statistics

event

$$P(\text{soccer}) = \frac{\text{soccer}}{\text{total}} = \frac{3}{10} = 0.3$$

Sample space

Coin example I

an EXPERIMENT

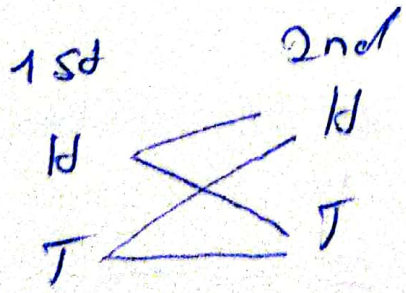
probability of landing on heads : P(heads)

" " " " tails : P(tails)

$$P(\text{heads}) = \frac{1}{2} = 0.5$$

---

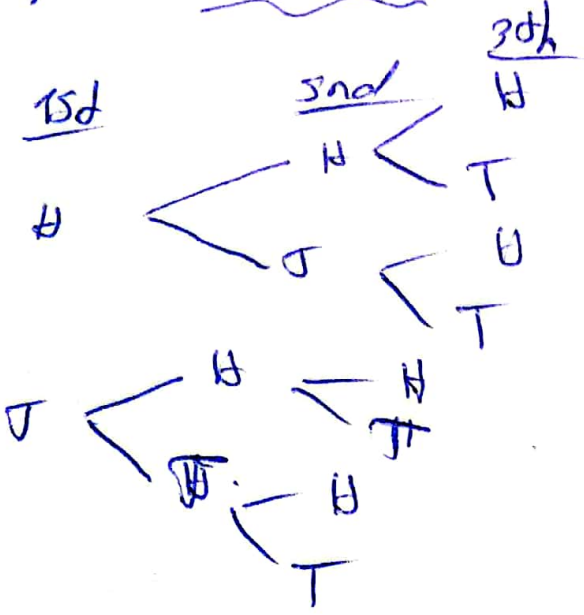
coin exp 2 (2 coins)



both heads = $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25$

P(HH)

~~P(HH)~~ 3 coins



H H H
H H T
H T H
H T T
T H H
T H T
T T H
T T T

$P(HHH) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 12.5\% = 0.125$

Dice examples

$P(6) = \frac{1}{6}$

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

---

complement of probability

$P(1 - \text{soccer}) = 1 - \frac{3}{10} = \frac{7}{10} = 0.7$

complement rule

$P(A') = 1 - P(A)$

$P(HHH) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

$P(HHH') = 1 - \frac{1}{8} = \frac{7}{8}$

$P(\text{not } 6) = 1 - \frac{1}{6} = \frac{5}{6}$

Sum of probability ~~(joint ev)~~

Soccer $P(S) = 0.3$

basketball $P(B) = 0.4$

$P(S) \text{ or } P(B) = 0.3 + 0.4 = 0.7$

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

$P(S \cup B) = P(S) + P(B)$ → joint events

Dice exp:
what is the probability of obtaining an even number or a 5?
even: 2, 4, 6

$P(5 \text{ or even}) = \frac{4}{6} = \frac{2}{3} \quad \frac{1}{6} + \frac{3}{6} = \frac{4}{6} = \frac{2}{3}$

2 Dice (U = union)

Sum of 7 or sum of 10

$1+6=7$
$2+5=7$
$3+4=7$
$4+3=7$
$5+2=7$
$6+1=7$

$4+6=10$
$5+5=10$
$6+4=10$

$$P(\text{sum } 7 \text{ or sum } 10) = \frac{6}{36} + \frac{3}{36} = \frac{9}{36} = \frac{1}{4}$$

difference of 2 or difference of 1

$6-4=2$ $4-6=2$
$5-3=2$ $3-5=2$
$3-1=2$
$3-1=2$
$4-2=2$
$2-4=2$

$6-5=1$ (2)
$5-4=1$ (2)
$4-3=1$ (2)
$3-2=1$ (2)
$2-1=1$

$$P(d1 \cup d2) = \frac{10}{36} + \frac{8}{36} = \frac{18}{36} = \frac{1}{2} = 0.5$$

(disjoint events)

$P(S) = 0.6$
$P(B) = 0.5$

$$P(S \cup B) = P(S) + P(B) - P(S \cap B)$$

$P(S) = 6$

$P(B) = 5$

$P(S \cap B) = 3$

total = 10

$$P(S \cup B) = P(S) + P(B) - P(S \cap B)$$

$$= \frac{6}{10} + \frac{5}{10} - \frac{3}{10} = \frac{8}{10} = \frac{4}{5} = 0.8$$

Dice

Sum 7 or diff 1

$$P(\text{Sum } 7 \text{ or diff } 1) = \frac{6}{36} + \frac{10}{36} - \frac{2}{36} = \frac{14}{36}$$

Independence

chess moves → dependence

coin → independence

Product rule (independent events)

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

---

coin exp:

what is the probability of landing on heads five times?

$$P(5 \times H) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$$

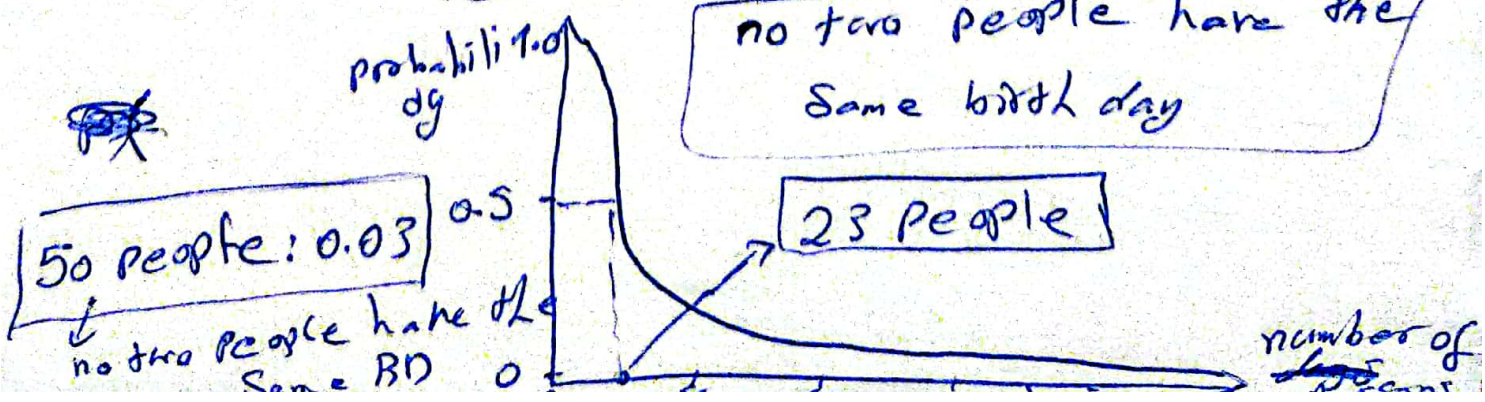
Dice exp:

2 Dice both 6:

$$P(2 \times 6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

$$P(10 \times 6) = \left(\frac{1}{6}\right)^{10}$$

Birthday Problem



product rule (Independent events)

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

product rule (dependent events)

