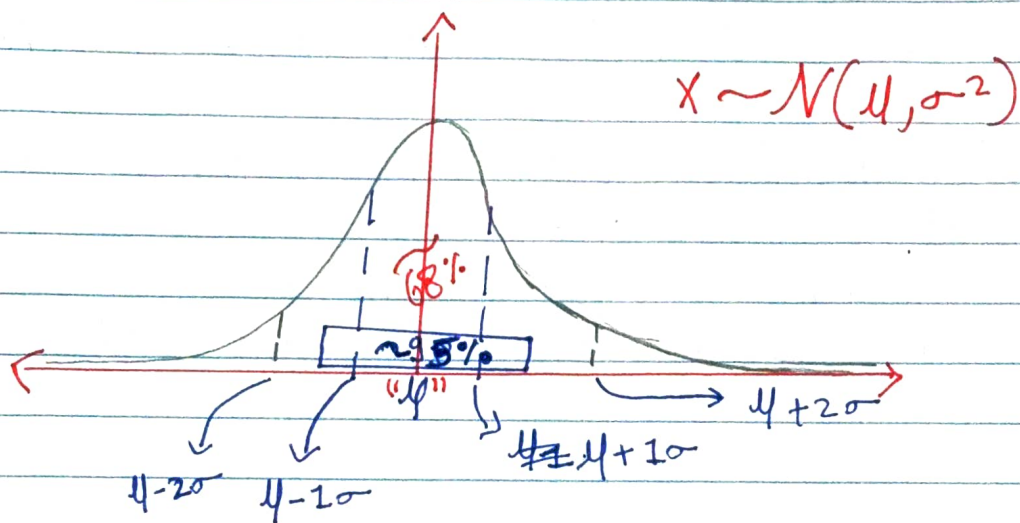
→ sampling distribution of sample mean.



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

So, all you have is $\bar{x}$ & let's assume that you also know (σ) ,

Now, step (1) - learning about z-scores,



$\Rightarrow$ for any normal curve roughly " $\sim$ " 68% of the curve lies within 1 standard deviation of the mean, & 95% of the curve lies within 2 standard deviation of the mean.

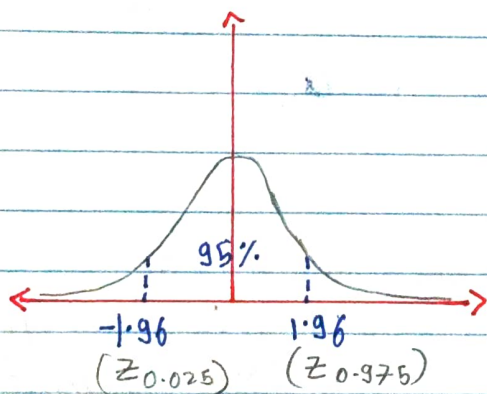
$\Rightarrow$ These points like 1, 2 in front of σ or coefficient of σ are called "z-values" or "z-scores" or "z-statistics". The point 2 standard deviation greater than the mean has a "z-score" of "2", & the point "one" standard deviation below the mean has the z-score of "-1".

To find the z -scores you first have to standardize the distribution.

$$\frac{X - \mu}{\sigma} = z \sim N(0, 1)$$

Now the distribution becomes z distribution, The z -distribution has the mean of 0 & variance 1.

if you want exactly 95% of distribution, the the z -scores to 2 decimal places are, $(-1.96 \text{ \& } 1.96)$ not $(-2 \text{ \& } 2)$.



$\Rightarrow$ These -1.96 & $+1.96$ are called critical values, they are cut-off points inside of which an exact Percentage (95%) of a probability distribution is contained, To find these values you'd either look them up in a lookup table or use a software library.

$\Rightarrow$ The first critical value is $(z_{0.025})$, this notation means, find me the z -score of the point which has 2.5% of the distribution curve to its left. The second critical value is $(z_{0.975})$.

it says to find the z-score of a point which has 97.5% curve to its left.

So you want your critical values to exclude 5% values. This is just significance level " α "

$$Z_{\frac{\alpha}{2}} = Z_{(0.025)} \quad \& \quad Z_{1-\frac{\alpha}{2}} = Z_{0.975}$$

"for example let's say you have $\alpha = 0.10$ "
then

$$Z_{\frac{\alpha}{2}} = Z_{(0.05)} \quad , \quad Z_{1-\frac{\alpha}{2}} = Z_{(0.950)}$$

are your critical values

Now,

if you have a non-standardized, since the distribution isn't normalised, you need to multiply it by standard deviation.

$\Rightarrow$ for normal distribution standard deviation is 1 so if you multiply it by 1 or not it doesn't matter, but when you don't have normal distribution you have to multiply by standard deviation.

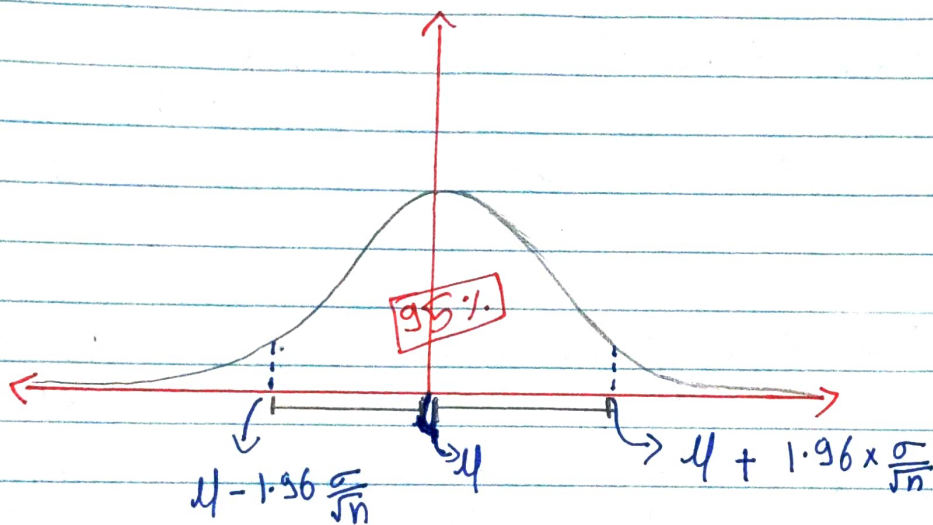
$$\text{so } \Rightarrow \boxed{-1.96 \times \sigma \quad \& \quad 1.96 \times \sigma}$$

