26.12.2022

YZV 231E

Probability Theory & Stats

Week 14

Gü.

Recap:

$\boxed{\text{MLE}}$: Maximum Likelihood Estimation :   Note: θ is not an r.v.

Model w/ unknown parameters $\underline{\theta}$;

$$X \sim \underbrace{p_X(x;\underline{\theta})}_{\text{a family of parameterized models}} \longrightarrow \left.\begin{array}{l} p_X(x;\theta_1) \\ p_X(x;\theta_2) \\ \vdots \\ p_X(x,\theta_N) \\ \vdots \end{array}\right\} \text{ some possibilities.}$$

MLE picks θ that makes the data most likely : $\arg\max_{\theta} p_X(x;\underline{\theta}) = \hat{\theta}$.

Compare to Bayesian approaches to estimation: Θ is an r.v.

$\boxed{\text{MAP}}$: $\arg\max_{\theta} p_{\Theta|X}(\theta|x) = \hat{\theta}$

$$p_{\Theta|X}(\theta|x) = \frac{\overbrace{p_{X|\Theta}(x|\theta)}^{\text{likelihood}}\,\overbrace{p_\Theta(\theta)}^{\text{prior}}}{p_X(x)}$$

find θ most likely under the posterior distrib.

Note: MLE & MAP appear the same when we have a uniform prior, but in principle they are very different.

$\boxed{\text{LMS}}$ $\doteq E_{\Theta|X}[\Theta|X] = \hat{\theta}$
estimator

— Sample Mean Estimator of θ

r.v. $\boxed{\hat{\theta}_n = \dfrac{X_1 + \cdots + X_n}{n}}$: $\boxed{\text{point}}$ estimator.

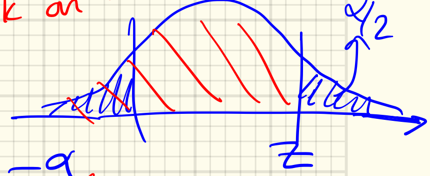
→ Properties of an estimator: Unbiased $E[\hat{\theta}] = \theta$, Bias $= \hat{\theta} - \theta$. , Consistent $\hat{\theta} \xrightarrow[\text{in prob}]{} \theta$, "small" MSE $\approx Var(\hat{\theta}) + (Bias)^2$

— $(1-\alpha)$ Confidence Interval (CI)

$$P\Big(\underset{\text{r.v.}}{\hat{\theta}_n^-} \leq \theta \leq \underset{\text{r.v.}}{\hat{\theta}_n^+}\Big) \geq \underbrace{1-\alpha}_{\text{e.g. } 95\%}, \quad \forall \theta.$$

pick an α → fix z.

Construction of the CI : w/ $\boxed{\text{CLT}}$:



Confidence Interval for the Sample mean $\hat{\theta}_n$

$$P\Big(\underbrace{\hat{\theta}_n - \frac{z\,\sigma}{\sqrt{n}}}_{\hat{\theta}_n^-} \leq \theta \leq \underbrace{\hat{\theta}_n + \frac{z\,\sigma}{\sqrt{n}}}_{\hat{\theta}_n^+}\Big) \approx \underbrace{1-\alpha}_{95\%}$$

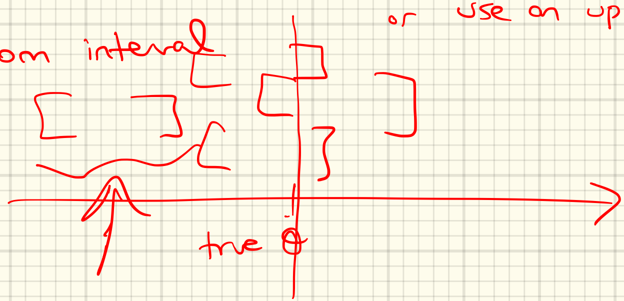
$\alpha = 0{,}05$

where z is s.t $\underbrace{\Phi(z)}{} = 1 - \dfrac{\alpha}{2} = 0.975$

– need an estimate of the variance σ^2; either use sample variance, or use an upper bound if any.

CI → $[\hat{\theta}_n^-, \hat{\theta}_n^+]$: random interval

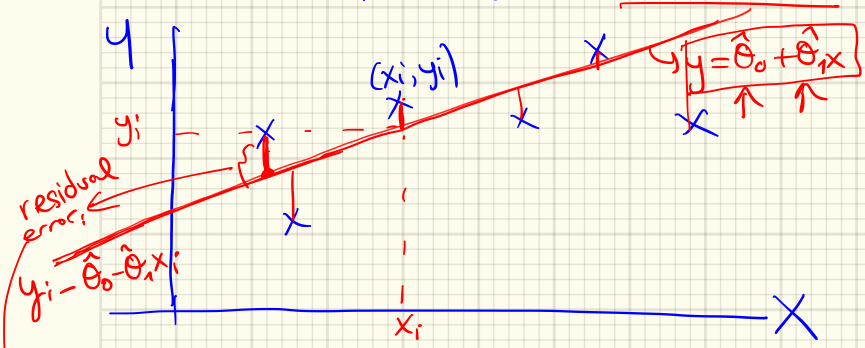
$\hat{\theta}_n^-$: r.v. $\hat{\theta}_n^+$: r.v.



REGRESSION:

Let X be your TYT exam score
Let Y be your ITU GPA

Q. Is there a relation between the two?



Goal: Find the "best" model to explain the data

Always ask: "optimal" or "best" w.r.t. which criterion (measure)?

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) = \{x_i, y_i\}_{i=1}^{n}$

Model: $y_i \approx \boxed{\theta_0} + \boxed{\theta_1} x_i$

unknown parameters that define our model.

Minimize the Residual error: $y_i - \theta_0 - \theta_1 x_i$:

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (\underbrace{y_i - \theta_0 - \theta_1 x_i}_{\text{residual error}})^2$$

★ Cost: sum of squared errors in predictions.