2 dices first 6 and sum 10

$$P(6 \cap \text{sum}10) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

(6) (4)

$$P(\text{sum}=10 \mid 1\text{st } 6) = P(1\text{st } 6) \cdot P(\text{sum}=10 \mid 1\text{st } 6)$$

$$P(A \cap B) = P(A) \cdot P(B|A)$$

General product rule

~~school is~~

Total = 100

S = 60

room1 = 50
room2 = 30

$P(S) = 0.4$

$P(\text{not } S) = 0.6$

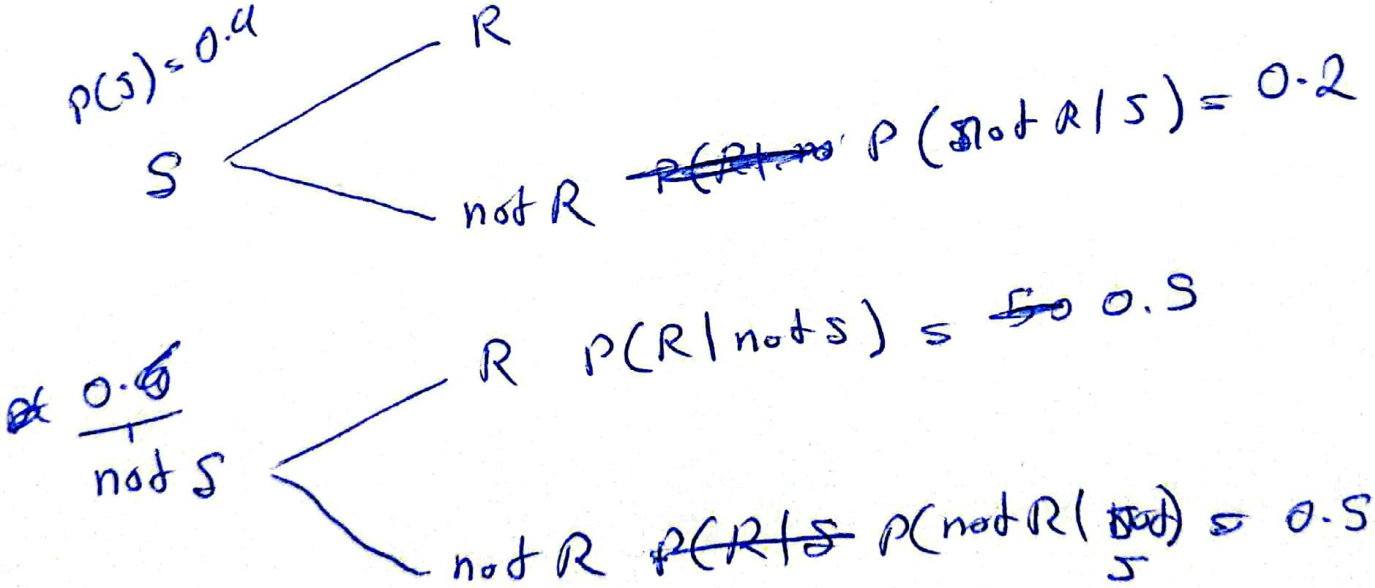
$P(R|S) = 0.8$

$P(S \cap R) = P(S) \cdot P(R|S) = 0.4 \times 0.8 = 0.32$

$P(\text{not } S \cap R) = P(\text{not } S) \cdot P(R|\text{not } S)$

$= 0.6 \times 0.5 = 0.30$

$P(R|S) = 0.8$



Bayes Theorem

$$P(\text{sick} \mid \text{diagnosed sick}) = \frac{\text{sick and diagnosed sick}}{\text{sick and diagnosed sick} + \text{healthy and diagnosed sick.}}$$

$P(\text{sick}) = 0.01\%$ sick = A

$P(\text{not sick}) = 99.99\%$ diagnosed sick = B

$P(\text{diagnosed sick} | \text{sick}) = 99\%$

$P(\text{diagnosed sick} | \text{not sick}) = 1\%$

$$P(A|B) = \frac{P(\text{sick}) \cdot P(\text{diagnosed sick} | \text{sick})}{P(\text{sick}) \cdot P(\text{diagnosed sick} | \text{sick}) + P(\text{not sick}) \cdot P(\text{diagnosed sick} | \text{not sick})}$$

$$P(A|B) = \frac{0.01 \times 0.99}{0.01 \times 0.99 + 0.9999 \times 0.1} = 0.0098$$

$P(A) = 0.01\%$

$P(A') = 99.99\%$

$P(B|A) = 99\%$

$P(B|A') = 1\%$

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}$$

prior	event	Posterior
$P(A)$	E	$P(A\|E)$

Naive assumption (in Bayes theorem)

$$P(A \mid w_1, \ldots, w_n) = \frac{P(A) \cdot P(w_1 \mid A) \cdots P(w_n \mid A)}{P(A) \cdot P(w_1 \mid A) \cdots P(w_n \mid A) + P(A') \cdot P(w_1 \mid A') \cdots P(w_n \mid A')}$$

Exps:

1) Image recognition:

$P(\text{Cat} \mid \text{image}) = P(\text{Cat} \mid \text{Pixel}_1, \text{Pixel}_2, \ldots, \text{Pixel}_n)$

2) classification in medicine

$P(\text{healthy} \mid \text{symptomes and history})$

Random Variables

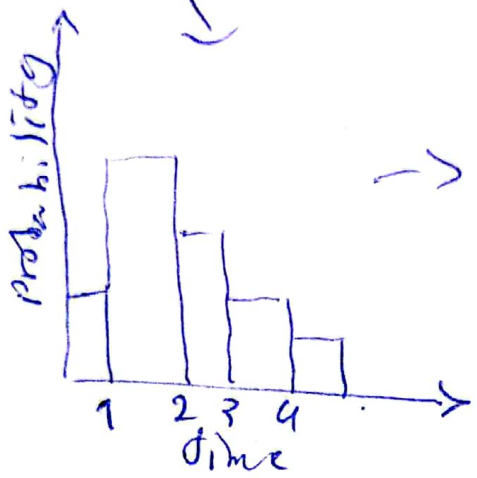
Exp, flip a coin $\begin{cases} H, P(H) = 0.5 \\ T, P(T) = 0.5 \end{cases}$

if $H \rightarrow X = 1$
if not $H \rightarrow X = 0$

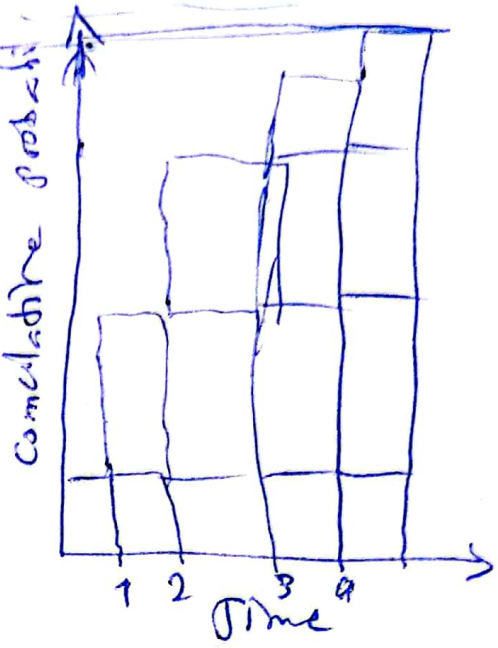
X is a random variable and in this case can be 1 or 0

