YZV 231E

10.01.2022

Probability Theory & Stats

Gü.

Recap:

MLE : Maximum Likelihood Estimation:

model w/ unknown parameters

↳ $X \sim p_X(x;\theta)$:

$p_X(x,\theta_1)$, $p_X(x,\theta_2)$, ⋮ } a family of parameterized of models

pick θ that makes data likely:

$$\max_\theta \; p_X(x,\theta)$$

Compare to Bayesian

– MAP estimation :

$$\max_\theta \; \underbrace{p_{\Theta|X}(\theta|x)}_{\text{posterior}} = \max_\theta \frac{p_{X|\Theta}(x|\theta)\, \overbrace{p_\Theta(\theta)}^{\text{prior}}}{p_X(x)}$$

→ Sample mean Estimate of θ

r.v. $\hat{\Theta}_n = \frac{X_1 + \cdots + X_n}{n}$ →

→ $(1-\alpha)$ confidence Interval

$$P\left(\hat{\Theta}_n^- \le \theta \le \hat{\Theta}_n^+\right) \ge 1-\alpha \quad , \forall\theta$$

e.g. 95%

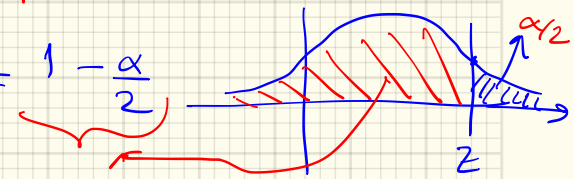
Note: MLE & MAP appear the same when we have a uniform prior in MAP.

Construction of CI : Easy w/ CLT :

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

$$P\left(\hat{\theta}_n - \frac{z \cdot \sigma}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + \frac{z \cdot \sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

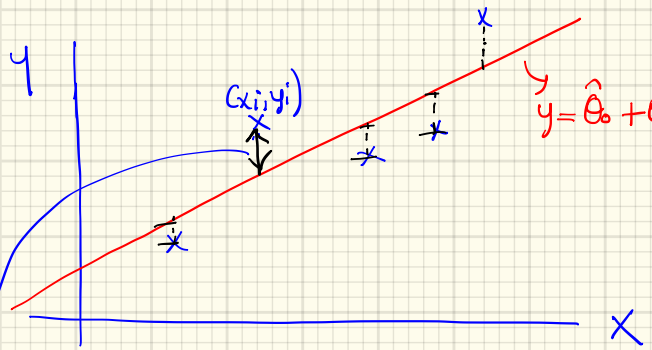
where z is s.t. $\Phi(z) = 1 - \frac{\alpha}{2}$



pick an α → fix z.

need an estimate of the variance σ^2.

REGRESSION :



Let x be your TYT exam score (High school GPA)
Let y be your ITU GPA

Is there a relation between the two?

- Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Model : $y \overset{\sim}{\sim} \theta_0 + \theta_1 x$

θ_0, θ_1 : unknown parameters that define our model.

- Find the "best model" to explain the data :
  ("always ask : w.r.t. which criterion ?")
  "best" $\rightleftarrows$ "optimal"

minimize residual error : $y_i - \theta_0 - \theta_1 x_i$

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

residual error

cost : sum of squared error in predictions.

→

→ Probabilistic interpretation:

→ $y_i = \theta_0 + \theta_1 x_i + w_i$, $w_i \sim \mathcal{N}(0, \sigma^2)$, i.i.d.

Do MLE ; $P_{Y|X}(x, y; \theta)$: likelihood

$w \rightarrow$ $c\, e^{-w_i^2/2\sigma^2}$

$\rightarrow$ $c \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \right\}$ (w_i^2)

w_i's are indep. $\left(e^{-w_1^2/2\sigma^2} \cdot e^{-w_2^2/2\sigma^2} \cdots \right)$

take a log

Maximize w.r.t. θ_0 & θ_1.

↳ After you take log → this is the same cost as in (★)

$\Rightarrow$ Linear Regression (Least Squares) $\equiv$ MLE where $w_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Linear Regression:

- Model : $y \approx \theta_0 + \theta_1 x$

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

- Solution : set the derivatives of the cost fn. to zero:

exercise: obtain θ_0 & θ_1.

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

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

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

$\approx$ Covariance (X, Y)

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

(*)

Interpretation of the Solution

Assume a model: $Y = \theta_0 + \theta_1 X + W$ w/ X, W are independent w/ zero mean.