Probabilistic Interpretation:

model the GPA score $\rightarrow$ $$y_i = \theta_0 + \theta_1 x_i + \boxed{w_i}, \quad w_i \sim \mathcal{N}(0, \sigma^2)$$

w_i: random noise

$\rightarrow w_i$'s are indep. $\forall i$.

$\mathcal{N}(0,\sigma^2)$: choose a specific probabilistic model.

Want to do $\boxed{\text{MLE}}$ estimation:

Write a likelihood fn. $\quad P_{y,x;\theta}(x; y; \theta)$

likelihood $\sim$ probability

likelihood of w $\quad w \sim c\, e^{-w_i^2/2\sigma^2}$

c: normalization coef.

$\sim$ likelihood of y $\quad p(y;x,\theta) \sim C \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\underbrace{y_i - \theta_0 - \theta_1 x_i}_{w_i})^2\right.$

write this for all samples $y_1, \dots, y_n$.

a sum b/c w_i's are independent.

maximize w.r.t. θ_0, θ_1: ~~you can~~ take a logarithm

$\rightarrow$ this is the same cost as in ★ previous page.

$\therefore$ Linear Regression $\equiv$ MLE where $\underbrace{w_i}_{\text{i.i.d.}} \sim \mathcal{N}(0, \sigma^2)$

Linear Regression:

model: $y \approx \theta_0 + \theta_1 X$

optimization prob → $$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Solution: set the derivatives of the cost function to zero:

exercise, derive these $\hat{\theta}_0$ & $\hat{\theta}_1$ expressions.

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \dots + y_n}{n}$$

$$\boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

$$\boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Covariance X, Y

$\approx E[(X - E[X])(Y - E[Y])]$

(*)

or through a probabilistic interpretation:

Our model: $Y = \theta_0 + \theta_1 X + W$, $X \perp\!\!\!\perp W$ are independent w/ zero mean. (for simplicity)

exercise: try w/o this assumption.

$$E[Y] = \theta_0 + \theta_1 E[X] + 0$$

$$\theta_0 = \underbrace{E[Y]} - \theta_1 \underbrace{E[X]}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \cdot \bar{x}$$ ✓ given $\hat{\theta}_1$ already

Now, Obtain $\hat{\theta}_1$ (estimate):

$$Y \cdot X = \theta_0 \cdot X + \theta_1 X^2 + W \cdot X$$

$$\underbrace{E[Y \cdot X]} = \underbrace{\theta_0 \cdot E[X]}_{\cancel{\ }\,0} + \theta_1 \cdot \underbrace{E[X^2]} + \underbrace{E[W \cdot X]}_{E[W] \cdot E[X] \,=\, 0}$$

for zero mean r.v.s

$$= \; Cov(X, Y) = \theta_1 \cdot Var(X)$$

$$\rightarrow \theta_1 = \frac{Cov(X, Y)}{Var(X)}$$

$\longrightarrow$ $cov(X, Y) \approx \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

$\longrightarrow$ $Var(X) \approx \frac{1}{n} \sum_i (x_i - \bar{x})^2$

→ Multiple Linear Regression:

eg. include more variables that may affect your ITU GPA: your high school GPA, years of education of parents, family income, ... or such

Data: $\left(x_i^{(1)},\ x_i^{(2)},\ x_i^{(3)},\ y_i\right)$

$x_i^{(1)}$ → high school gpa

$x_i^{(2)}$ → tyt.

$x_i^{(3)}$ → education years of parents

y_i → itu gpa

Model: $\boxed{y_i \sim \theta_0 + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)}}$: linear fn. of all the variable

not squared!

superscript, not to the power!

multiple explanatory variables in our model.

$$\min_{\theta_0,\theta_1,\theta_2,\theta_3} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1 x_i^{(1)} - \theta_2 x_i^{(2)} - \theta_3 x_i^{(3)}\right)^2$$

take derivatives w.r.t. $\theta_0, \theta_1, \theta_2, \theta_3$, set to 0, you get a system of linear equations.

In Matrix notation: you can get a closed form solution.

Digression: In vector notation: set $\underline{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}_{\substack{n\times 1 \\ 4\times 1}}$ $\underline{x}_i = \begin{bmatrix} 1 \\ x_i^{(1)} \\ x_i^{(2)} \\ x_i^{(3)} \end{bmatrix}_{\substack{n\times 1 \\ 4\times 1}}$

m data points $\{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_m\}$:

$$\boxed{\min_{\underline{\theta}} \|\underbrace{\underline{y}}_{\text{vector}} - \underbrace{\underline{\underline{X}}}_{\text{matrix}}\,\underbrace{\underline{\theta}}_{\text{vector}}\|^2}, \quad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}_{m\times 1}$$

m: # data instances

$$\underline{\underline{X}}_{m\times n} = \begin{bmatrix} - & \underline{x}_1^T & - \\ & \underline{x}_2^T & \\ & \vdots & \\ & \underline{x}_m^T & \end{bmatrix}_{m\times n} \quad (1\times n \text{ rows})$$

$$f(\underline{\theta}) = \left(\underline{y}^T - (\underline{\underline{X}}\,\underline{\theta})^T\right)\left(\underline{y} - \underline{\underline{X}}\,\underline{\theta}\right)$$

$$f(\underline{\theta}) = \underline{y}^T\underline{y} - \underline{y}^T\underline{\underline{X}}\,\underline{\theta} - \underline{\theta}^T\underline{\underline{X}}^T\underline{y} + \underline{\theta}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\theta}$$

take deriv. w.r.t. $\underline{\theta}$

$$\nabla_{\underline{\theta}} f = \begin{bmatrix} \frac{\partial f}{\partial \theta_0} \\ \frac{\partial f}{\partial \theta_1} \\ \vdots \\ \frac{\partial f}{\partial \theta_n} \end{bmatrix} = 0$$

$$-2\underbrace{\underline{y}^T\underline{\underline{X}}}_{=\underline{\underline{X}}^T\underline{y}} + 2\,\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\theta} = 0$$

$$\boxed{\underline{\theta} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{y}}$$