* Cumulative distribution

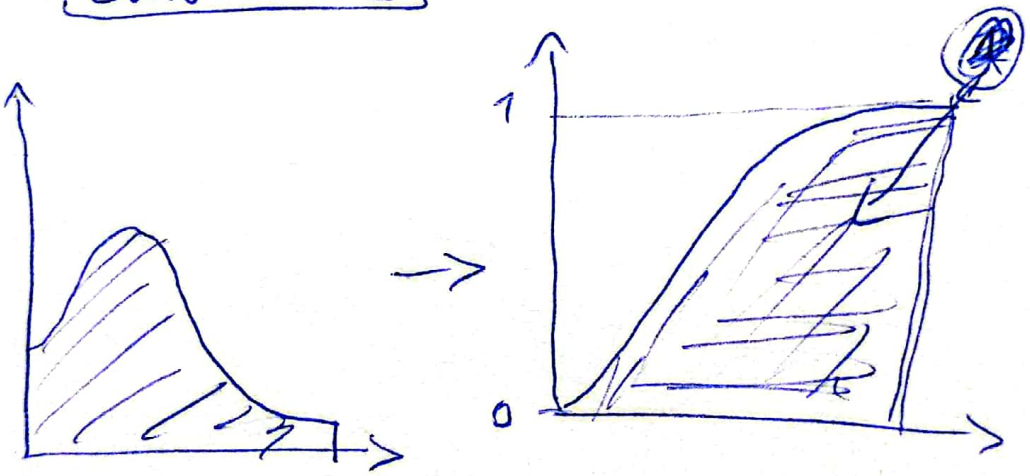
Discrete

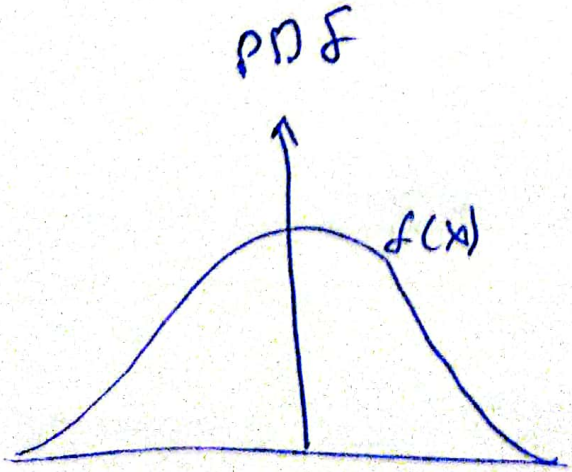


①

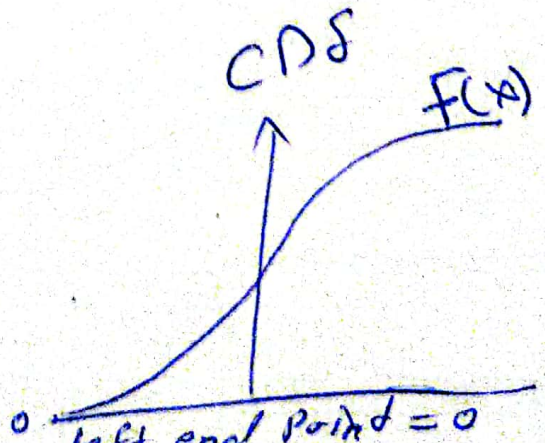


Continuous





area = 1
always positive



left end point = 0
right end point = 1
always positive and increasing

Binomial distribution

Binomial coefficient $= \binom{n}{k}$ counts all the combinations for landing k heads in n coin tosses.

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}$$

~~PDF~~

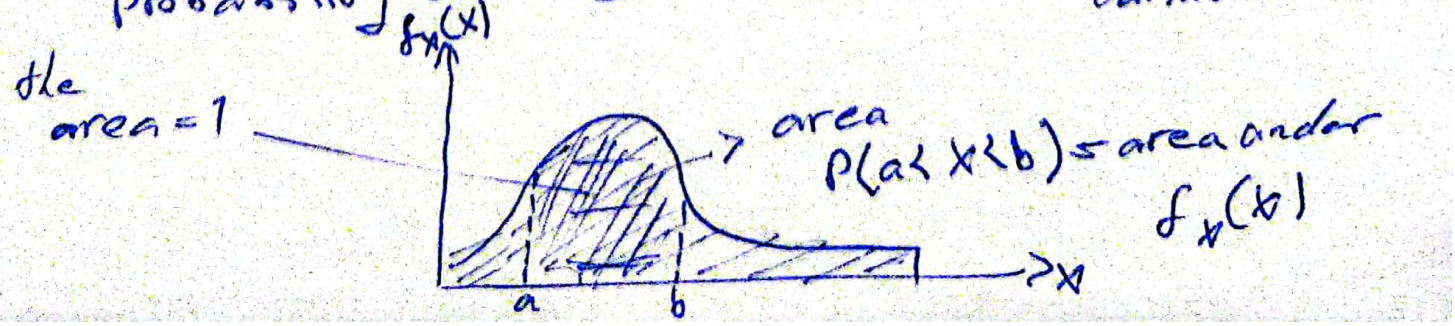
probability mass function → discrete variables

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

$X \sim \text{Binomial}(n, p)$

n and p are called the parameters of the binomial distribution.

probability density function (PDF) → continuous variables



Random variables
- Discrete (can take only a countable number of values)
- continuous (take values on an interval)

Random variables:

~~X=2~~

X = number of defective items in a shipment

↙

uncertain outcome

Deterministic Variables

X = 2 if f(x) = x²

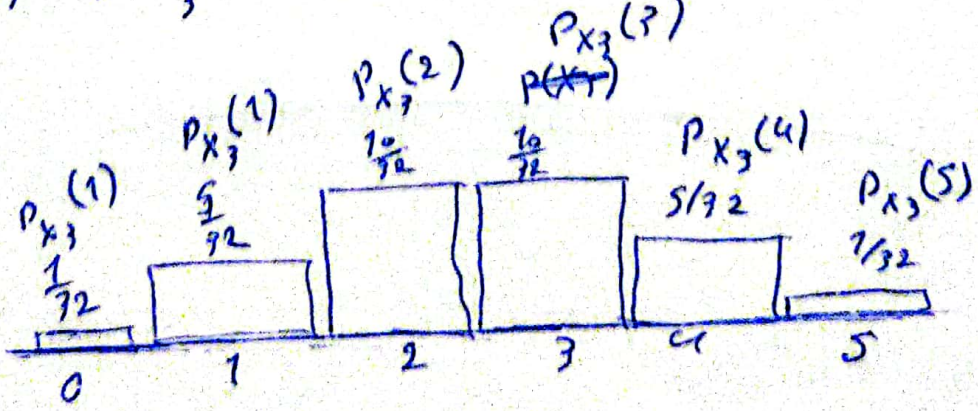
↙

fixed outcome

---

probability mass function (PMF)

$$P_X(x) = P(X = x), \quad x \in 0, 1, 2, 3, 4, 5$$



Binomial Distribution with very large n = normal distribution.

normal (Gaussian) distribution $= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

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

X is normal distributed

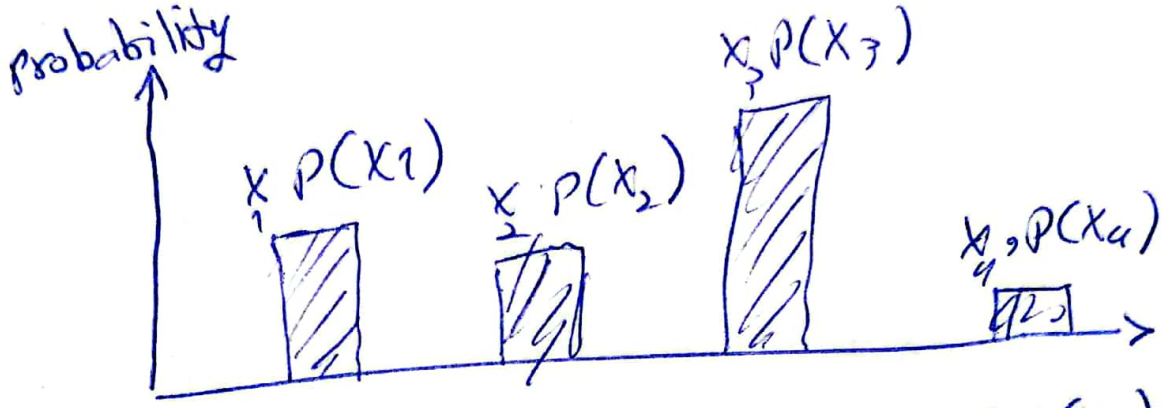
Standardization $\Rightarrow$ $Z = \frac{X-\mu}{\sigma}$