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

$$\theta_0 = E[Y] - \theta_1 E[X]$$

estimate w/ sample mean

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$: we assumed we have $\hat{\theta}_1$.

To obtain θ_1 estimate:

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

$$E[Y \cdot X] = 0 + \theta_1 \mathrm{Var}(X) + \cancel{E[W]E[X]}^{\,0}$$

for zero mean r.v.s $= \mathrm{Cov}(X, Y) = \theta_1 \cdot \mathrm{Var}(X)$

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

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

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

→ You may want to include more variables that your ITU GPA may depend on: family income, high school gpa, years of education of parents..

Multiple Linear Regression:

- Data: $(x_i^1, x_i^2, x_i^3, y_i)$, $i = 1, \dots, n$
  (x^1: hist, x^2: high-sch gpa, x^3: education years of parents, y: itu gpa)
- Model: $y \approx \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \theta_3 x^3$: linear fn. of all the variables
- Formulation: (x^1, x^2, x^3: multiple explanatory variables)

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

(take deriv. w.r.t. parameters, set to 0, you get a system of linear eqns.

→ In matrix notation: ⟶

Digression:

In vector notation: set $\underline{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$ $n \times 1$ (4×1)

$\underline{x}_i = \begin{bmatrix} 1 \\ x_i^{(1)} \\ x_i^{(2)} \\ x_i^{(3)} \end{bmatrix}$ $n \times 1$ (4×1)

$$\min_{\underline{\theta}} \underbrace{\| \underline{y} - \underline{\underline{X}}^T \underline{\theta} \|^2}_{= f(\underline{\theta})} \quad \min \text{ w.r.t. } \underline{\theta}$$

$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}_{m \times 1}$; $\underline{\underline{X}} = \begin{bmatrix} | & | & & | \\ \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_m \\ | & | & & | \end{bmatrix}_{n \times m}$

data instances

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

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

$\underline{\underline{X}}^T = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix}$ ↑

derivative w.r.t. $\underline{\theta}$

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

$$-2 \underline{\underline{X}} \cdot \underline{y} + 2 \underline{\underline{X}} \underline{\underline{X}}^T \underline{\theta} = 0$$

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

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

Let $\underline{\underline{X}} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix}$

$$\underset{n\times 1}{\theta} = \underbrace{(\underset{m\times n}{X}^T \underset{m\times n}{X})^{-1}}_{n\times n} \underbrace{\underset{n\times m}{X}^T}_{n\times m} \underset{m\times 1}{y}$$

: Normal Equations give us the θ estimates for linear regression.

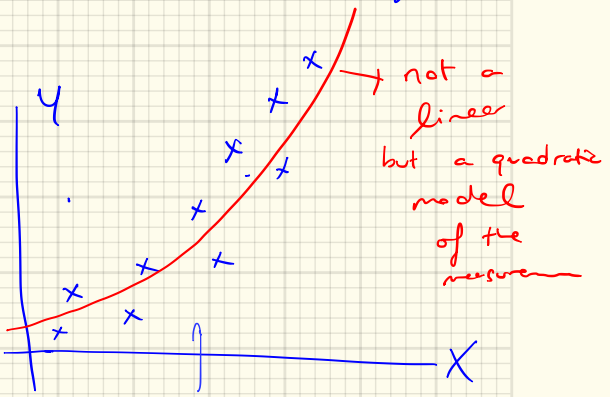
$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}_{m\times n}$$

$$y \approx \theta_0 + \theta_1 h(x)$$

Nonlinear functions of the Data

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

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



Same

- Formulation

- Data points : $(y_i, h(x_i))$
- model : $y = \theta_0 + \theta_1 \cdot h(x)$

$$\min_{\theta} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1 \underbrace{h(x_i)}_{\text{fn. of } x} \right)^2$$

eg. x^3

Still a linear model

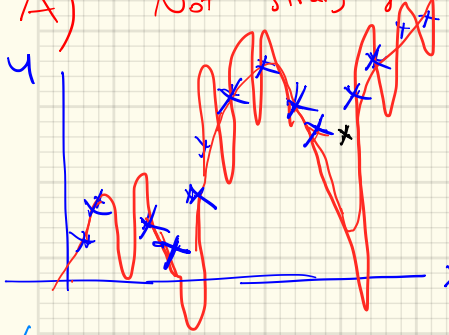
Q) In your linear regression, how do we choose $h(x)$'s ?