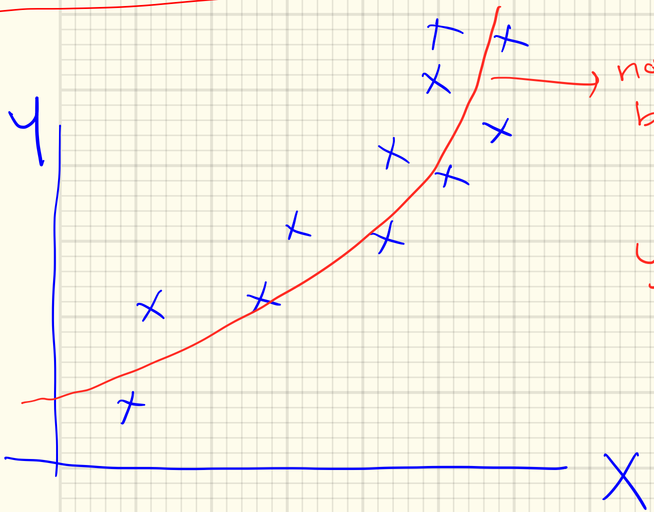
a closed form solution for multiple linear regression.

$$\underline{\theta}_{n\times 1} = \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1}_{n\times n} \underline{\underline{X}}^T_{n\times m} \, \underline{y}_{m\times 1}$$: Normal Equations

gives us the $\underline{\theta}$ vector estimates for linear regression.

$$\underline{\underline{X}} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix}_{m\times n}$$

row vector for each data instance



not a linear but a quadratic model of the measurements.

$$y \approx \theta_0 + \theta_1 \cancel{x} \; h(x)$$

Now, use: nonlinear functions of the data.

Model: $$y \approx \theta_0 + \theta_1 \boxed{h(x)}$$

still, this is a linear regression

eg. $$y \approx \theta_0 + \theta_1 x^2$$

→ Same formulation

• Data points: $(y_i, h(x_i))$

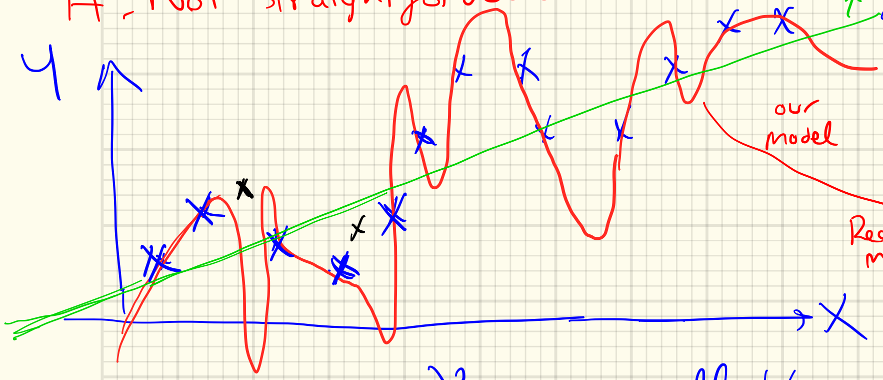
• model: $y_i = \theta_0 + \theta_1 h(x_i)$, $\forall i = 1, \dots, m$

$$\min_{\substack{\theta \\ (\theta_0, \theta_1, \dots)}} \sum_{i=1}^{m} \Big(y_i - \theta_0 - \theta_1 \underbrace{h(x_i)}_{\text{fn. of } x \text{: e.g. } x_i^3 \to 3^{rd}}\Big)^2$$

still a linear model.

Q. In your linear regression, how do you choose $h(x)$'s?

A. Not straightforward:



$y_i = \theta_0 + \theta_1 x_i$

you try to go thru all data points by using a large # parameters, e.g. use an 8th degree polynomial

Red model: $y_i = \sum_{j=0}^{8} \theta^{(j)} x_i^j \to h(x_i)$

$= \theta^{(0)} + \theta^{(1)} x_i + \theta^{(2)} x_i^2 + \theta^{(3)} x_i^3 + \dots + \theta^{(8)} x_i^8$

$\left(y_i - \sum_{j=0}^{8} \theta^{(j)} x_i^j\right)^2$: error is small b/c we have lots of parameters.

Q: Is the red model a good model? No!

→ Your model cannot generalize to a new data point well!

→ Overfitting problem!

→ Choosing complex $h(.)$ vs simple $h(.)$?

Q. How complex? How many explanatory variables?

→ open & extensive research topic,

— When you have a $\boxed{\text{few data points}}$, avoid using too many parameters in your model.

→ Good rule: Start w/ simpler models ≡ a few parameters, especially when you have a few data points. Gradually increase later w/ more data etc.

Notes For these θ_i, people also report confidence interval (not covered in this class)

— R^2: measure of explanatory power of the model in your linear regression.

— Standard error estimates of σ^2 → uncertainty in the model.

$$y = \theta_0 + \theta_1 X + \omega \longrightarrow \sigma^2 \text{: variance of the noise}$$

related to R^2

$$\boxed{\frac{Var(Y|X)}{Var(Y)}} < 1 \quad \text{naturally}$$

adding knowledge of Y, how much of the randomness in Y is reduced.