"Joint Distribution" (Discrete)

$P_{XY}(x,y) = P(X=x, Y=y)$

"Expected Value of a function"

$$E[X] = x_1 P(x_1) + x_2 P(x_2) + x_3 P(x_3) + x_4 P(x_4)$$



$$E[f(x)] = f(x_1)P(x_1) + f(x_2)P(x_2) + f(x_3)P(x_3) + f(x_4)P(x_4)$$

Exp: $E[x^2] = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}$ ~~$= E[x^2]$~~

~~$E[x] = E[x_i] + E[x_j]$~~

in general: $E[x_1 + x_2] = E[x_1] + E[x_2]$

$$E[\text{maddches}] = E[x_1] + E[x_2] + \dots + E[x_n] = 7$$

Variance formula:

$$Var(X) = E[(X-\mu)^2]$$
$$= E[X^2] - E[X]^2$$

Standard deviation:

$$std(X) = \sqrt{Var(X)}$$
$$= \sqrt{E[(X-\mu)^2]}$$
$$= \sqrt{E[X^2] - E[X]^2}$$

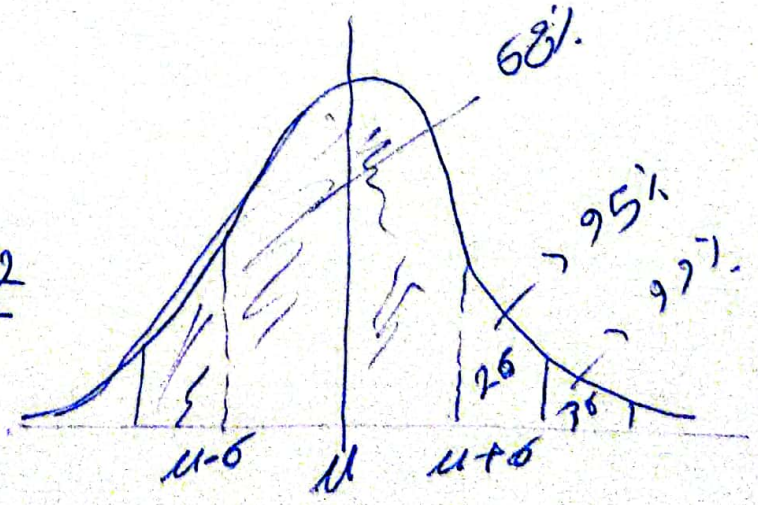
Normal distribution

μ = center of the bell

σ = spread of the bell

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

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



"Sum of Gaussians"

Sum of two or more Gaussian distributions is also Gaussian distributed.

$$R = (T + L) \sim N(\mu_T + \mu_L, \sigma_T + \sigma_L)$$

In General: $W = aX + bY$

independent $\begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim N(\mu_Y, \sigma_Y^2) \end{cases}$

$$W \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2\right)$$

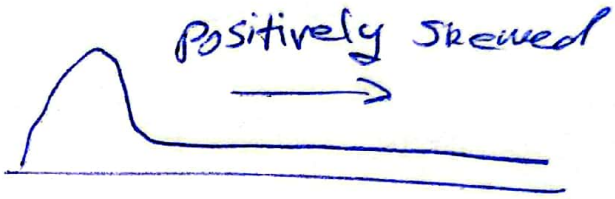
"Moments of a distribution"

$$E[X] = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$$
$$E[X^2] = p_1 x_1^2 + p_2 x_2^2 + \cdots + p_n x_n^2$$
$$E[X^3] = p_1 x_1^3 + p_2 x_2^3 + \cdots + p_n x_n^3$$
$$\cdots$$
$$E[X^n] = p_1 x_n^n + p_2 x_n^n + \cdots + p_n x_n^n$$

Skewness

$$\text{Skewness} = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] > 0$$

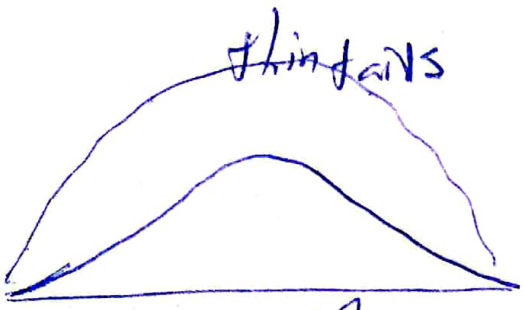


$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = 0$$

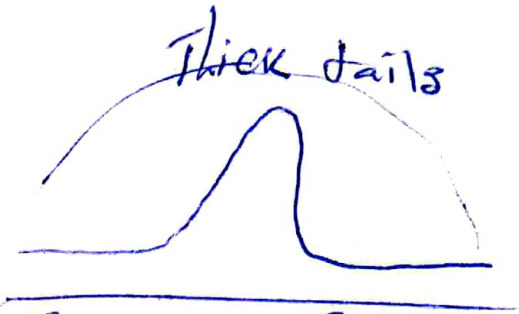
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] < 0$$

kurtosis

~~$E[X^4]$~~ $\text{kurtosis} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$