Step (2)

You know that, thanks to Central Limit theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \cdot, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad \left. \vphantom{\frac{\sigma}{\sqrt{n}}} \right\} \begin{array}{l} \text{also said} \\ \text{as "standard} \\ \text{error"} \end{array}$$

$\bar{X} \Rightarrow$ sampling distribution of sample mean.



margin error $\therefore$ for 95% Confidence interval $\Rightarrow$

$$\boxed{1.96 \cdot \frac{\sigma}{\sqrt{n}}}$$

Margin of error in general,

$$\boxed{Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}$$

Now, that you have a confidence interval margin error, you are set to find confidence interval,

Your goal was to find lower & upper limit that bound " μ ", the population mean.

$$\Rightarrow \text{lower limit} < \mu < \text{Upper limit}$$

now, you have calculated your sample mean will fall within the range,

$$\mu - 1.96 \times \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Subtract μ & $\bar{x}$ from all the term you will get,

$$-\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

multiplying by (-1) & also flipping the inequality because multiplying by (-1) we get

$$\bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Now, Confidence interval, is

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Summary for confidence interval, Calculation steps

- Steps :-
- Find the sample mean
 - Define a Desired confidence interval level $(1-\alpha)$.
 - Get the critical value $(z_{1-\frac{\alpha}{2}})$
 - Find the standard error $(\frac{\sigma}{\sqrt{n}})$
 - find the margin of error
 - add/subtract margin of error to the sample mean

$$\text{Confidence interval} = \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

- Assumptions :-
- Sample used should be random
 - Sample size > 30 or population is approximately normal.

Calculating sample size

$\Rightarrow$ let's say you are given margin error & you are told find the ~~min sample~~ sample size required to achieve this margin.

So, margin error = $z_{1-\frac{\alpha}{2}} = 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

given

$$\Rightarrow z_{1-\frac{\alpha}{2}} = 1.96, \sigma = 25$$

& margin error $\Rightarrow 3$

So, putting it in eqn we get, $\Rightarrow 1.96 \times \frac{25}{\sqrt{n}} = 3, n \approx 267$

In general we use the formula,

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{MOE} \right)^2$$

MOE $\Rightarrow$ Margin of Error.

Difference between Confidence & ~~transcript~~ Probability

So, $\Rightarrow$ The population mean μ is fixed, not random
 $\Rightarrow$ The sample mean $\bar{x}$ is random - it changes with each example.
 $\Rightarrow$ Therefore, the confidence interval is also random - it changes each time based on it.
So the randomness lives in the process, not in the population mean.

$\Rightarrow$ The correct statement is

"We are 95% confident that the interval from 170 to 174 cm contains the true mean"

In correct statement is:-

"There's 95% probability that the true mean is between 170 & 174 cm."

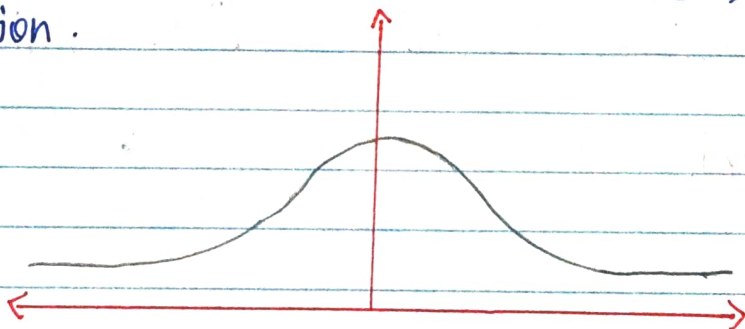
$\Rightarrow$ because when we calculate the sample mean ($\bar{x}$) the randomness is gone, now the population mean is in the interval or not.

Confidence interval when " σ " is unknown

⇒ When sigma " σ " is unknown, in the we use standard deviation of the sample denoted by " s "

⇒ Note that, the " s " is the standard deviation of the one sample you took, not the ~~sample~~ standard deviation of The sampling distribution of sample mean.

⇒ & if you are taking " s " as the standard deviation then you don't use Z-score, you use a t-score & this t-score doesn't follow the normal distribution, it has thicker tails, it follows t-distribution.



⇒ Now there are several t-distributions based on the number of sample size, which is called degree of freedom,

$$\text{degree of freedom} = n - 1$$

$n \rightarrow$ sample size

as, the degree of freedom increases, the t-distribution approaches normal distribution.

finally the formula is

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

$t \rightarrow$ t-distribution

$s \rightarrow$ sample mean

$n \rightarrow$ sample size.

Confidence intervals for proportions (P).

Let's say there are "30" people in our sample size ~~is~~ that we took from Statistopia.

~~30~~ 24 of them use bicycle &
6 of them use cars

Proportion for bicycle is $= \frac{x}{n} = \frac{24}{30} = 0.8$ or 80%.

Now the problem is how do we calculate 95% Confidence interval for this sample proportion?

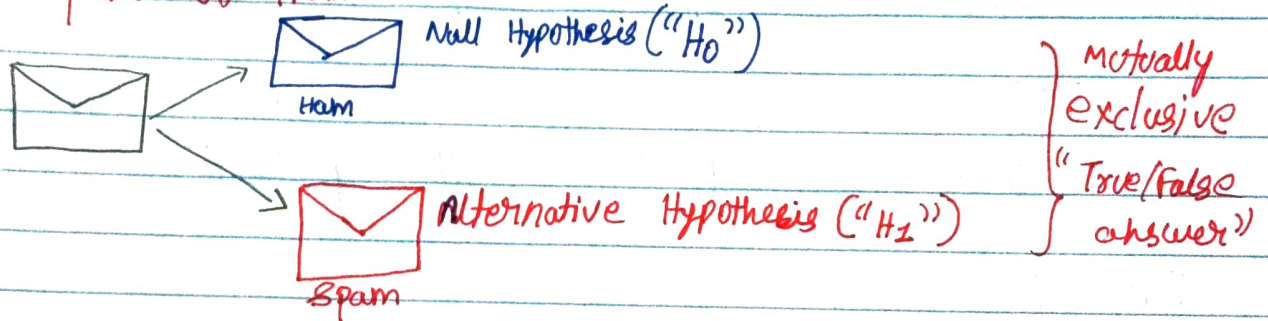
So, confidence interval $= \hat{p} \pm \text{margin of error}$

$$\text{margin of error} = Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hypothesis Testing

example :-

let's say you have an email spam detector & what it does is it determines if a given email is spam or Ham.