A) Not straightforward : b/c eg. you try to go over all data points by using a large # of parameters.

eg. use an pth degree polynomial

$$y_p = \sum_{j=0}^{p} \theta^j \underbrace{x_i^j}_{h(x_i)}$$

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



$\left(y_i - \sum_j \theta^j x_i^j\right)$ → error is small → we have lot of parameters

→ Q : is this a good model? No!

→ Choosing complex $h(\cdot)$ ≡ how many exploratory variables you should be using?

How much complex?

→ Open topic.

Good rule

→ Start w/ simpler models, especially when you have few data points. → a few parameters

---

Notes: In practice;

→ confidence intervals for θ_i: (not covered here)

→ "Standard error" estimates of σ: noise in the model.

$$Y \approx \theta_0 + \theta_1 X + W \rightarrow \sigma^2$$

(σ^2: variance of the noise)

uncertainty in the model

→ R^2: measure of the explanatory power of the model for linear regression.

→ $\frac{Var(Y|X)}{Var(Y)}$ → (variance / uncertainty) in Y conditioned on X.

< 1 naturally.

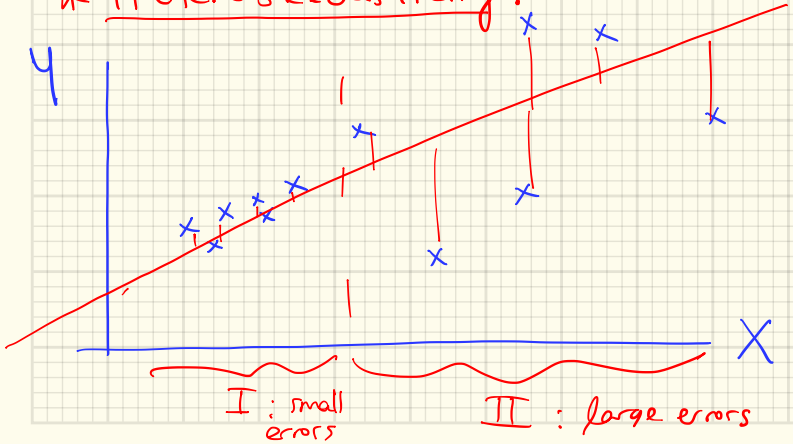
→ (uncertainty) in Y / Variance of Y. w/o knowing any other information

e.g. "60% of students GPA is explained by the TYT score."

↓ in statistical studies.

Using Linear Regression : Some pitfalls.

* **Heteroskedasticity:**



A good fit in Region I,

I) has small variance

II) has large variance.

X-space has very different variances in different regions.

→ Be careful (not covered here).

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

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

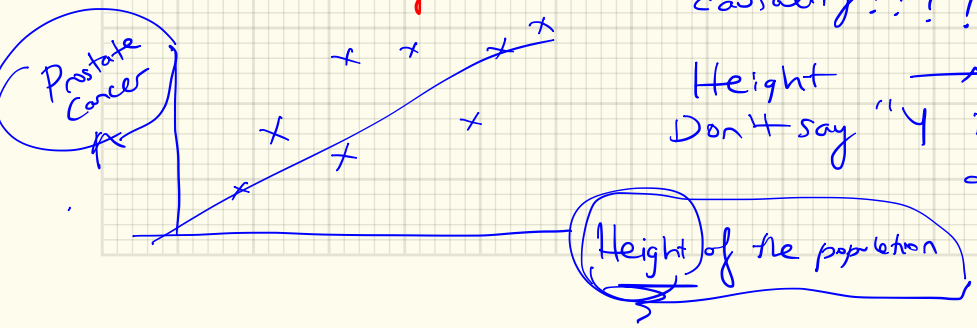
your TYT_1 & TYT_2 (2 exam scores) are close to eachother

$y_{GPA} = \theta_0 + \theta_1 TYT_1$

or $= \theta_0 + \theta_2 TYT_2$

avoid such redundancy in explanatory variables.

* **Causality !!** people use linear regression to conclude causality!!!!