$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \text{small}$$



$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \text{ large}$$

↓

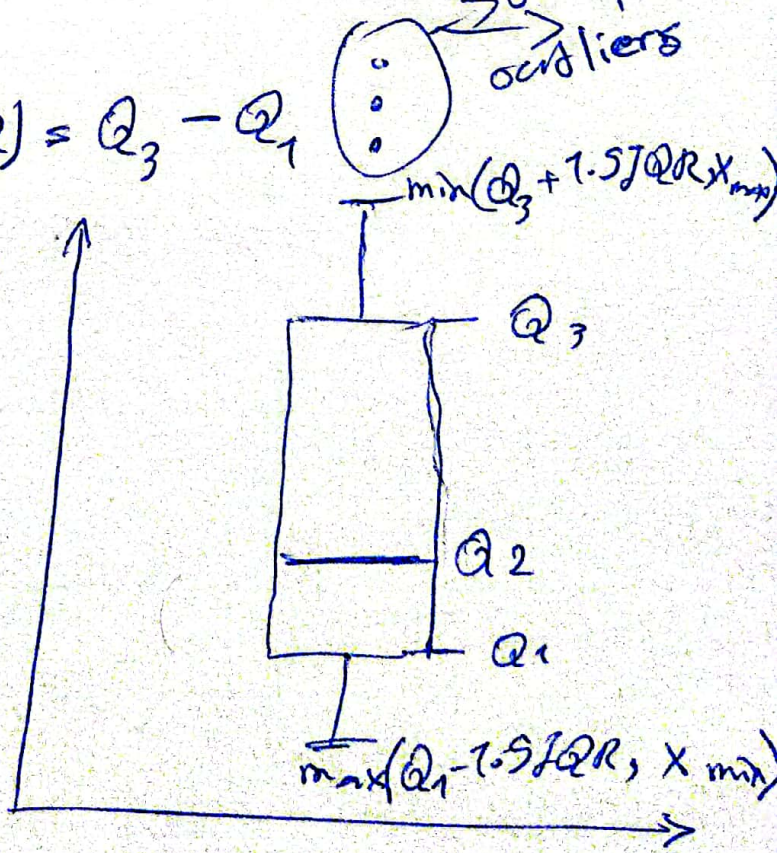
kurtosis

Box-plot

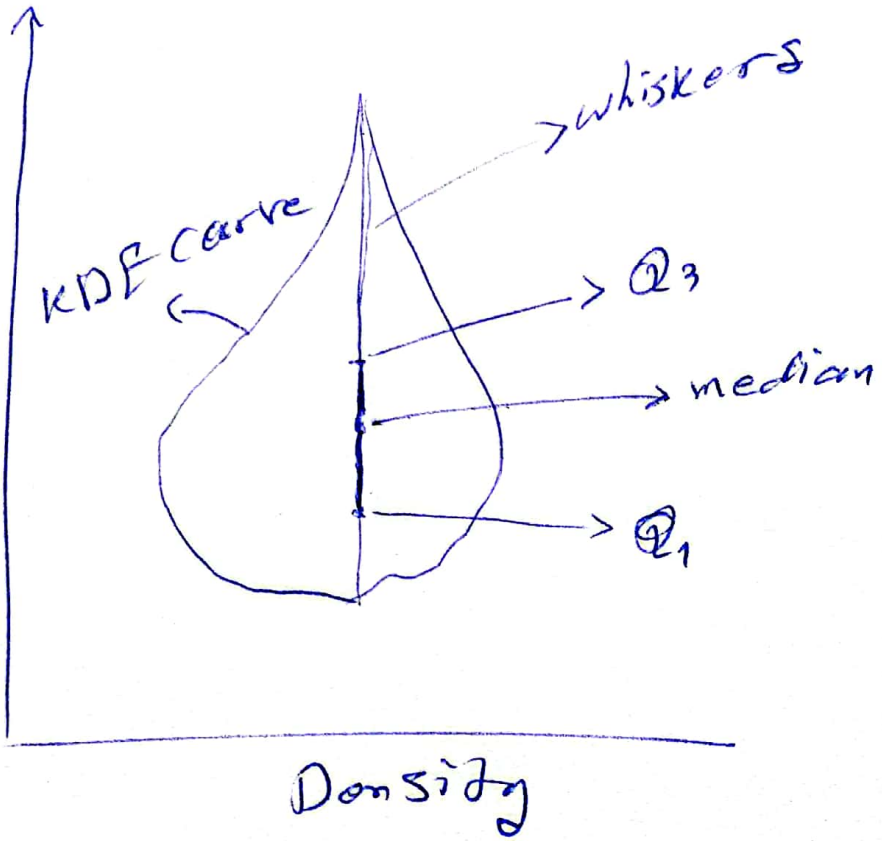
$Q_1 = q_{0.25}$ $Q_2 = q_{0.5}$ $Q_3 = q_{0.75}$

Interquartile range $(IQR) = Q_3 - Q_1$

$x_{min} = ?$ $x_{max} = ?$

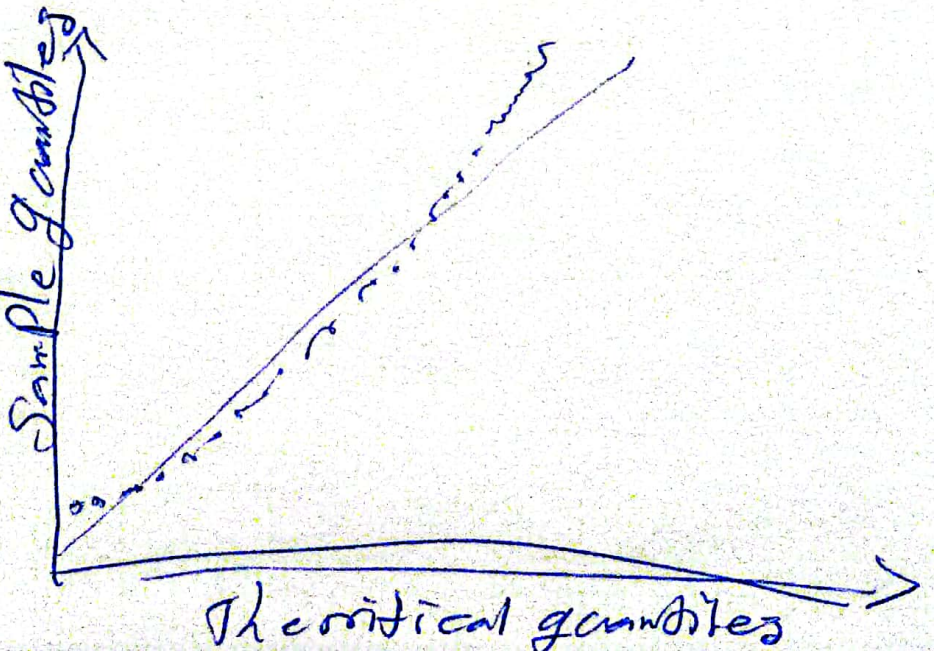


Violin-Plot



QQ Plots

check normality (Gaussian distribution)



"joint Distribution" (discrete)

$$P_{xy}(x,y) = P(X=x, Y=y)$$

for dependent discrete variables

EXP:

Age (Y) n=20

Weight (X) n=20	45	46	47	48
7	2/10	2/10	0	0
8	0	0	2/10	1/10
9	0	0	0	3/10

(all combination of probabilities)

(discrete joint)

$$P_{xy}(x,y) = P(X=x, Y=y) = P(X) \cdot P(y)$$

Independent random variables

discrete

marginal distribution

$$P_Y(y_i) = \sum_i P_{XY}(x_i, y_j)$$

$$P_X(x_i) \sum_j P_{XY}(x_i, y_j)$$

---

discrete conditional distribution

~~$P_{Y|X=i}(y) = P(Y=y \mid X=i)$~~

$$P_{Y|X=x}(y) = \frac{P_{XY}(x,y)}{P_X(x)}$$

- $P_{XY}(x,y)$ → joint PDF of X and Y
- $P_X(x)$ → marginal distribution of X
- $P_{Y|X=x}(y)$ → conditional PDF of Y

Continuous conditional distribution

conditional PD of of Y ← joint PD of of X and y

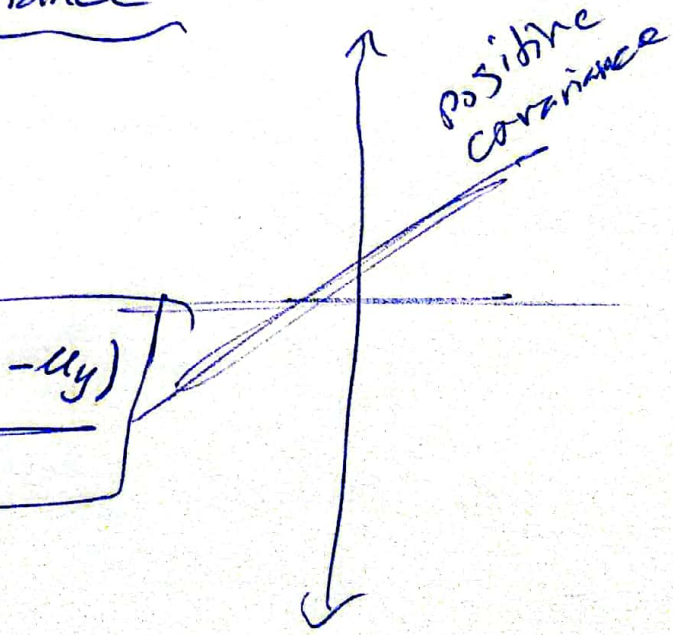
$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}$$