& By default we are gonna assume ~~that~~ that email's are ham, & the reason for this is that it's much worse to delete a good email than to accidentally put a spam email on your inbox.

& that our null hypothesis (H_0) we assume safely that nothing is happening.

& then we have an alternative ~~hyp~~ hypothesis "~~#1~~" " H_1 " & that's the one we are ~~try~~ trying to identify.

If there is strong evidence that the email is spam, we reject the null hypothesis & accept the alternative.

However, if there isn't enough evidence to ~~prove~~ the email is spam, we fail to reject the null hypothesis. This doesn't

mean we have proven the email is ham, only that we don't have sufficient evidence to conclude it is spam.

TYPE I & TYPE II ERROR

TYPE I Error $\therefore$ also called (false positive), it happens when you reject " H_0 " null Hypothesis when it was true.

TYPE II Error $\therefore$ also called (false -ve), it happens when you don't reject " H_0 " null Hypothesis.

TYPE I & TYPE II error's Don't have same impact on the problem

for example, if you send a ham email to spam it's a very big problem, but if you send a spam email into ham category it's not that big of a problem

$\Rightarrow$ So, what is the greatest probability of type I error you are willing to tolerate?

$\Rightarrow$ well the maximum probability of type one error is called Significance level " α ". it's a probability so it will only take values btw 0 & 1.

$\Rightarrow$ So, let's say significance level $\boxed{\alpha = 0.05}$. This means on average you will decide ham email is spam 5% of the time.

Right tailed, left tailed & 2 tailed tests

⇒ what is test statistic?

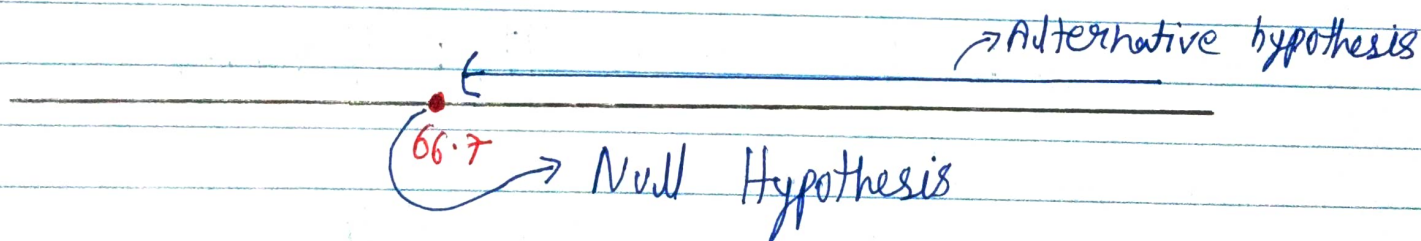
⇒ It is the information about the population parameter under study.

⇒ for example " μ " mean.

Right-tailed test or Right tailed hypothesis -:

Ex ⇒ [The mean height for 18yo in the 70s was 66.7 in.]
"It's a right tailed hypothesis when the alternate hypothesis is lying right side of the null hypothesis"

$$\text{So, } H_0: \mu = 66.7 \quad \text{vs} \quad H_1: \mu > 66.7$$



Type I error :- Determine $\mu > 66.7$, when population mean did not change,

Type II error :- Determine μ Do not reject $\mu = 66.7$, when in true $\mu > 66.7$.

Left tailed test :-

"It's a Left tailed hypothesis when the alternate hypothesis is lying left side of the null hypothesis."

$$\text{ex} \Rightarrow H_0: \mu = 66.7 \text{ vs } H_1: \mu < 66.7$$

Type I error: Determine $\mu < 66.7$, when population mean didn't change.

Type II error: Don't reject that $\mu = 66.7$ when true $\mu < 66.7$.

2 tailed test / Null Hypothesis :-

"It's a Null Hypothesis when the sample deviates a lot"

$$\text{ex} \Rightarrow H_0: \mu = 66.7 \text{ vs } H_1: \mu \neq 66.7$$

$$\left. \begin{array}{l} \text{if } \Rightarrow \bar{x} >> 66.7 \\ \text{or} \\ \bar{x} << 66.7 \end{array} \right\} \rightarrow \text{Reject } H_0$$

Type I error: Determine $\mu \neq 66.7$, when population mean didn't change.

Type II error :- Don't reject that $\mu = 66.7$ when true $\mu \neq 66.7$.

P-value

Example :- Heights

"The mean height for 18 y/o in the 'US' in the 70s was 66.7 inches,"

"So the intuition is if the sample mean falls too far away from ^{null} hypothesis then we reject the null hypothesis (H_0)"

Now, also assume that

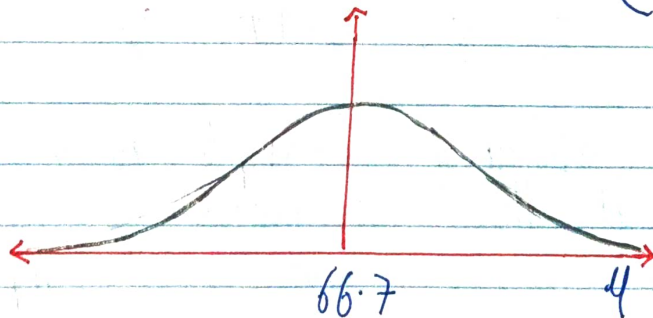
$$\sigma = 3$$

$$n = 10$$

& we assume " H_0 " which says the height didn't changed is true,

Now,

if H_0 is true: $\bar{X} \sim N(1, \frac{3^2}{10})$



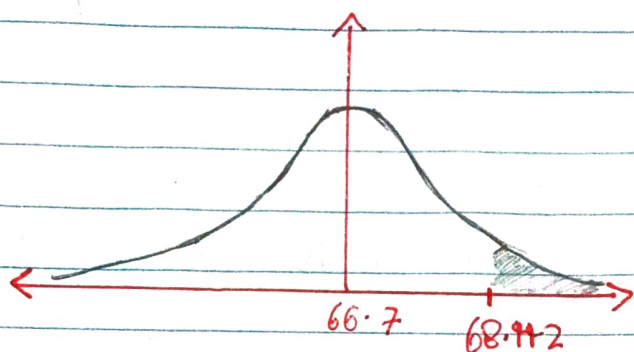
So, we ~~assume~~ ask the question,

⇒ How likely was your sample if H_0 is true?