Height → prostate

Don't say "Y is caused by X according to your model !!"

good linear

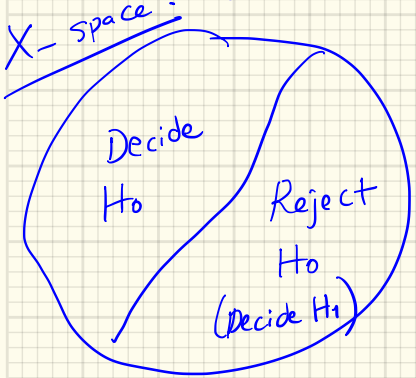
Hypothesis Testing:

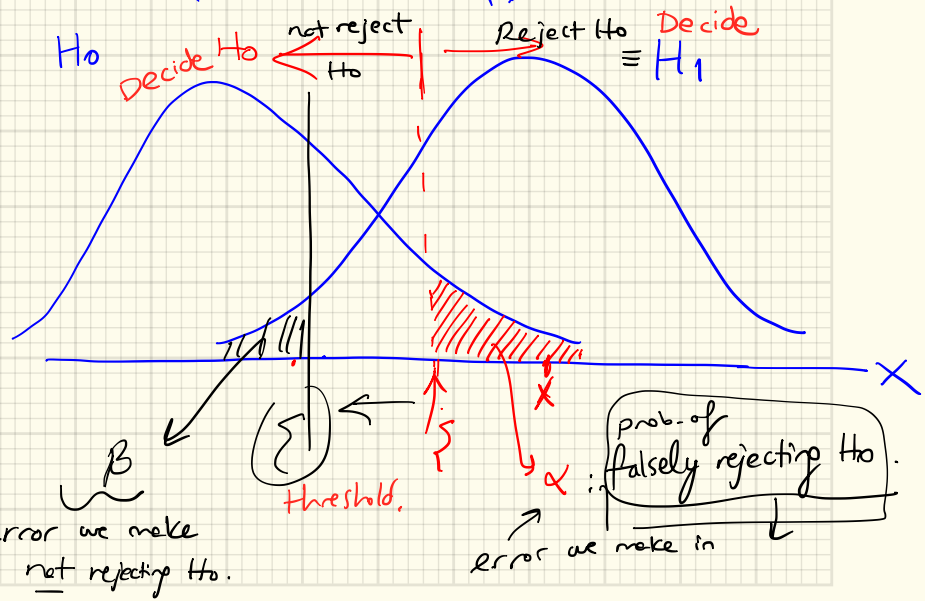
We have a null (H_0) & and an alternate (H_1) hypothesis

Coin: fair or not: $H_0 : p = \frac{1}{2}$ vs. $H_1 : p = \frac{3}{4}$ or $p \neq \frac{1}{2}$

Testing coin's fairness ($p = \frac{1}{2}$) against $p \neq \frac{1}{2}$ or $p = \frac{3}{4}$.

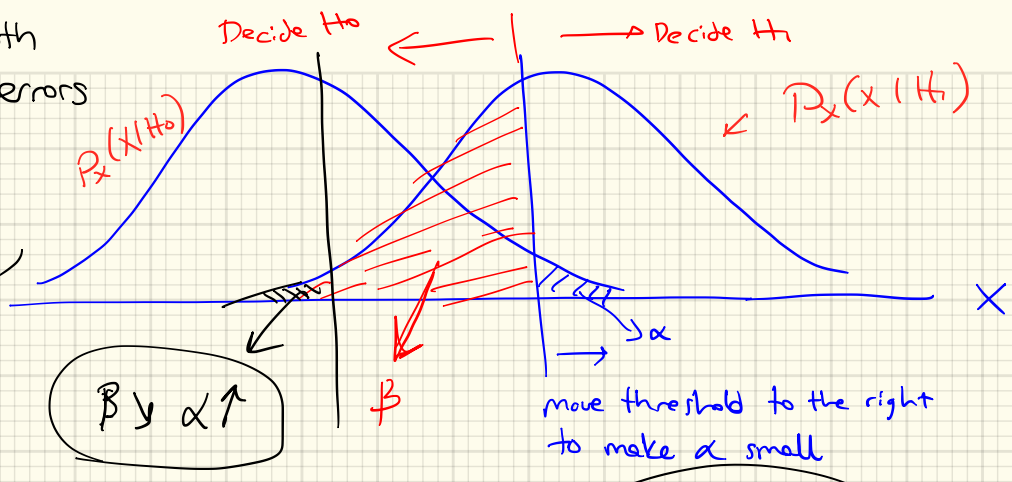
H_0 vs H_1.





We want both α & β errors to be as small as possible,

$\Rightarrow$ a trade-off



Q. How to set the threshold?

A. Likelihood Ratio Test (LRT): look at posterior prob. of the hypothesis

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

pick the hypothesis that is more likely, given the data.