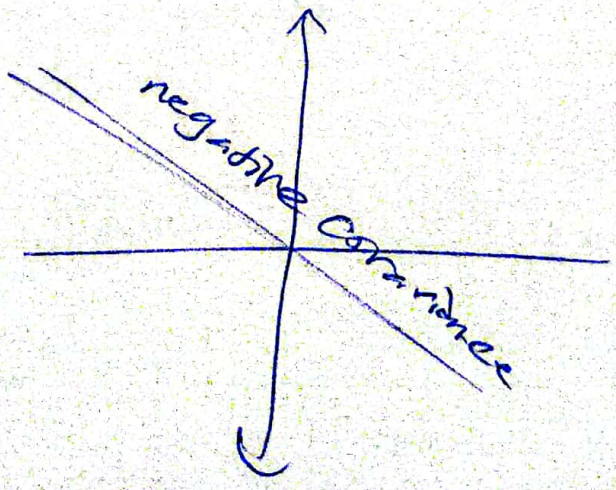
$f_X(x)$ ← marginal distribution of X

Covariance

$$cov(X,y) = \sum xy$$

$$cov(X,y) = \frac{\sum_{i=0}^{n} (x_i - \mu_x)(y_i - \mu_y)}{n}$$





Covariance of a probability distribution

$$Var(X_\theta) = \sum_{i=1}^{N} (x_i - a_x)^2 \cdot P(x_i)$$

$$cov(x,y) = \sum P_{xy}(x_i, y_i)(x_i - a_x)(y_i - a_y)$$

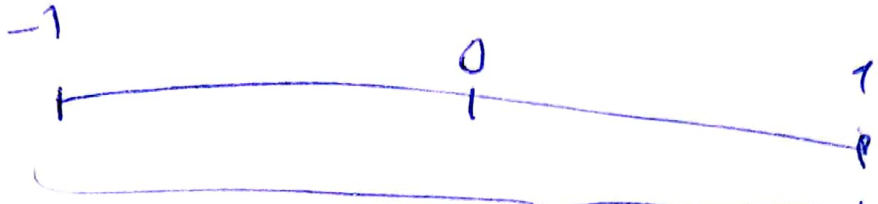
$$= E[XY] - E[X]E[Y]$$

Covariance matrix

	X	Y
X	Var(X)	cov(X,y)
Y	cov(X,y)	Var(Y)

Σ = Sigma or covariance matrix

"Correlation coefficient"



range of correlation coefficient

$$\text{Correlation coefficient} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

it is actually the standardized covariance.

or

$$\text{corr\_coeff} = \frac{cov(X,Y)}{\sqrt{var(X)} \cdot \sqrt{var(Y)}}$$

multivariate Gaussian Distribution

for a single variable

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

if W, H were independent

$$f_{HW}(h,w) = f_H(h) f_W(w)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_H} e^{\frac{-1}{2}\frac{(h-\mu_H)^2}{\sigma_H^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_W} e^{\frac{-1}{2}\frac{(w-\mu_W)^2}{\sigma_W^2}}$$

$$= \frac{1}{2\pi\sigma_H\sigma_W} e^{-\frac{1}{2}\left(\frac{(h-\mu_H)^2}{\sigma_H^2} + \frac{(w-\mu_W)^2}{\sigma_W^2}\right)}$$

$$= \frac{1}{2\pi \det \Sigma^{\frac{1}{2}}} \exp\left(\frac{-1}{2}\left(\begin{bmatrix} h \\ w \end{bmatrix} - \mu\right)^T \Sigma^{-1} \left(\begin{bmatrix} h \\ w \end{bmatrix} - \mu\right)\right)$$

$$f_X(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$|\Sigma|$ → determinant of the cov matrix

covariance matrix spread of the bell

mean vector $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$

Every dataset you work with in machine learning is a sample (not a population)

Population size = N

Sample size = n

Population mean = μ

Sample mean = $\bar{X}$

Population proportion

$$P = \frac{\text{number of items with a given characteristic } (X)}{\text{population } (n)}$$

Sample proportion

$$\hat{p} = \frac{x \rightarrow \text{number of items in sample}}{n \rightarrow \text{sample size}}$$

Sample Variance

Population Variance

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

N ← population size

μ ← population mean

Variance estimation

& $n = m$

Sample(X) → $Var(X) = \frac{\sum (X - \bar{X})^2}{n-1}$ → estimated variance

Sample Variance

$$Var(X) = \frac{1}{n-1} \sum (X - \bar{X})^2$$

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.

n = number of samples

X_i = some estimate X for a sample size i

as $n \longrightarrow \infty$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow E[X] = \mu_X$$

conditions:

1) Sample is randomly drawn

2) Sample size must be sufficiently large.

3) Independent observations.

Central limit theorem

As the number of observation (n) increases, the probability distribution of the sample becomes closer to a Gaussian distribution. even if the original variables (population) are not normally distributed.

$$\bar{X} \approx \mu = np$$

$$\sigma^2 \approx np(1-p)$$

$$\text{As } n \to \infty \quad \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]}{\sigma_X}\sqrt{n} \sim N(0,1^2)$$

$$\text{As } n \to \infty \quad \frac{\sum_{i=1}^{n} X_i - nE[X]}{\sqrt{n}\,\sigma_X} \sim N(0,1^2)$$

"Point estimation"

"maximum likelihood estimation" (MLE)

"find the model that most likely produced the data"

EXP:

model 1 — P(Data | model 1)
model 2 — P(data | model 2)
model 3 — P(data | model 3)

Data → maximize P(Data | model)

Maximum-Likelihood Bernoulli Example

Data: 8 heads 2 Tails — coin

Models: model 1: P(H) = 0.7, P(T) = 0.3 } model 2: P(H) = 0.5, P(T) = 0.5 } model 3: P(H) = 0.3, P(T) = 0.7

→ maximum P(model | Data)

an a better model

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

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

↑ log likelihood

we try to maximize the log likelihood, which can be interpreted as maximizing the likelihood.

$$\frac{d}{dP}\left(8\log(P) + 2\log(1-P)\right) = \frac{8}{P} + \frac{2}{1-P}(-1)$$

if $\frac{8}{P} + \frac{2}{1-P}(-1) = 0 \Rightarrow \hat{P} = \frac{8}{10}$

if n = coins

k = heads

$X_1 \; X_2 \cdots X_{n-1} \cdot X_n$

$X = (X_1, \ldots, X_n)$

$X_i \sim \text{Bernoulli}(p)$

likelihood

$$L(p; X) = P_p(X = x) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

if $x_i = 1$, $p^{[x_i]}(1-p)^{[1-x_i]} = p$

if $x_i = 0$, $p^{[x_i]}(1-p)^{[1-x_i]} = (1-p)$

$\sum_{i=1}^{n} x_i$ = # heads $\qquad n - \sum_{i=1}^{n} x_i$ = # tails

So:

$$L(p; x) = P_p(X = x) = \prod_{i=1}^{n} p_{x_i}(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\left(\sum_{i=1}^{n} x_i\right)} (1-p)^{\left(n - \sum_{i=1}^{n} x_i\right)}$$

likelihood.

$$L(p;x) = \log\left(p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i}\right)$$

$$= \left(\sum_{i=1}^{n} x_i\right) \cdot \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \log(1-p)$$

find the maximum:

$$\frac{d}{dp} L(p;x) = \frac{d}{dp}\left(\left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \log(1-p)\right)$$

$$= \frac{\sum_{i=1}^{n} x_i}{p} + \frac{n - \sum_{i=1}^{n} x_i}{1-p}(-1) = 0$$

$$\Rightarrow \boxed{\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}} \rightarrow \text{mean of the population}$$

---

maximum likelihood: Gaussian Example:

The best model ~~of~~ is the mode where the mean and variance of the ~~population~~ is ~~the~~ equal to sample

mean and ~~of~~ variance of the population

Suppose $X = (X_1, X_2, \ldots, X_n)$ Gaussian distributed

Mean = μ

Variance = σ^2

So; $X_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$

$$L(\mu, \sigma; X) = \prod_{i=1}^{n} f_{X_i}(x_i) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}$$

find the values of μ and σ that maximize the likelihood $L(\mu, \sigma; X)$

$$C(\mu, \sigma) = \log(L(\mu, \sigma; X))$$

$$L(\mu, \sigma) = \log\left(\frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}\right)$$

$$= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$$

Partial derivative of ~~and~~ log likelihood with respect to μ and σ

$\frac{\partial}{\partial \mu}$

$$\frac{d}{d\mu} L(\mu,\sigma) = \frac{-1}{2} \frac{\sum_{i=1}^{n} 2(x_i-\mu)}{\sigma^2}(-1)$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

$$\frac{d}{d\sigma} L(\mu,\sigma) = -\frac{n}{\sigma} - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)(-2)\frac{1}{\sigma^3}$$

$$= -\frac{n}{\sigma} + \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^3}$$

equating the partial derivatives to 0

$$\frac{d}{d\mu} L(\mu,\sigma) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

since $\sigma > 0$ $\quad \sum_{i=1}^{n} x_i - n\mu = 0$

so: $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ ← sample mean

MLE for mean

$$\frac{d}{d\sigma} L(\mu,\sigma) = -\frac{n}{\sigma} + \left(\sum_{i=1}^{n} (x_i - \mu)^2\right) \frac{1}{\sigma^3} = 0$$

since $\sigma > 0$:

$$\frac{d}{d\sigma} L(\mu,\sigma) = -n + \left(\sum_{i=1}^{n} (x_i - \mu)^2\right) \frac{1}{\sigma^2} = 0$$

$$\boxed{\hat{\mu} = \bar{X}} \rightarrow$$

$$\frac{d}{d\sigma} L(\mu,\sigma) = -n + \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right) \frac{1}{\sigma^2} = 0$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

so:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

MME for the standard deviation

Polynomial regression

this model has low loss but it is not the appropriate one.

the best loss but it is not the appropriate one.

Loss	Equations	Penalty
10	$y = cx + 3$	$L_2 = 4^2 = 16$
2	$y = 2x^2 - 4x + 5$	$L_2 = 2^2 + (-4)^2 = 20$
0.1	$y = cx^{10} + 9x^8 - 2x^6 + 3x^4 - 8x^2 - 10x + 4$	$L_2 = 4^2 + (-9)^2 + 3^2 + (-8)^2 + 5^2 + (-10)^2 = 246$

(38)

near loss
= ~~abc~~ loss + Penalty

~~abc~~

15 + 10 = 20

2 + 20 = 22

0.1 + 246 = 246.1

→ Regularized error: $LL + \lambda(a_n^2 + a_{n-1}^2 + \cdots + a_1^2)$

→ best model to describe the system

Regularization term in general

model: $y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 n + a_0$

log_loss = LL

L_2 Regularization ~~term~~ Error: $a_n^2 + a_{n-1}^2 + \cdots + a_1^2$

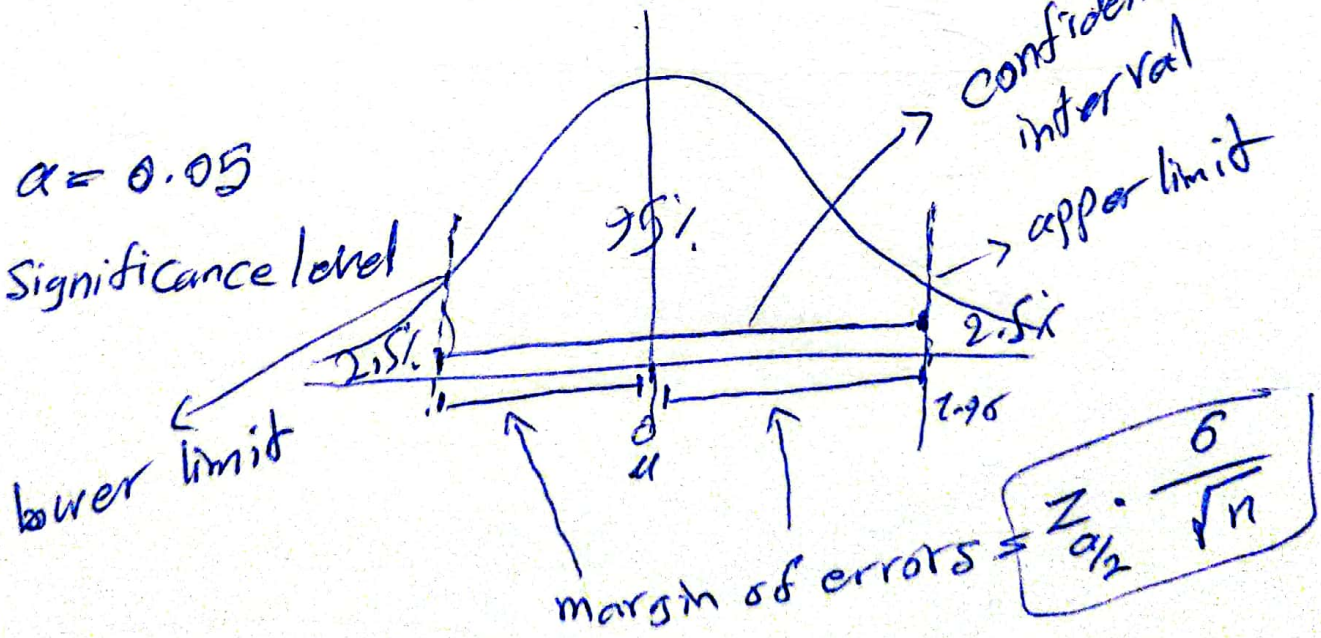
Regularization Parameter: λ → a small number

Statistical Inference

- Bayesian
- frequentist (maximum likelihood...)

Z-distribution

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



if $\mu = 0$ and $\sigma = 1$ ($X \sim N(0,1)$)

then $I_{\alpha/2} = \boxed{1.96}$ → critical value

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

confidence interval calculation

1) find the sample mean

→ pop size

N = 6000 adults

→ sample std

$\begin{cases} n = 49 \\ \bar{x} = 170 \text{ cm} \end{cases}$ $\sigma = 25$ cm

2) find a desired confidence level (1−a)

find 95% confidence interval

3) get the critical value ($I_{a/2}$)

for 95% → $I_{a/2} = 1.96$

4) find the Standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$

$\frac{\sigma}{\sqrt{n}} = \frac{25 \text{ cm}}{\sqrt{49}} = \frac{25}{7}$

5) find the margin of error

$I_{a/2} \cdot \frac{a}{\sqrt{n}} = 1.96 \times \frac{25}{7} = 7$ ← margin of error

6) Add/subtract the margin of error to the Sample mean.

confidence interval = $\bar{X} \pm I_{a/2} \cdot \frac{a}{\sqrt{n}}$

confidence interval = 170 cm ± 7 cm

So 163 cm, 177 cm

mean of the population

- The confidence interval contains the true population parameter approximately 95% of the time.

confidence interval - Student's t-distribution → for means

when σ known

$$\text{confidence interval} = \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