⇒ If the answer is very unlikely, then reject H_0

Right-tailed test for Gaussian Data (known σ)

Now, remember of our sample mean was $\bar{x} = 68.442$,



$$H_0: \mu = 66.7 \text{ v.s. } H_1: \mu > 66.7$$

Goal: Type I error probability $< \alpha = 0.05$

Type I error: Determine $\mu > 66.7$, when population mean didn't change

$$\Rightarrow P(\bar{x} > 68.442 \mid \mu = 66.7) ?$$

why we take $\bar{x} > 68.442$,

So if we consider an observed sample mean of 68.442 is enough to reject " H_0 " of course, you would also reject H_0 if you saw a sample mean greater than 68.442,

The probability you are looking for is the shaded area which has the value of 0.0332.

So, $P(\bar{x} > 68.442 \mid \mu = 66.7) = 0.0332$, This is also called "P-value" the probability you just calculated

Notice that this value is less than 0.05,
 $0.0332 < \alpha$

So, we reject the null hypothesis

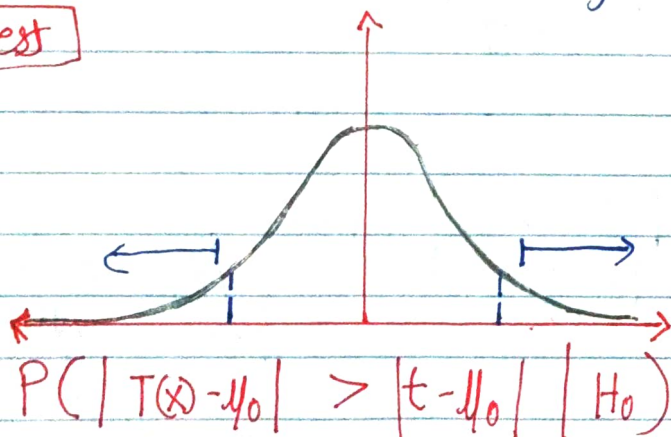
P-value definition "formal" -:

A P-value is the probability, assuming H_0 is true, that the test statistic takes on a value as extreme as or more extreme than the value observed.
 "A small P-value says that your sample landed on the tail of the distribution under H_0 , so they are pretty unlikely to happen"

Decision rule:

if $P\text{-value} < \alpha$ reject H_0 (& accept H_1 as true)
 if $P\text{-value} > \alpha$ don't reject H_0

Two tailed test



Here $\Rightarrow T(x) \rightarrow$ Test statistic

$t \rightarrow$ observed statistic

$\mu_0 \rightarrow$ Population mean

$H_0 \rightarrow$ Null Hypothesis

$H_0 \rightarrow \mu = 66.7$ for this example"

In the height case we can write,

$$P(|\bar{x} - 66.7| > |68.442 - 66.7| | H_0)$$

for left tailed test

let's say $\bar{x} = 64.222$ $P(T(x) < t | H_0)$, in this case,
 $P(\bar{x} < 64.222 | \mu = 66.7)$

Tests Using the z-statistic

So far, we used " $\bar{X}$ " statistic

$$\text{If } H_0 \text{ is true: } \bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

$$\text{(z-Statistic)} \quad z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Now after standardizing,

$$\text{if } H_0 \text{ is true: } \cancel{z = \bar{X} - \mu_0} \\ z \sim N(0, 1)$$

$$\text{Here } z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Right tailed test using the z-statistic

using the previous height example,

$$\sigma = 3 \quad \bar{n} = 68.442 \\ n = 10$$

Here,

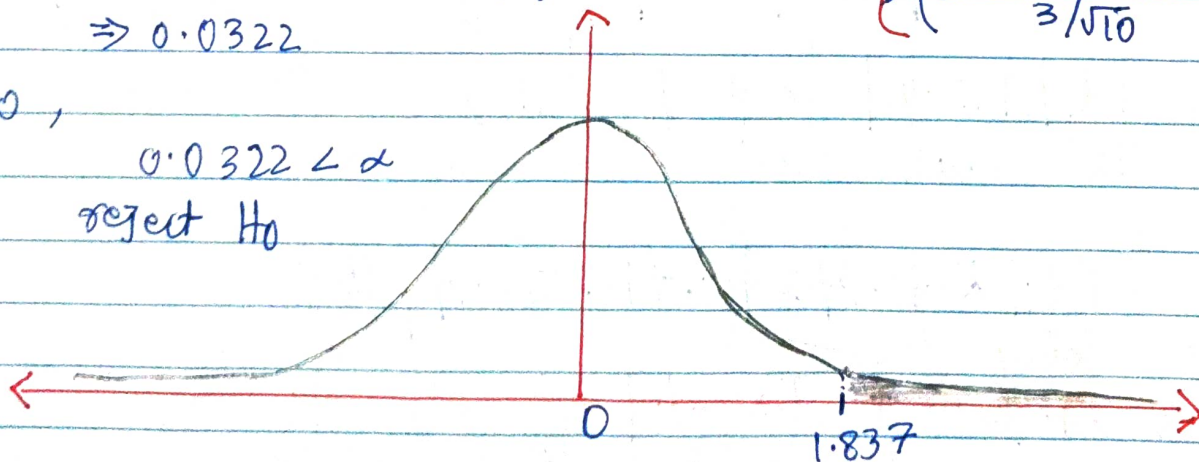
$$P(\bar{X} > 68.442 \mid \mu = 66.7) \quad \text{becomes} \\ \Rightarrow P(Z > 1.837 \mid \mu = 66.7)$$

$$\Rightarrow 0.0322$$

$$\left\{ \frac{(68.442 - 66.7)}{3/\sqrt{10}} = 1.837 \right\}$$

So,

$$0.0322 < \alpha \\ \text{reject } H_0$$

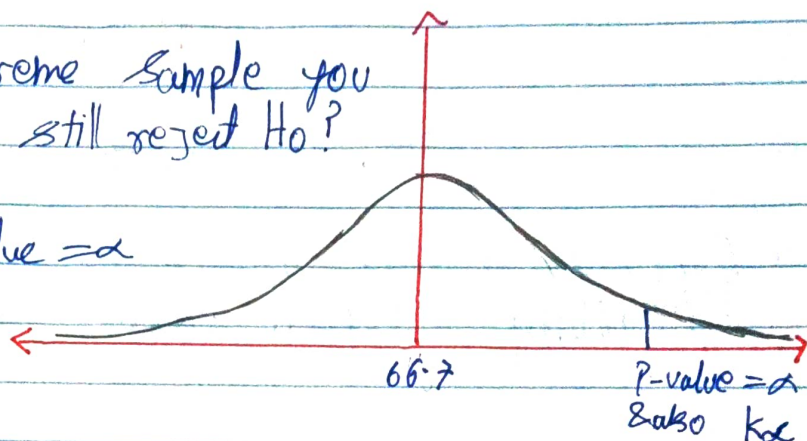


P-value

Critical values

So, what is the least extreme sample you could get that you would still reject H_0 ?