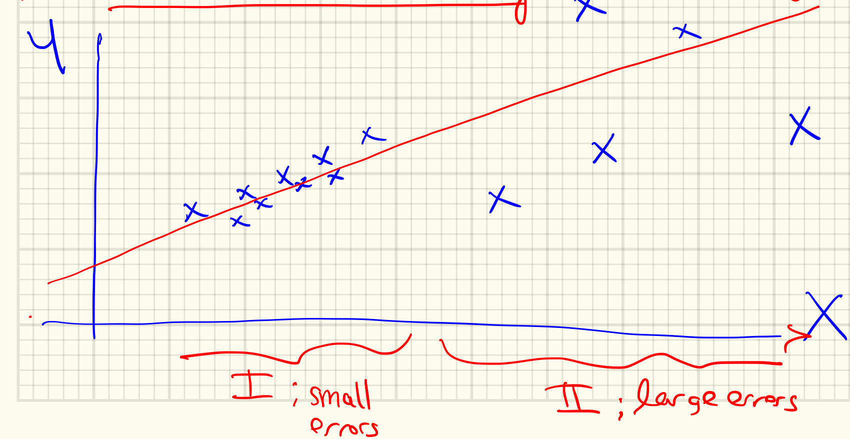
If this is small, including X as my explanatory variable for the target variable Y helps / improves our predictions on Y.

60% of the student's ITU GPA is explained by the TYT score.

Some Pitfalls in Using Linear Regression:

* Heteroskedasticity:

A linear model that you fit to this data.



A good fit in Region I.

model:
I) has small variance
II) has large variance.

— Be careful w/ this problem. (not covered in this class.)

eg.
— Need to incorporate varying Var(W).

* Multi-Collinearity: Multiple explanatory variable; they are closely w/ each other.

eg. model $Y_{GPA} = \theta_0 + \underline{\theta_1}\,\underline{TYT_1} + \underline{\theta_2}\,\underbrace{TYT_2}_{AYT}$.

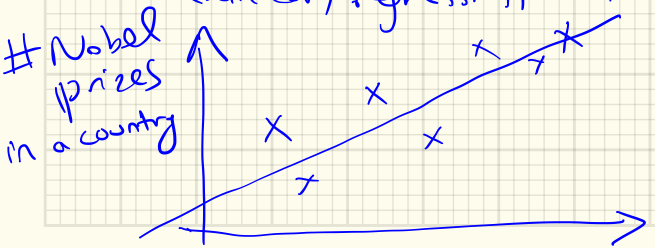
Your TYT & AYT (2 exam scores) are close to each other.

→ Correlated → Redundancy.

$$Y_{GPA} = \theta_0 + \theta_1 TYT_1$$
$$Y_{GPA} = \theta_0 + \theta_2 TYT_2$$

avoid such redundancy in explanatory variables. b/c they create sensitivity of the model to small changes in the data.

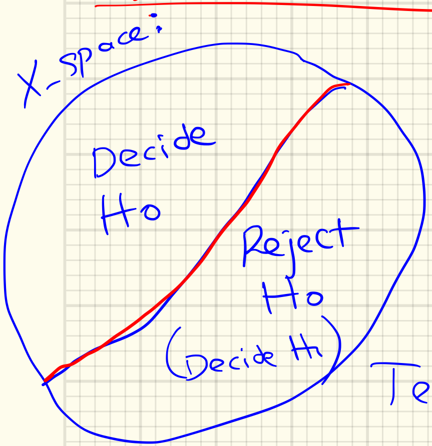
* Causality: Do not use (linear) regression to conclude causality!



Never say that Y is CAUSED by X according to your linear regression model

Hypothesis Testing:

We have a null H_0 X and an alternate H_1 hypothesis.



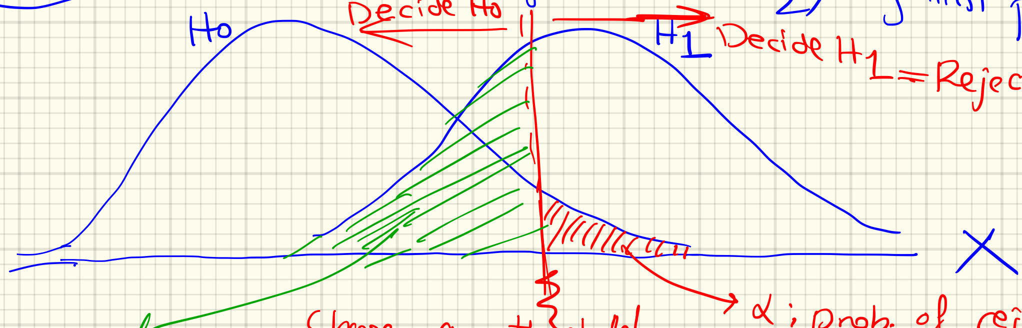
Coin: Fair or not;

$$H_0 : p = \frac{1}{2} \quad \text{vs.} \quad H_1 : p \neq \frac{1}{2}$$ single alternate hypothesis