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

$$\frac{P(X=x|H_1)\,P(H_1)}{P(X=x)} > \frac{P(X=x|H_0)\,P(H_0)}{P(X=x)} \quad \text{rewrite} \Rightarrow$$

$$L(x) = \frac{P(X=x \mid H_1)}{P(X=x \mid H_0)} > \frac{P(H_0)}{P(H_1)} \qquad \text{(LRT)}$$

likelihood ratio $\longrightarrow$ compare to a threshold ξ.

- In a Bayesian setting: threshold $\xi \propto$ prior probabilities of the two hypothesis.
- In a non-Bayesian setting, we don't have prior probabilities

$$\xi \quad \frac{P(X=x \mid H_1)}{P(X=x \mid H_0)} \overset{\text{prob.}}{>} \xi \quad \overset{\text{i.t.o. pdf}}{\equiv} \quad \frac{P_X(x; H_1)}{P_X(x; H_0)} > \xi$$

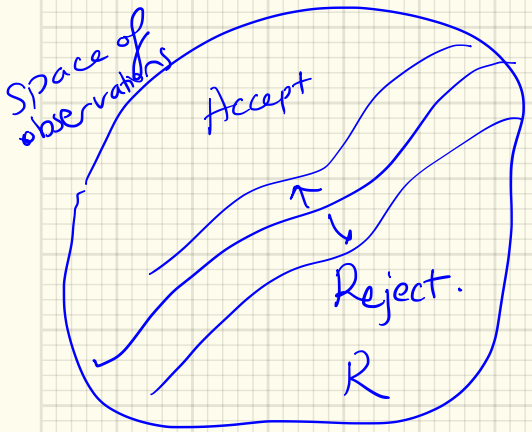
if this ratio is large,
it is unlikely that my observation X occured under H_0,
$\therefore$ reject H_0.

Simple Binary Hypothesis Testing:

- Default hypothesis:
  - ↳ Null Hypothesis H_0: $X \sim p_X(x; H_0)$
  - Alternative Hypothesis H_1: $X \sim p_X(x; H_1)$

Want to check whether H_0 is false or not.

Whether to reject or not reject the null hypothesis.



Designing the hypothesis test:

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

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

1) Choose a rejection region R

Reject H_0 iff data $X \in R$.

→ Structure of the test: Likelihood Ratio test

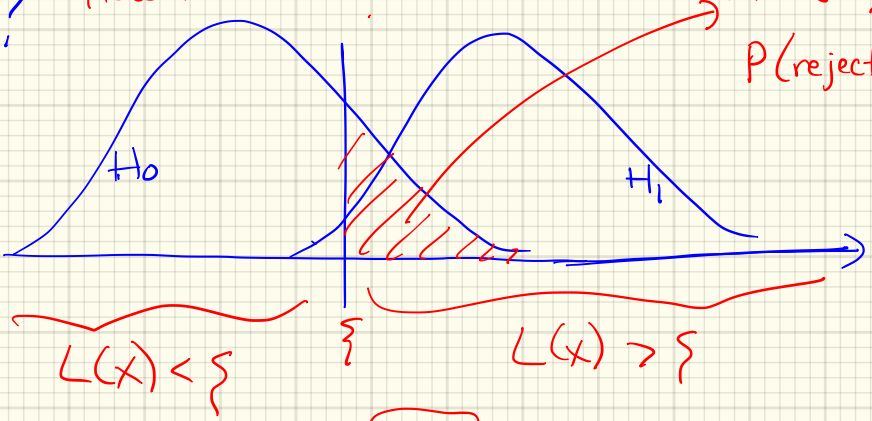
1) $L(x) = \dfrac{P_x(x; H_1)}{P_x(x; H_0)} > \xi$