So, sample that has $p\text{-value} = \alpha$ are critical values.
denoted as " k_α "



you can create a decision rule based on critical value for example, in the height if ($k_{0.05}$) of critical value is 68.26, then ~~you~~ if you sample any mean which is greater than 68.26, you can reject the null hypothesis.

$$k_\alpha = \text{Quantile}(1 - \alpha)$$

we can write k_α as quantile of $1 - \alpha$, because that's the same thing.

so $k_{0.05} = Q_{1-0.05}$

$$k_{0.05} = Q_{0.95}$$

in probability, we write

$$0.05 = P(\bar{X} > k_{0.05} | \mu = 66.7)$$

If " α " increases then The critical value or k -value increases.

Decision rule for two-tailed test

Reject H_0 if $t > k_{\alpha_1}$ or
 $t < k_{\alpha_2}$

$$k_{\alpha_1} = Q_{1-\frac{\alpha}{2}}, \quad k_{\alpha_2} = Q_{\frac{\alpha}{2}}$$

- Defining a test in terms of critical values makes determining Type II error probabilities for the decision rule.

Power of a test & Type II error (β)

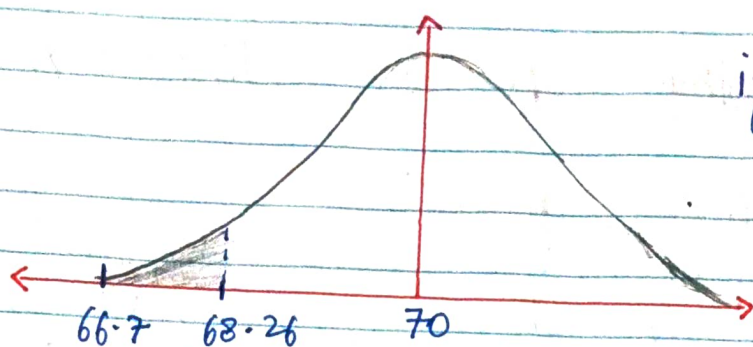
$\Rightarrow$ So, what is the Type II error, it's the β error when we accept null hypothesis even though it was wrong,

$\Rightarrow$ So, considering height example, Type II error can happen for any value bigger than 66.7.

So, what is the Type II error probability if the true value is $\mu = 70$?

Think about it we are not going to reject the null hypothesis if the mean we got is less than k_{α} "critical value", so that is the error,

& $P(\bar{X} < 68.26 | \mu = 70)$



if $\mu = 70$
 $\bar{X} \sim N\left(70, \frac{3^2}{10}\right)$

So, the shaded ~~is~~ area is the probability of type 2 error.

$P(\bar{X} < 68.26 | \mu = 70) = 0.0333 = \beta$
 The probability of type 2 error is called " β " Beta

→ This probability doesn't depend on your observed sample, it depends on the significance level you choose.

To characterize the chances of making a good decision we need something & it's called "Power test",

It tells us ~~are~~ the probability when we reject null hypothesis when it's not true, or wrong, it's given by,

$$(1 - \beta) = \text{Power test}$$

Steps for performing Hypothesis Testing :- A summary

Step 1 → State your Hypothesis

- Null Hypothesis : the baseline $\rightarrow H_0 : \mu = 66.7$
- Alternative hypothesis : the Statement you want to prove $\rightarrow H_1 : \mu > 66.7$

Step 2 → Design your test

- Decide the test statistic to work with.
like $\rightarrow$ sample mean " $\bar{x}$ ".
- Decide the significance level $\rightarrow \alpha = 0.05$
- ~~Compute~~

Step 3 → Compute the observed statistic (based on your sample)

$$\bar{x} \rightarrow 68.442$$

Step 4 → Reach a Conclusion:

- if the P-value is less than the significance level reject H_0 .

if $\downarrow \alpha$, $\uparrow \beta$, if you ^{decrease} increase the probability of type one error the probability of type II error increases.

T-distribution

We use T-distribution when we don't know the population variance which is the most common case instead we take "s" sample variance,

So, the t-statistic looks like,

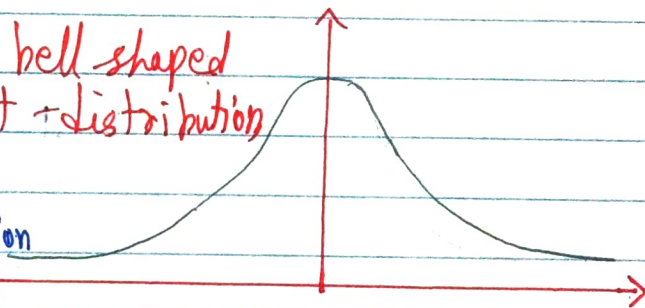
$$T\text{-statistic} \rightarrow T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

just, replace " σ " with "s".

T-statistic follows a t-distribution.

⇒ it is still bell-shaped
but follows t-distribution
which means,

The tails of the distribution
are thicker because it
reflects the extra uncertainty.