or

H_1 : $p = 0.55$, $p = 0.6$, $p = 0.65$ many alternate hypothesis

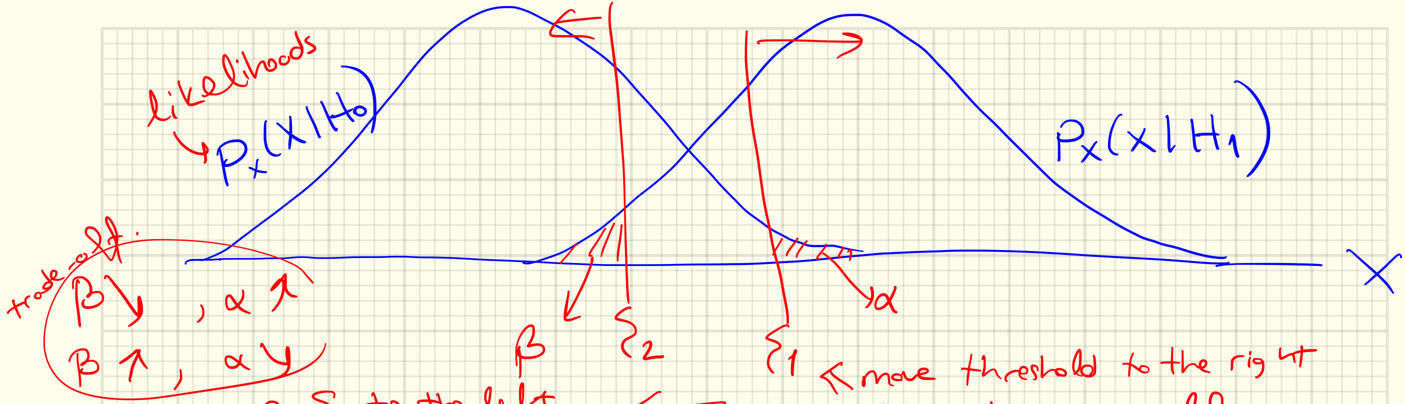
Test coin's fairness $(p = \frac{1}{2})$ against $p \neq \frac{1}{2}$



Choose a threshold

α ; prob. of rejecting H_0 when it was true.

β : error I make for not rejecting H_0, actually when H_0 should have been rejected.



move ξ to the left to make β small.

$\Leftarrow$ move threshold to the right to make α small

$\Rightarrow$ want both α, β to be small; however $\exists$ a trade-off.

Q. How to set the threshold ξ?

A. $\boxed{\text{Likelihood Ratio Test (LRT)}}$; Compare the posterior prob. of the hypothesis.

Choose H_1 : if $P(H_1 | X=x) > P(H_0 | X=x)$

Pick the hypothesis which is more likely, given the data.

$$P(H_1 \mid X=x) \gtrless P(H_0 \mid X=x)$$

using Bayes

In a Bayesian setting (MAP), use Bayes rule:

$$\frac{P(X=x \mid H_1)\, p(H_1)}{p(X=x)} > \frac{P(X=x \mid H_0)\, P(H_0)}{P(X=x)}$$

$$\Rightarrow L(x) \triangleq \boxed{\frac{P(X=x \mid H_1)}{P(X=x \mid H_0)}} > \boxed{\frac{P(H_0)}{P(H_1)}} \quad \text{(LRT)}$$

likelihood ratio → compare to a threshold ξ.

ξ : threshold $\propto$ ratio of the prior prob. of the hypothesis. (Bayesian view)

↑ Bayesian setting

— In a non-Bayesian setting; don't have prior probabilities

Still $\left\{ \dfrac{P_X(X=x\,;H_1)}{P_X(X=x\,;H_0)} > \xi \right.$ so we re-write i.t.o. pdf of x.

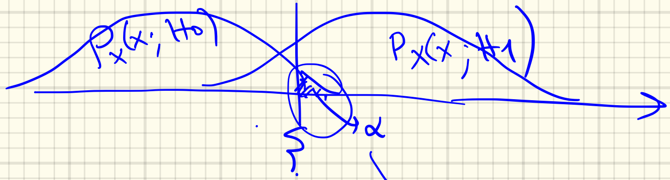
Q: If ($L(x)$) this ratio is large, $\boxed{Q}$: Is it likely that my observations X occured under H_0?

No!

No ; it is unlikely that observations X occured under H_0.

$\therefore$ Reject H_0.

— Threshold ξ trades-off 2 types of error :



Choose ξ. s.t.

$$P(\text{Reject } H_0 ; H_0) = \alpha.$$

$$1 - CDF_X(\xi) = \alpha.$$

we fix α, eg. $\alpha = 0.05$ $\rightarrow$ find ξ (threshold)

$\rightarrow$ then β is already fixed.

Simple $\boxed{\text{Binary Hypothesis Testing}}$

Want to make a decision whether to Reject or Not Reject the null hypothesis

- (Default) Null Hypothesis H_0 : $\underset{\text{r.v.}}{X} \sim P_X(x; H_0)$
- Alternative Hypothesis H_1 : $X \sim P_X(x; H_1)$

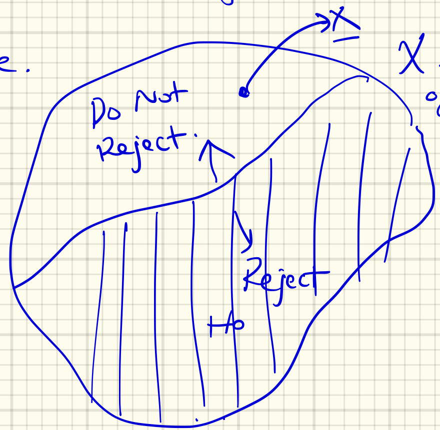
Designing the Hypothesis Test (checking whether H_0 is false or not)

1) Structure of the test : shape of the dividing curve.

ex. Likelihood Ratio Test :

$$\frac{P_X(x; H_1)}{P_X(x; H_0)}$$

2) Given the shape, where to place the division ?



X – space of observations

1) LRT : Reject H_0 if :