High value of $L(x)$ ≡ likelihood of H_1 > likelihood of H_0

2) ξ ? How to choose ξ ? → Fix α ; choose ξ so that

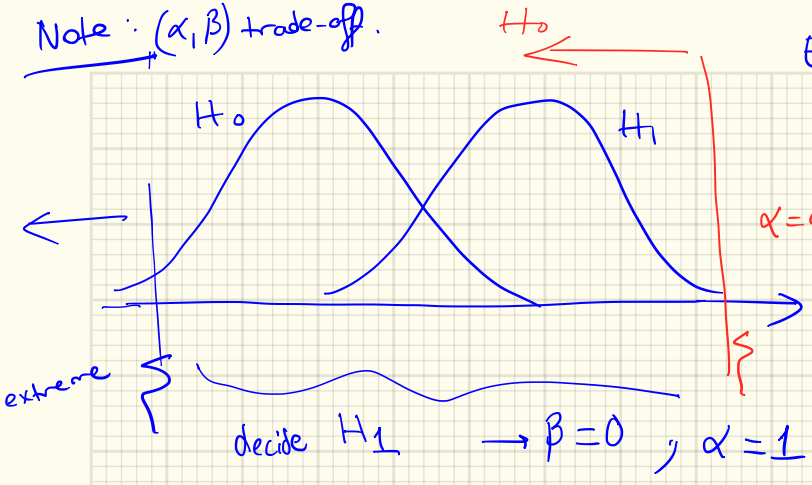
$P(\text{reject } H_0 ; H_0) = \alpha$

↑ prob. of false rejection.

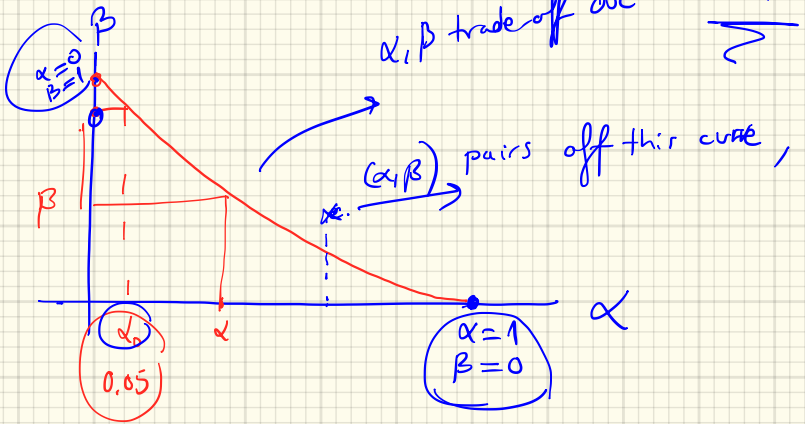


eg. set $\alpha = 0.05$ → 5 %

Note: (α, β) trade-off.



Extreme cases for α & β probabilities, moving ζ to the right & left

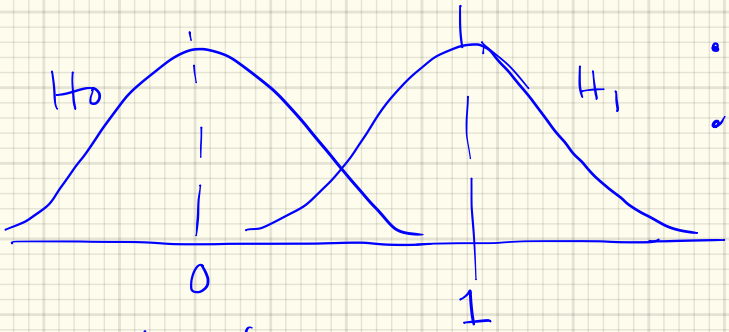


α, β trade-off due to LRT

(α, β) pairs off this curve, choosing them w/ other than LRT method.

Ex: Hypothesis Test on Normal Means:

- n data points (i.i.d.)
- You have 2 normal distributions w/ different means:



- $H_0: X_i \sim \mathcal{N}(0, 1)$
- $H_1: X_i \sim \mathcal{N}(1, 1)$

1) • Likelihood ratio test, Rejection region H_1

$$\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n}(X_i-1)^2/2\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_i X_i^2/2\right\}} > \xi$$

H_0

Do some algebra to simplify to : exercise →

1) LRT test: Reject H_0 if : $\sum_i X_i > \xi'$

a test "statistic":

↳ summarizes our measurements into a single number

$\left(\xi' = \log \xi + \frac{n}{2}\right)$

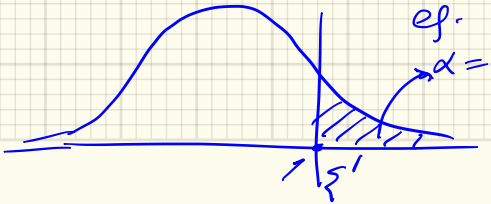
(If this statistics is large, it's evidence to reject H_0.)

2) How to choose ξ'? We set prob. of false rejection to a certain probability, eg. 5%.

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

→ Sum of X_i's have a Normal distrib. (bc X_i's have a normal distrib.)