T-distribution has a parameter called

Degree of freedom (" ν ") → geek letter "nu"

$\nu = n - 1$, if as the " ν " increases the
t-distribution approaches gaussian distribution. "Here we take
n-1 to find sample variance in the denominator".

t-Test's

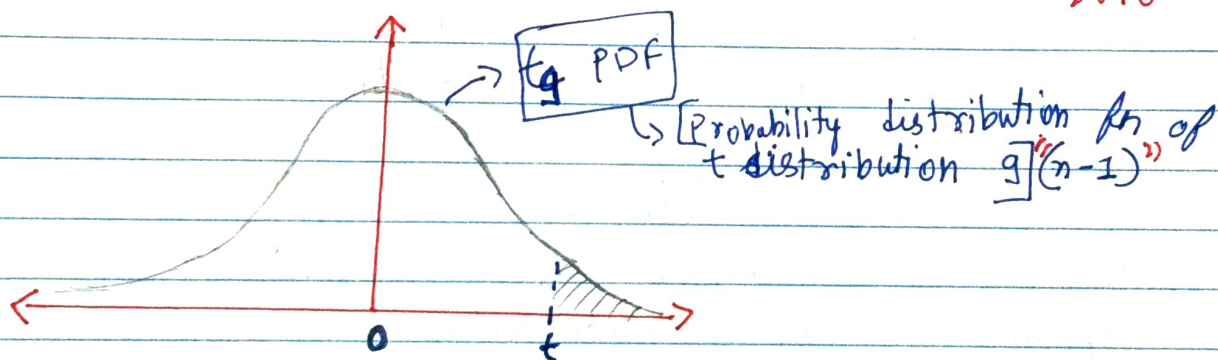
Let's take the height example:-

The mean height for 18 y/o in the US in the 70s was 66.7 inches.

[So, this time you don't know " σ "]

A small drawback is that all the computations will have to be made in "T" statistics not in the mean. So instead of plotting the distribution around 66.7, you will need to plot them around zero.

$$H_0: \mu = 66.7, \text{ if } H_0 \text{ is true: } T = \frac{\bar{X} - 66.7}{S/\sqrt{10}} \sim t_9$$



Right-tailed, Test for Gaussian Data (unknown σ)

$$H_0: \mu = 66.7 \text{ vs } H_1: \mu > 66.7, \quad n=10, \quad \alpha = 0.05$$

given,

$$\bar{X} = 68.992$$

$$S = 3.113$$

$$t = \frac{68.992 - 66.7}{3.113/\sqrt{10}} = 1.770$$

$$\text{So, } P\left(\frac{\bar{X} - 66.7}{S/\sqrt{10}} > 1.770 \mid \mu = 66.7\right) = 0.052$$

Now,

$$0.052 > \alpha \quad \left\{ \begin{array}{l} \text{So don't reject null} \\ \text{hypothesis} \end{array} \right\}$$

Two Sample t-Test

Two Sample hypothesis testing tells us how to perform hypothesis testing or how to compare samples from different population.

Independent Two-Sample t-Test

Considering the example,

You are comparing height of 18 y/o in the US with height of 18 y/o in Argentina.

Group US

$$n_x = 10 \quad \{\text{sample size}\} \quad \leftarrow \rightarrow$$

$$\bar{x} = 68.442 \quad \{\text{sample mean}\} \quad \leftarrow \rightarrow$$

$$s_x = 3.113 \quad \{\text{sample variance}\} \quad \leftarrow \rightarrow$$

Group Argentina

$$n_y = 9$$

$$\bar{y} = 65.949$$

$$s_y = 3.106$$

Goal: if the population mean is "US" is different from the population mean in Argentina.
for right tailed test,

null hypothesis " H_0 " $\Rightarrow \mu_{us} = \mu_{ar}$

Alternate hypothesis " H_1 " $\Rightarrow \mu_{us} > \mu_{ar}$

Writing this in terms of mean difference we get,

$$H_0: \mu_{us} - \mu_{ar} = 0 \quad \text{vs} \quad H_1: \mu_{us} - \mu_{ar} > 0$$

Some assumptions for independent two-sample t-Test.

- All people in the sample from the two groups are different
- Each sample person in both samples are independent.
- Populations are normally distributed.

$$X \sim N(\mu_{us}, \sigma_{us}^2), \quad Y \sim N(\mu_{arg}, \sigma_{arg}^2)$$

"How does the difference between both sample means distribute?"

Since it's a linear combination of gaussian then it will also be a gaussian.

$$\bar{X} - \bar{Y} \sim N(\mu_{us} - \mu_{arg}, \frac{\sigma_{us}^2}{10} + \frac{\sigma_{arg}^2}{9})$$

$\mu \rightarrow \mu_{us} - \mu_{arg}$
σ^2 or variance $\rightarrow \frac{\sigma_{us}^2}{10} + \frac{\sigma_{arg}^2}{9}$

Standardizing we get,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_{us} - \mu_{arg})}{\sqrt{\frac{\sigma_{us}^2}{10} + \frac{\sigma_{arg}^2}{9}}} \sim N(0, 1^2)$$

↳ The part in under root is standard deviation

But when, " σ " standard deviation is unknown,
So we replace it with the sample standard deviation

Now it will look like $\rightarrow$

$$T\text{-statistic, } T = \frac{(\bar{X} - \bar{Y}) - (\mu_{us} - \mu_{arg})}{\sqrt{\frac{S_x^2}{10} + \frac{S_y^2}{9}}} \sim t_v$$

$$\text{Degree of freedom} = \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{\left(\frac{S_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{S_y^2}{n_y}\right)^2}{n_y - 1}}$$

A Right tailed test $\therefore$

$$H_0: \mu_{us} - \mu_{arg} = 0 \text{ vs } H_1: \mu_{us} - \mu_{arg} > 0$$

$H_0 \rightarrow$ Says population mean of US & Argentina are same

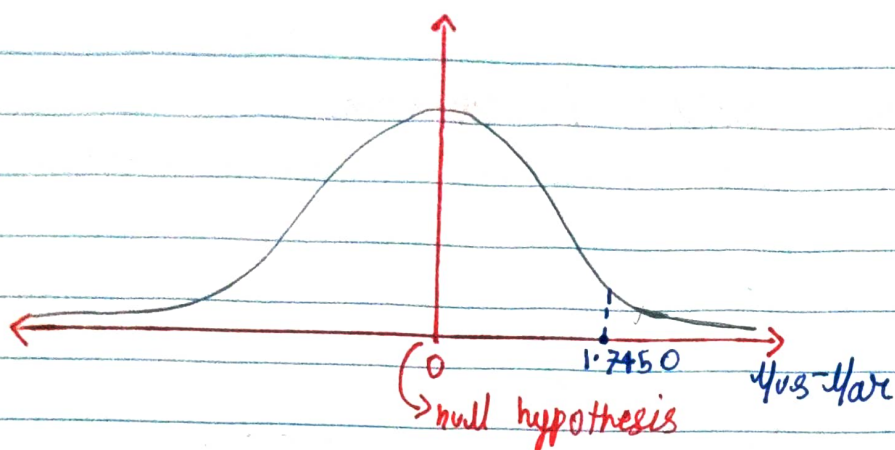
$H_1 \rightarrow$ Says Population mean of US is greater than that of Argentina.