$$L(x) = \frac{P_X(x; H_1)}{P_X(x; H_0)} > \xi$$

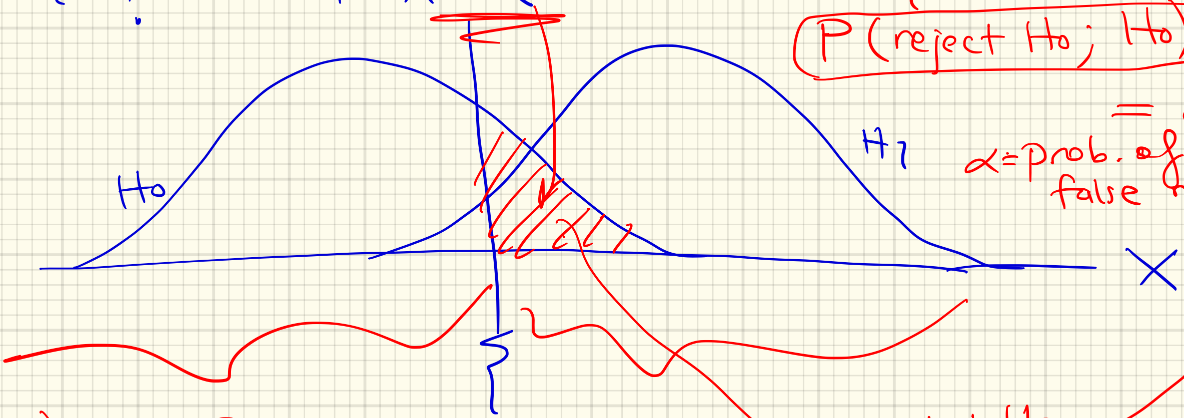
2) How to choose ξ ?

2) ξ ?

Fix $\alpha \longrightarrow$ choose ξ so that

$$\boxed{P(\text{reject } H_0 ;\ H_0) = \alpha}$$

$=$ α = prob. of false rejection



$L(x) < \xi$; Do not Reject Ho.

$L(x) > \xi$; reject Ho

e.g. set $\alpha = 0.05$ (5%) $\rightarrow$ that sets ξ.

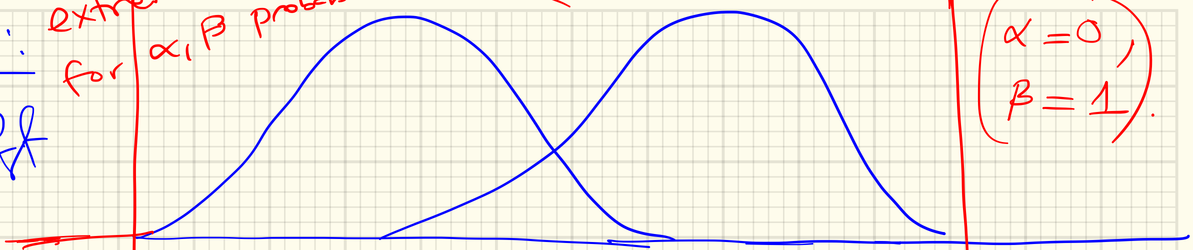
Note:

(α,β) trade off

extreme cases for α, β probabilities

H_0

$(\alpha = 0, \beta = 1)$.

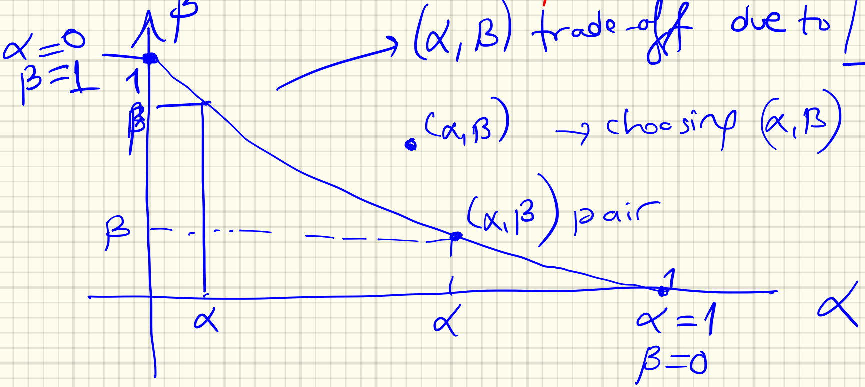


extreme

always decide $H_1 \rightarrow (\beta = 0, \alpha = 1)$

extreme

not rejecting H_0 at all.



(α,β) trade-off due to $\boxed{LRT}$

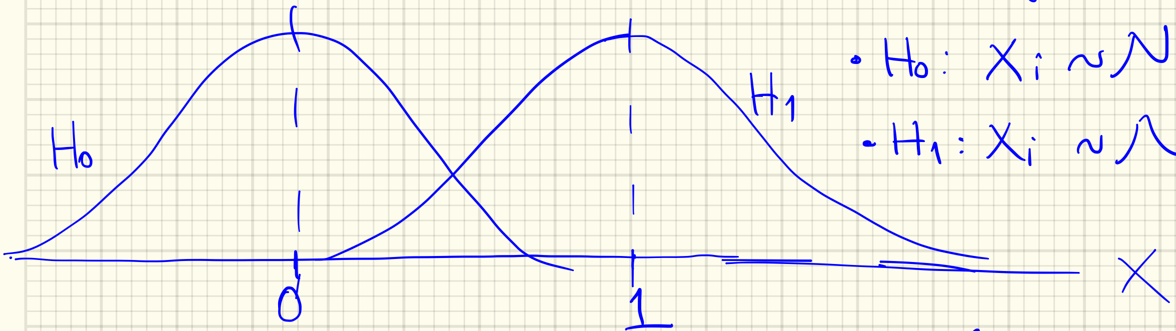
• (α,β) → choosing (α,β) w/ method other than LRT.

Note: Theoretically
For a given α value,
LRT minimizes
the probability β.

Ex: Hypothesis Test on Normal Means

• n data points, X_i: i.i.d. and normal

You have 2 normal distributions w/ different means



• $H_0: X_i \sim \mathcal{N}(\boxed{0}, 1)$

• $H_1: X_i \sim \mathcal{N}(\boxed{1}, 1)$

1) Likelihood Ratio Test; Reject H_0 if:

$$\frac{p_X(x; H_1)}{p_X(x; H_0)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n} (x_i - 1)^2/2\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n} x_i^2/2\right\}} > \xi$$

H_1

H_0

exercise: do some algebra to simplify to.