σ → pop-std

n → num sample

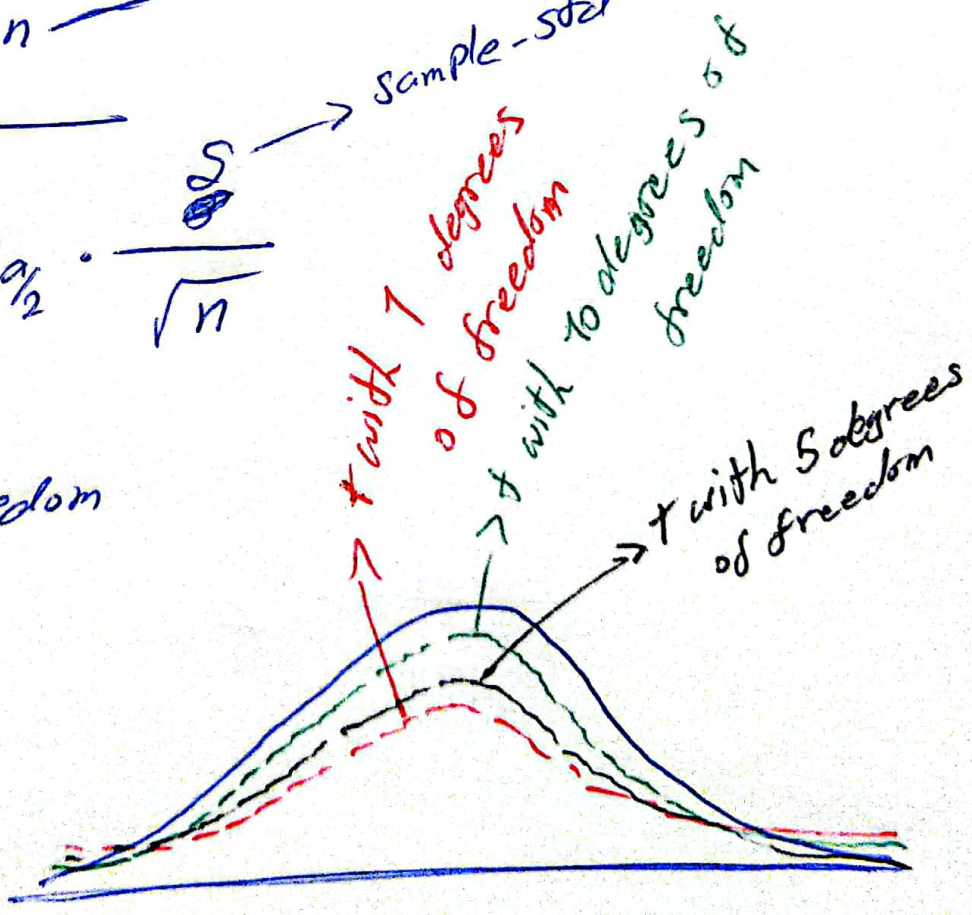
when σ unknown → t-Distribution

$$\text{confidence interval} = \bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

s → sample-std

degrees of freedom

$n - 1$



$X = 24$

$n = 30$

$\hat{p} = \frac{X}{n} = \frac{24}{30} = 80\%$

confidence interval for proportion $= \hat{p} \pm$ margin of error

margin of error $= I_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ← standard error

margin of error $= 1.96 \times \sqrt{\frac{0.8 \times 0.2}{30}}$

$= 0.14$

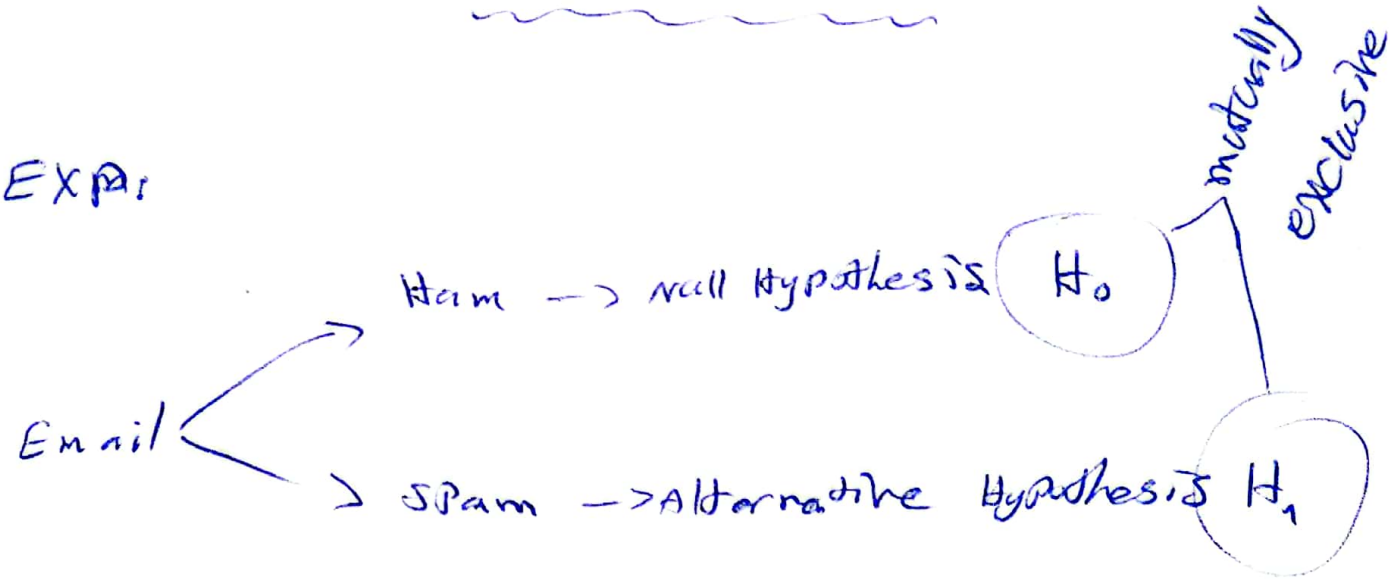
confidence interval $= 0.8 \pm 0.14$

$0.66 < p < 0.94$

$66\% < p < 94\%$

Hypothesis Testing

Exp:



Email is assumed to be Ham

Type I and Type II errors

Exp:

false positive

if you send a regular email to spam Box: Type ~~one~~ I error

if you send a spam Email to the regular Box: type II

↓

false negative

decision	H_0 True	H_0 false
reject H_0	Type I error	correct
Don't reject H_0	correct	Type II error

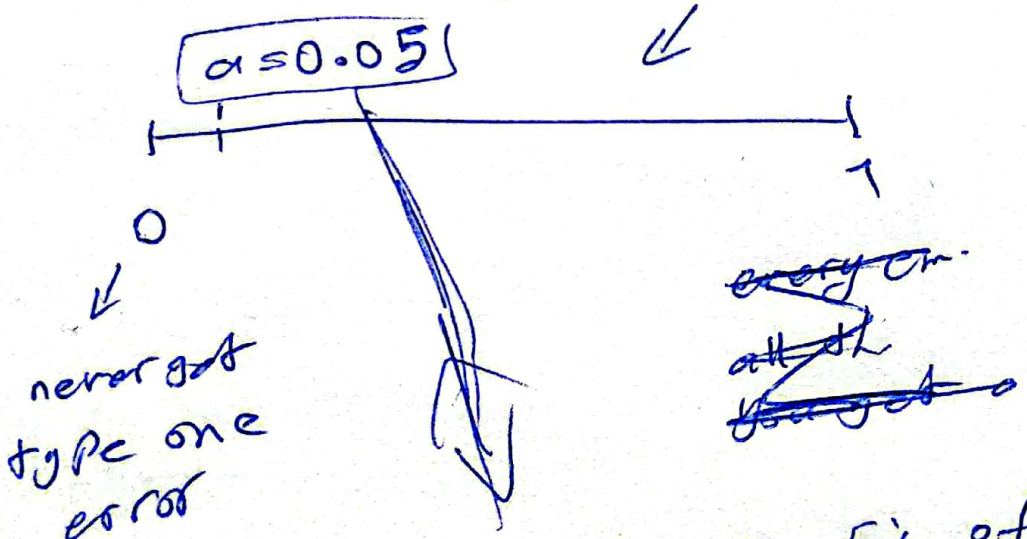
Type one error is more Dangerous

Ex: Sending a regular email to spam is worse than not sending a spam email to the regular inbox.

- what is the greatest probability of type I error are willing to tolerate?

↓

significance level (α)



a Ham email is spam 5% of times

Right- left- and two-tailed test

Right-Tailed Test ⟶ $H_0: \mu = 66$ vs. $H_1: \mu > 66$

Left-Tailed Test ⟶ $H_0: \mu = 66$ vs. $H_1: \mu < 66$

Two-tailed Test ⟶ $H_0: \mu = 66$ vs. $H_1: \mu \neq 66$

P-value

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.

if P-value $< a$ reject the H_0 (and accept H_1)

if P-value $> a$ do not reject H_0

$T(X)$: test statistic T: observed statistic

$H_0 : \mu = \mu_0$