Considering, $\alpha = 0.05$

$$\text{If } H_0 \text{ is true: } T = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\frac{S_x^2}{10} + \frac{S_y^2}{9}}} \sim t_{16.8}$$

$\bar{X} = 68.442$	$\bar{Y} = 65.949$
$S_x = 3.113$	$S_y = 3.106$
$n_x = 10$	$n_y = 9$

$$t = \frac{(68.442 - 65.949) - 0}{\sqrt{\frac{(3.113)^2}{10} + \frac{(3.106)^2}{9}}} = 1.745$$



$$P\text{-value} : P(T > 1.7450 \mid \mu_{us} - \mu_{arc} = 0) = 0.0495$$

So,

$$0.0495 < \alpha$$

So, we reject H_0

For two tailed test,

$$P\text{-value} : P(|T| > 1.7450 \mid \mu_{us} - \mu_{arc} = 0) = 0.0991$$

$$0.0991 > \alpha,$$

Don't reject H_0

for left tailed test,

P-value: we have to take a different mean & then do calculation since the instructor didn't do left tailed test so i don't have values & can't do test on my own.

Paired t-Test

⇒ it is used in hypothesis testing when two groups are not independent or simply dependent.

ex ⇒ You want to test the effectiveness of a training plan to loose weight, The first group will be people before the training program, the ~~see~~ second group is the people after training program.

In this case the samples are paired.

Now you're interested in the difference between pair of samples, $(X_i - Y_i)$
let's say,

$\begin{cases} X_i \text{ is the person before the training program} \\ Y_i \text{ is the person after the training program} \end{cases}$

Now, for all people in the group, what we need to do is sum all the differences & divide it by the number of participants, & that will give us the mean, which is the average weight loss,

So, let's call, $(X_i - Y_i) = D_i$

$$\text{Now, mean "D"} = \frac{\sum_{i=1}^n (D_i)}{n}$$

If X_i, Y_i are gaussian ⇒ $D_i = X_i - Y_i$ is gaussian
D $D_i \text{ iid } \mathcal{N}(\mu_D, \sigma_D^2)$

after standardizing, we get

$$\frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{10}} \sim N(0, 1^2) \quad \left. \begin{array}{l} \text{But } \sigma_D \text{ is unknown} \\ \text{So, } \sigma_D \rightarrow S_D \end{array} \right\}$$

S_D = Sample Standard deviation.

$$S_D = \sqrt{\frac{\sum_{i=1}^{10} (D_i - \bar{D})^2}{10 - 1}}$$

Now,

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{10}} \sim t_9$$

→ This is our test statistic.

Now, let's say " $\bar{d} = 1.09$ "
" $S_D = 1.485$ "

$$t = \frac{1.09 - 0}{1.485 / \sqrt{10}} = 2.321$$

Right tailed test

$$H_0 : \mu_D = 0 \text{ vs } H_1 : \mu_D > 0$$

Our null hypothesis is there is no weight change.
& " α " The significance level = 0.05

$$\Rightarrow P(T > t | \mu_0 = 0)$$

$$\Rightarrow P\left(\frac{\bar{D} - \mu_0}{S_0/\sqrt{10}} > 2.321 \mid \mu_0 = 0\right) = 0.0227,$$

Now ~~0.0227~~ α , So we reject H_0 .

~~0.0227~~ $\downarrow$

$$0.0227 < \alpha$$

This means that weight loss program is good.

ML application of Hypothesis testing

A/B Testing

Test for proportions

- $\Rightarrow$ "P" is the population proportion of individuals in a particular category (i.e. probability of the coin landing heads)
- $\Rightarrow$ "P₀" is the population proportion under the null hypothesis (i.e. P₀ = 0.5) we say probability of heads didn't change as null hypothesis.
- $\Rightarrow$ X is the observed number of individuals in the sample from the specified category (i.e. heads).

⇒ n is the sample size (i.e. no. of coin toss)

⇒ $\hat{p} = \frac{x}{n}$ is the sample proportion for the observed sample x .

we know that, according to CLT

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

Standardizing

$$\text{Test Statistic} \Rightarrow (Z) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} \sim N(0, 1)$$

Observed Statistic (z) ⇒ it will be the value that comes after we put known values in Test Statistic.

Now, hypothesis testing,

⇒ Right-tailed test: $[H_0: p = p_0 \text{ vs } H_1: p > p_0]$

$$P\text{-value} = P(Z > z)$$

⇒ left-tailed test: $[H_0: p = p_0 \text{ vs } H_1: p < p_0]$

⇒ Two-tailed test: $[H_0: p = p_0 \text{ vs } H_1: p \neq p_0]$
 $P\text{-value} = P(|Z| > |z|)$

Two Sample test for proportions

→ A common application for two sample test for proportions is in A/B testing.

given an example,

Imagine you want to compare the proportion of households that own a car in Chicago (P_1) with proportion of households that do in New York (P_2). Note that P_1 & P_2 are Population proportions.

Step ① Calculate proportions

Chicago
Car owners = 62 = X
(n₁) Total people checked = 100

New York
Car owners = 58 = Y
(n₂) Total people checked = 120

$$\hat{P}_1 = \frac{62}{100} = \frac{X}{100}, \quad \hat{P}_2 = \frac{58}{120} = \frac{Y}{120}$$

Now, $\hat{P}_1 \sim N(H_0: P_1 = P_2 \text{ vs } H_1: P_1 \neq P_2)$

$$\hat{P}_1 = \frac{62}{100} \sim N\left(P_1, \frac{P_1(1-P_1)}{100}\right),$$

$$\hat{P}_2 = \frac{58}{120} \sim N\left(P_2, \frac{P_2(1-P_2)}{120}\right)$$

Now, the difference between $\hat{P}_1$ & $\hat{P}_2$ is

$$\hat{\Delta} = \hat{P}_1 - \hat{P}_2 \sim N\left(P_1 - P_2, \frac{P_1(1-P_1)}{100} + \frac{P_2(1-P_2)}{120}\right)$$

Since in hypothesis we have taken $P_1 = P_2$, so

$$\hat{\Delta} = \hat{P}_1 - \hat{P}_2 \sim N\left(0, P(1-P)\left(\frac{1}{100} + \frac{1}{120}\right)\right)$$

Standardizing $\hat{\Delta}$ you get,

$$Z \Rightarrow \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2) \rightarrow 0}{\sqrt{P(1-P)\left(\frac{1}{100} + \frac{1}{120}\right)}} \sim N(0, 1)$$

In the above standardized eqn we don't know the value of 'P'.
we don't have Population proportion but we have sample proportion.