1) LRT test

Reject H_0 if : $\sum_i X_i > \xi'$

a test "statistic"

Summarizes our measurements into a single number

$\xi' = \log \xi + \frac{n}{2}$

Intuitive here, if $\sum_i X_i$ is large → evidence to reject H_0.

2) How to choose ξ' ? Set prob of false rejection to a certain probability α.

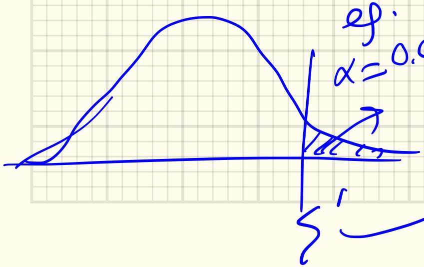
eg. 5%.

$$P\left(\sum_{i=1}^{n} X_i > \xi' ; H_0\right) = \alpha$$

X_i's normal → sum X_i's

$\sum X_i$ → Normal distrib.

→ Use Normal tables



eg: $\alpha = 0.05$

→ $\xi' = 1.96$

If $\sum_i X_i > 1.96$; Reject H_0.

< 1.96 ; Do Not Reject H_0.

Ex: Hypothesis Test on Normal Variances.

- n data points X_i, i.i.d.

$H_0 : \mathcal{N}(0, \boxed{1})$

$H_1 : \mathcal{N}(0, \boxed{4})$

same mean but different variances.

LRT: Rejection (of H_0) region:

$$\frac{\text{Density of data under } H_1}{\text{Density under } H_0} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_i X_i^2 / 2(4)\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_i X_i^2 / 2(1)\right)} > \xi$$

Do algebra to simplify to:

* Reject H_0 if $\underbrace{\boxed{\sum_i X_i^2}}_{\text{test statistic}} > \xi'$

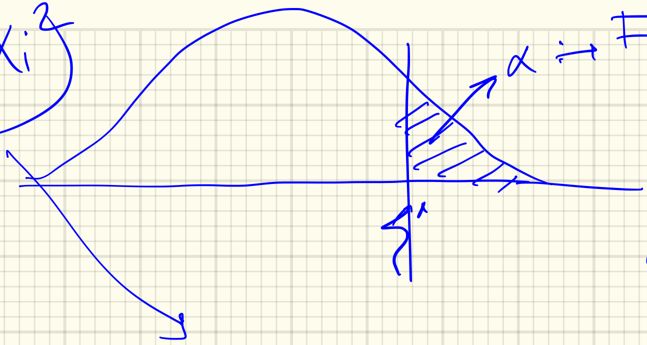
* Find ξ' s.t. $P\left(\sum_i X_i^2 > \xi' ; H_0\right) = \alpha$.

Distribution of $\sum_i X_i^2$ is known: χ^2 (Chi-squared distrib)

Tables are available

Recall: derived distrib.

$\sum_i X_i^2$



$\alpha \to$ Fix the tail prob. of the χ^2-distrib.
e.g. to 95^{th} percentile.
From χ^2-tables, read off the ξ' that corresponds to 0.95.

Note: Your "statistic": $\sum_i X_i^2$

If $\sum_i X_i^2 > \xi'$: Reject H_0

$\leq \xi'$: Do not Reject H_0.

Composite Hypothesis : eg. coin → is it fair or unfair?

→ You make n tosses of the coin.

You get $S = 474$ Heads in $n = 1000$ tosses?

Is the coin fair?

$H_0 : p = \frac{1}{2}$ (fair) vs $H_1 : p \neq \frac{1}{2}$ (unfair)

$p = 0.51, 0.52, \ldots$

$\boxed{\text{Expected value} : \frac{n}{2}}$: half heads half tails

H H H T T H T T T . . . . (1000)

(i) Pick a $\boxed{\text{statistic}}$:

come up w/ $S = \#$ Heads

: Design / pick a statistic $\equiv$ reasonable summary of your data

(ii) Pick shape of the rejection region :

1) $\equiv$ Decide how to make your decision.

$$\left| S - \underbrace{\frac{n}{2}}_{\text{expected value eg. } 500} \right| > \xi$$: Reject the hypothesis.

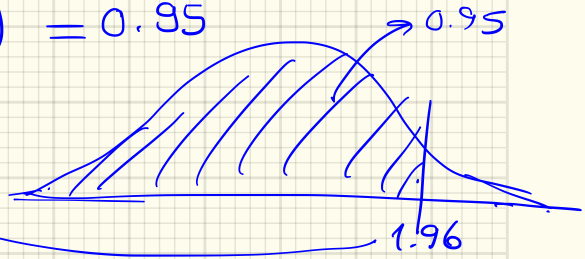
iii) Pick a significance level α (eg. $\alpha = 0.05$)

iv) Pick a threshold ξ s.t.

$$P(\text{reject } H_0 \,;\, H_0) = \alpha$$ ≈ probability of outliers

Using CLT: # heads, i.e. S statistic is Normal.

$$P(|S - 500| \le \xi \,;\, H_0) = 0.95$$



From the Normal table:

$\Phi(z) = 0.95 \to z = 1.96$.

normalize S

$$-1.96 \le \frac{S - 500}{\sqrt{Var(S)}} \le 1.96$$

$\underbrace{Var(S)}_{n \cdot \sigma^2}$

$1000 \cdot \frac{1}{4} = 250$ → use an upper bound on σ^2. (recall Bernoulli)

$$\boxed{S - 500 \le (1.96)\, 250 \approx 31 = \xi}$$

→

Test: $\rightarrow$ $|S - 500| \leq 1.96\sqrt{250} \simeq 31 = \xi$

For our ex. $S = 474 \rightarrow |S - 500| = 26 < \xi = 31$

$\rightarrow$ Do not Reject H_0 (at the 5% level of error.)

$\equiv \exists$ 5% chance that the data we got is an outlier.

Note: Say H_0 is Not Rejected rather then ~~H_0: accepted~~.