∴ Use Normal tables



$\xi' = 1.96$.

if $\sum_i X_i > 1.96$; reject H_0.

< 1.96 ; Not reject H_0.

Ex: Test on Normal Variances:

- n data points, i.i.d:

$$H_0: \mathcal{N}(0,1)$$
$$H_1: \mathcal{N}(0,4)$$

} same mean but different variances

- Intuitively, If X's have wide-spread (variance) small " , choose H_1 choose H_0.

- LRT; Rejection region:

$$\frac{\text{density Under } H_1}{\text{density Under } H_0} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_i X_i^2/2.4\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_i X_i^2/2\right\}} > \}$$

↳ Do algebra to simplify to:

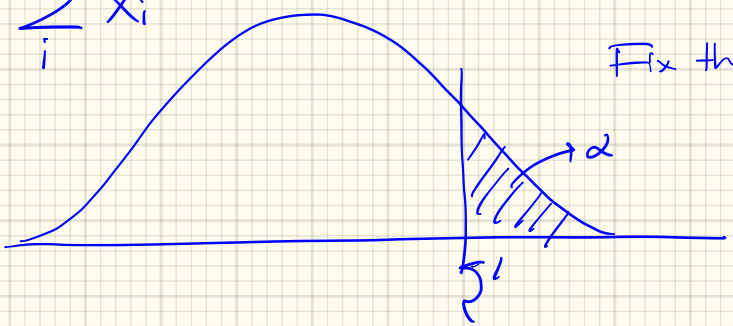
Reject H_0 if $\sum_i X_i^2 > \}'$

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 distribution)

Derived distrib.

$\sum_i X_i^2 \longrightarrow$ (Tables are available)



Fix the tail prob. of Chi-Squared distrib. eq. to 95th percentile

→ From χ^2-tables, read off ξ' that corresponds to 0.95.

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

Composite Hypothesis:

eg. you have a coin, you question: is it fair / unfair?

- You make n tosses
- → You get $S = 474$ heads in $n = 1000$ tosses ?

Is the coin fair?

~~binary hypothesis~~ (not $p = 0.5$ vs $p = 0.6$)

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

$p = 0.51, 0.52, 0.6, 0.7, 0.9$ many possible alternatives.

→ expected value: $\frac{n}{2}$: half heads, half tails) here : 500

(i) Pick a statistic : { H H T H T T T H . . . -

we did it → S = # Heads (statistic)

(ii) Pick shape of rejection region : $\left| S - \frac{n}{2} \right| > \{$ (500)

iii) Pick a Significance Level α: (eg. $\alpha = 0.05$)

iv) Pick a threshold ξ (critical value) s.t.

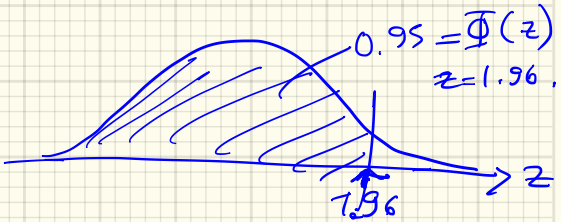
$$P(\text{reject } H_0 \,;\, H_0) = \alpha$$

$\alpha \approx$ prob of outliers.

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

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

Use normal tables



Recall.

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

$\text{Var}(S) = n \cdot \sigma^2$

$1000 \left(\frac{1}{4}\right) = 250$

$$S - 500 \leq 1.96\sqrt{250}$$

$\xi = 31$

$\rightarrow \quad |S-500| \leq 1.96\sqrt{250} \cong 31 \Rightarrow \xi = 31$

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

$\rightarrow$ Do Not Reject H_0 (at the 5% level)

$\nearrow$ 5% chance that the data we got is outlier.

Note: We tend to say

H_0 is **not rejected** rather than H_0: accepted.

$\equiv$

H_0 is the default hypothesis: we do not reject it until we see evidence contrary to it.

If we some evidence to the contrary, we reject H_0.

Ex: Is your die fair?

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

Null hypothesis → (fair die)

- N_i : observed occurences for each i.
- Roll your die n times; count $\#1$'s $\to N_1$, $\#2$'s $\to N_2$, $\vdots$, $\#6$'s $\to N_6$

Q: We observe N_i's; is my die fair?

or $\equiv$ is my pmf (above) valid?