So, since we have assumed that $P_1 = P_2 = P$,

we are essentially assuming that both the readings were taken from same sample (~~P₁ = P₂~~), ($P_1 = P_2$). So we write

$$\hat{P} = \frac{X+Y}{100+120} = \frac{X+Y}{220}$$

Putting, value of aggregated proportion we get,

$$Z \Rightarrow \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{100} + \frac{1}{120}\right)}} \sim N(0, 1)$$

$$\left\{ \begin{array}{l} \Rightarrow \text{we know that,} \\ \hat{P}_1 = \frac{X}{100}, \hat{P}_2 = \frac{Y}{120} \\ \hat{P} = \frac{X+Y}{220} \end{array} \right\} \quad \frac{\left(\frac{X}{100} - \frac{Y}{120}\right)}{\sqrt{\frac{X+Y}{220} \cdot \left(1 - \frac{X+Y}{220}\right) \cdot \left(\frac{1}{100} + \frac{1}{120}\right)}}$$

$$\Rightarrow Z = \frac{\left(\frac{62}{100} - \frac{58}{120}\right)}{\sqrt{\frac{62+58}{220} \cdot \left(1 - \frac{62+58}{220}\right) \cdot \left(\frac{1}{100} + \frac{1}{120}\right)}} = 2.0271$$

for, right tailed test, $\alpha = 0.05$
 $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 > 0$

$$P\text{-value} = P(|Z| > z \mid P_1 = P_2)$$

$$= P(|Z| > 2.0271 \mid P_1 = P_2) = 0.04265$$

So Now,

$$0.04265 < \alpha,$$

So reject H_0 .

left tailed test, $\alpha = 0.05$

$$H_0: P_1 - P_2 = 0 \text{ vs } H_1: P_1 - P_2 < 0$$

$$P\text{-value} = P(Z < z)$$

Right tailed test, $\alpha = 0.05$

$$H_0: P_1 - P_2 = 0 \text{ vs } H_1: P_1 - P_2 \neq 0$$

$$P\text{-value} = P(|Z| > |z|)$$

Now, the real ML stuff for A/B testing

ex $\Rightarrow$ Suppose, A company has a web page & it wants to test two different placing for the buy button.

first is, Design A & Design B

Aim: Decide if switching to strategy B leads to a higher purchase amounts,

To do this you select a group of people, as customers come in they are randomly assigned to design A or design B,

In A/B Testing most of them are assigned design "A" & a smaller number is assigned to Design "B" which is the new design because we don't know how good it will be,

Design A

$$n_A = 80$$

$$S_A = 10$$

$$\bar{x} = 50 \$$$

Design B

$$n_B = 20$$

$$S_B = 15$$

$$\bar{y} = 55 \$$$

$$H_0: \mu_A = \mu_B \quad \text{vs} \quad H_1: \mu_A \neq \mu_B \quad \mu_A < \mu_B$$

in difference of mean we can write

$$\alpha = 0.05$$

$$H_0: \mu_A - \mu_B = 0 \quad \text{vs} \quad H_1: \mu_A - \mu_B < 0$$

$$\text{If } H_0 \text{ is true: } T = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \sim t_{23.38}$$

$$t = \frac{(50 - 55) - 0}{\sqrt{\frac{10^2}{80} + \frac{15^2}{20}}} = \boxed{-1.914}$$

Now,

$$P(T \leq t | \mu_A = \mu_B)$$

or

$$P(T \leq t | \mu_A - \mu_B = 0) = 0.085$$

Now,

~~as~~

$$0.085 < \alpha$$

So we

don't reject null hypothesis.

A/B Testing: Conversion rates

Let's say you have 2 designs, now you want to check the conversion rates on those designs, which website has better conversion rates

Design-A

$$n_A = 80$$

$$x = 20$$

"Total customers"

"Conversions"

Design-B

$$n_B = 20$$

$$y = 8$$

$$\text{Now, } H_0: P_A - P_B = 0 \quad \text{vs} \quad H_1: P_A - P_B < 0$$

null hypothesis $\Rightarrow$ Both the conversions are same

Alternative hypothesis $\Rightarrow$ Conversion of website B is greater than that of A.

Here,

P_A = Conversion rate from Design A

P_B = " " " " " "

Let's say $\alpha = 0.05$

according to Law of Large numbers, if sample size is high enough then sample proportion tends to population proportion

$$\frac{X}{n_A} \rightarrow P_A$$

$$\longrightarrow \frac{X}{n_A} \sim N\left(P_A, \frac{P_A(1-P_A)}{n_A}\right)$$

$$\frac{Y}{n_B} \rightarrow P_B$$

$$\xrightarrow{\text{C.L.T}} \frac{Y}{n_B} \sim N\left(P_B, \frac{P_B(1-P_B)}{n_B}\right)$$

taking the difference we get,

$$\frac{X}{n_A} - \frac{Y}{n_B} \sim N\left(P_A - P_B, \frac{P_A(1-P_A)}{n_A} + \frac{P_B(1-P_B)}{n_B}\right)$$

Since, we assumed that $P_A = P_B = P$

$$\Rightarrow \frac{X}{n_A} - \frac{Y}{n_B} \sim N\left(0, \frac{P(1-P)}{n_A} + \frac{P(1-P)}{n_B}\right)$$

Standardizing, we get

$$Z = \frac{\left(\frac{X}{n_A} - \frac{Y}{n_B}\right) - 0}{\sqrt{\frac{P(1-P)}{n_A} + \frac{P(1-P)}{n_B}}} \sim N(0, 1^2)$$

But we don't know P , so replace

P , with estimation,

$$\hat{P} = \frac{X+Y}{n_A+n_B}$$

we get, test statistic

$$Z = \frac{\left(\frac{x}{n_A} - \frac{y}{n_B}\right) - 0}{\sqrt{(x+y) \left(1 - \frac{x+y}{n_A+n_B}\right)}} \sqrt{n_A n_B} \sim N(0, 1^2)$$

Now,

$$Z = -1.336$$

$$P(Z < z \mid p_A - p_B = 0) = 0.091$$

Since,

$$0.091 > \alpha,$$

we do not reject null hypothesis.