H_0: default hypothesis $\rightarrow$ we do not reject it until we see: evidence contrary to H_0.

Ex: Is your die fair? $i = 1, \ldots, 6$.

H_0: is a pmf: $\boxed{P(X=i) = p_i = \frac{1}{6}}$

Null Hypothesis → fair die.

- For each i $(1, \ldots, 6)$: N_i : # occurences for each

Roll your die $\boxed{n \text{ times}}$, count

$\# 1\text{'s} \to N_1$

$\# 2\text{'s} \to N_2$

$\vdots$

$\# 6\text{'s} \to N_6$

You Observe N_i's : Is your die fair?

Under H_0: I expect N_i's : $\boxed{N_i} = \frac{n}{6} = n \cdot p_i = n \cdot \frac{1}{6}$

1) Choose a form of Rejection region.

Reject H_0 if $T = \sum_i \frac{(\overbrace{N_i}^{\text{observed}} - \overbrace{n \cdot p_i}^{\text{expected}})^2}{n \cdot p_i} > \xi$

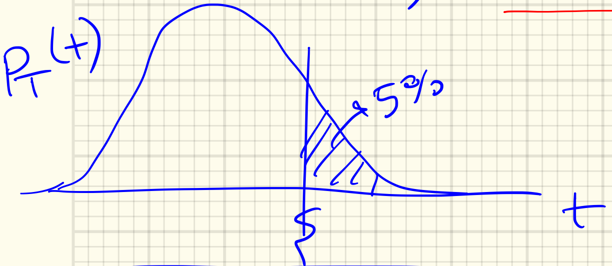
2) Choose ξ so that prob of false rejection 5%.

$$P(\text{reject } H_0 ; H_0) = 0.05$$

$$P(T > \xi ; H_0) = 0.05$$

We need distrib. of Test statistic: $\boxed{T = \sum_i \frac{(N_i - n p_i)^2}{n p_i}}$ $\Leftarrow$ derived distrib.

For large n, $T \sim$ a chi-squared distribution (widely used in statistical tests)



$\exists$ tables $\rightarrow$ set ξ. (threshold)

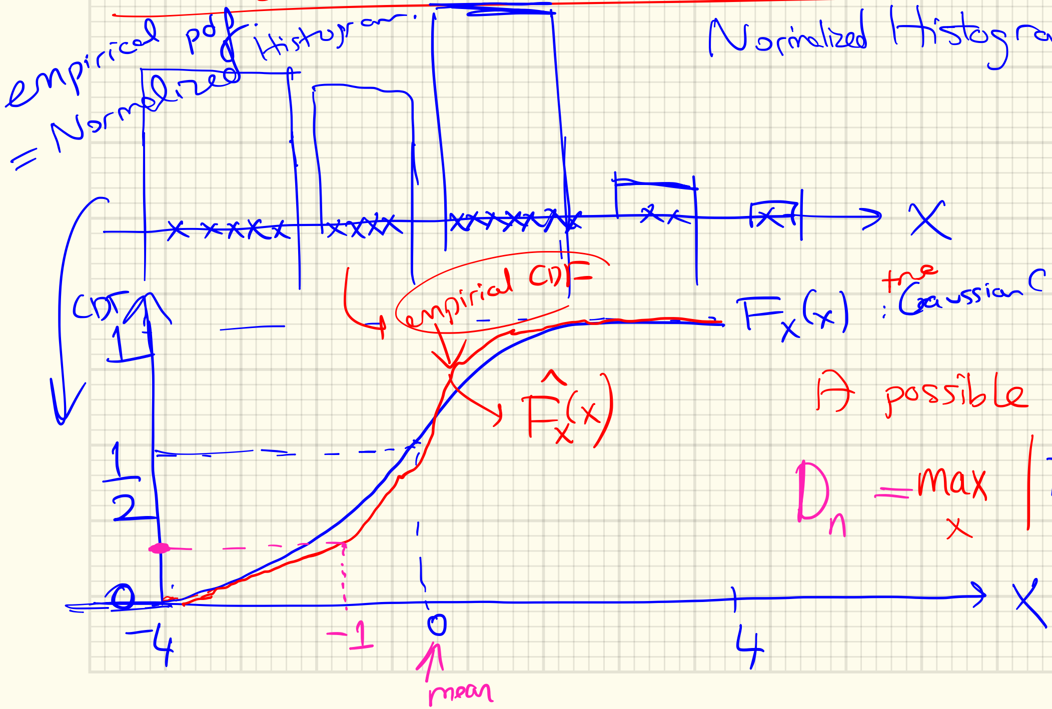
$$\boxed{\begin{array}{l} \text{Reject } H_0 \text{ if } T > \xi \\ \text{Do not Reject } H_0 \text{ if } T \leq \xi . \end{array}}$$

$\rightarrow$ Decide whether to Reject H_0 or Not!

Ex: Want to Test whether your data comes from a certain Gaussian distrib?

→ Kolmogorov - Smirnov (KS) Test; From CDF (empirical)

Normalized Histogram ($\simeq$ pdf) → CDF.



$H_0: X \sim N(0,1)$

∄ possible (KS test)

$$D_n = \max_x \left| F_X(x) - \hat{F}_X(x) \right|$$

If D_n is small → Do Not Reject H_0.

$\rightarrow P\left(D_n \geq \frac{1.36}{\sqrt{n}}\right) \approx 0.05.$

KS test is frequently used D_n has a known ~~calculated~~ distrib. $\rightarrow$ tabulated prob values of D_n.

$\xi = \frac{1.36}{\sqrt{n}}$ (n : # data points) w/ 5% rejection prob.

If $D_n \geq \xi \longrightarrow$ Reject H_0.

THE END.

$\rightarrow$ Now, you have learned the basics of statistical methods, any statistically-literate engineer/scientist should know about.