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

- Choose form of rejection region:

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

• Choose ξ so that prob of false rejection 5%

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

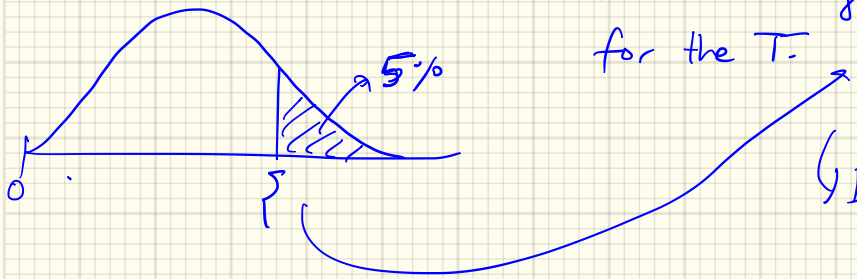
Specifically ↓

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

→ We need distribution of $T = \sum_i \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}$ ← derived distribution.

For large n, $T \sim$ a chi-squared distrib.

$P_T(t)$ under H_0.



For distrib. of T → there are tables for the T. Use them to find ξ.

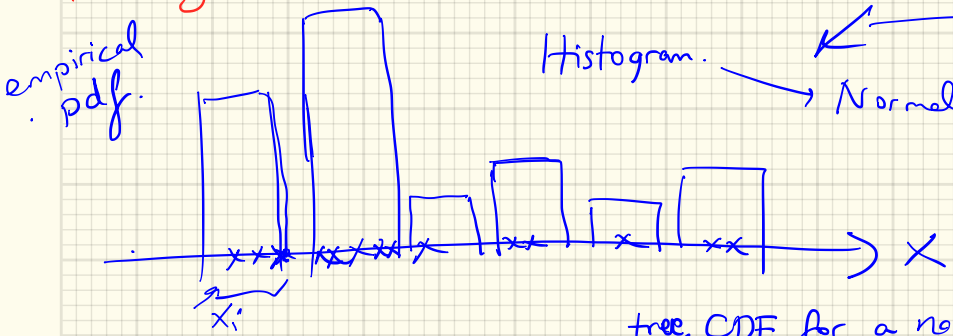
↳ Decide whether to Reject H_0 or Not Reject H_0.

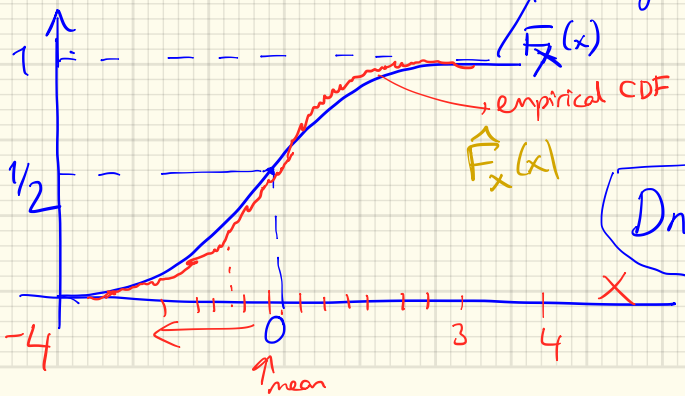
Now what happens

→ eg. When your data comes from a certain normal ? →
what could be a good statistic to use in a test :

Kolmogorov – Smirnov Test : Form CDF (empirical) F_X from your data.

Histogram. → Normalized Histogram ≈ pdf





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

A possible test :

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

If D_n is small → not reject H_0.

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

b/c this test is used frequently: people calculated distrib. of r.v. D_n.

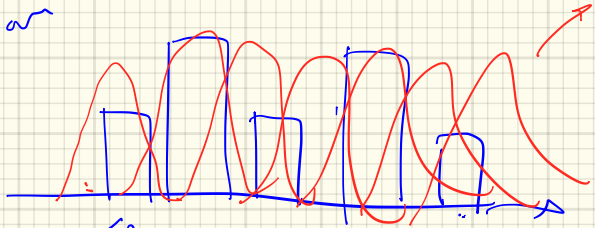
From tabulated prob. values of D_n:

$$\xi = \frac{1.36}{\sqrt{n}} \qquad n: \#\text{ data points}.$$

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

Note: **Kernel Density Estimation:** method to estimate unknown PDF from a histogram.

Histogram



$\Leftrightarrow$ Smoothed Histogram estimates thru weighted Gaussians.

$h(\cdot)$.

$\rightarrow$ Machine Learning / Data mining

Deep Learning $\Leftarrow$ Huge High-dim spaces of data + Huge # parameters