Data Analysis

(1

1 - Review on Prob. Theory and simulation

Def1: a random variable is just a mapping
$X : \Omega \longrightarrow \mathbb{R}$ where Ω is a probability space

$(\Omega, \mathcal{F}, \mathbb{P})$

Ω: sample space; $\mathcal{F}$: events; $\mathbb{P}$: probability

Remarks:

1. X is said to be discrete if its range $\# X(\Omega)$ is at most equal to $\mathbb{N} = \{0, 1, \dots\}$

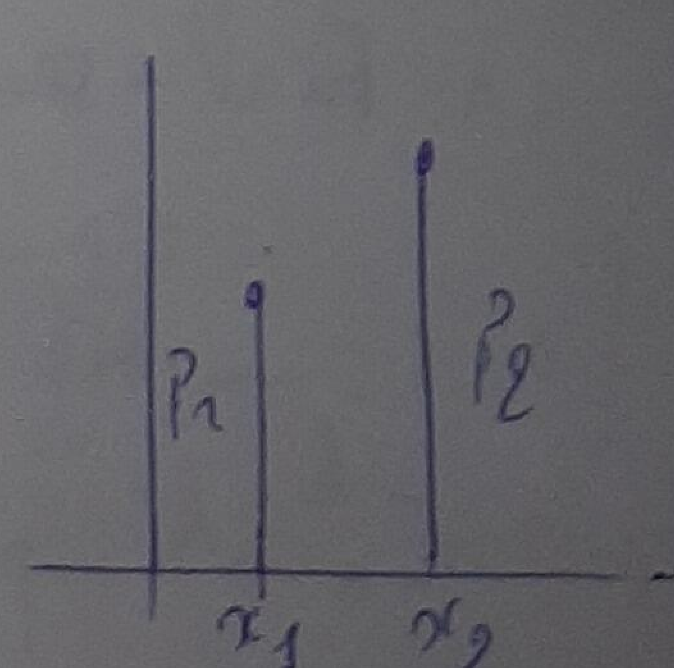
1-1. Any finite subset is enumerable

2. X is said to be continuous. if not

3. On associe à X v.a.d la loi de Prob
its "distribution function"
"mass"

$X(\Omega)$	P_X
x_1	$P_1 = \mathbb{P}\{X = x_1\}$
x_2	$P_2 = \mathbb{P}\{X = x_2\}$
⋮	⋮

$\{P_i\} \subset [0,1]$ "+" $\sum_{i \geq 1} P_i = 1$



Bar graph
diagrame à Batons

4. if X is continuous then $\exists$ a density function f_X such that:

+ $\int f_x = 1$

+ $\mathbb{P}\{a \leq X \leq b\} = \int_a^b f_x \, dx$

$(\Omega, \mathcal{F}, \mathbb{P})$ EA

$X : \Omega \longrightarrow \mathbb{R}$

les valeurs de X $\# X(\Omega) \leq \# \mathbb{N} \rightarrow$ discret

$\# X(\Omega) \simeq \# \mathbb{R} \rightarrow$ continue

2. Usual Distributions

2-1. Discrete Case :

Bernoulli Distribution :

- Let X be a d.r.v. It follows a Bernoulli distribution if

$X(\Omega) = \{0, 1\}$

If $p = \mathbb{P}\{X = 1\}$ $\quad X \sim B(p)$

generic situation

$\mathcal{R} \in \rightsquigarrow (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow$ A: non trivial event

$0 < \mathbb{P}(A) < 1$

$X : \Omega \rightarrow \{0, 1\}$

$$w \rightarrow \begin{cases} 1 & \text{si } w \in A \\ 0 & \text{sinon} \end{cases}$$

2-1.2 Binomial distribution :

$X \sim B(n, p)$, n trial-sens p: success prob

if : i) $X(\Omega) = \{0, 1, \dots n\}$

ii) $\mathbb{P}\{X = k\} = C_n^k \, p^k (1-p)^{n-k}$ $\quad k = 0, 1, \dots n$

Generic situation:
repeat a Bernoulli experiment n times and let X be the number of occurrences of A.

Then $X \sim B(n,p)$
Let's check:

i) $X = 0, 1, \dots n$
ii) $\mathbb{P}\{X=k\} = \mathbb{P}(\underbrace{A \cap A \cap A \dots}_{k \text{ fois}} \cap \underbrace{\bar{A} \cap \dots \bar{A}}_{n-k}) = C_n^k \, p^k (1-p)^{n-k}$

2-1-3 Geometric distribution

$X \sim \mathcal{G}(p)$ if

i) $X = 1, 2, 3, \dots$
ii) $\mathbb{P}\{X=k\} = p(1-p)^{k-1}$

• Generic Situation
repeat a given random experiment infinitely many times. and let X be the ~~...~~ position of the first occurrence of A (non trivial event)
Let's check:
. $X = 1, 2, \dots$
$\mathbb{P}\{X=k\} = \mathbb{P}(\underbrace{\bar{A} \cap \bar{A} \cap \dots \bar{A}}_{k-1} \cap A \cap \Omega \dots) = p(1-p)^{k-1}$

2-1-4 Pascal distribution: negative Binomial

$X \sim NB(n,p)$, $p \in]0,1[$ $n \in \mathbb{N}^*$

if
$X(\Omega) = n, n+1, \dots$
$\mathbb{P}\{X=k\} = C_{k+n-1}^{k} \, p^n (1-p)^k$

· Generic situation:
. repeat a Bernoulli experiment infinitly many times.
. let X be the number of needed repetitions up to obtaining n times A (A a given non trivial event)

Let's check

i) $X = n, n+1, \dots$

ii) $\mathbb{P}\{X = k\} = \mathbb{P}(\underbrace{A \cap A \dots}_{\substack{n-1\ A \\ \text{among } k-1}} \cap \overset{\text{position } k}{A}) = C_{k-1}^{n-1} \cdot p^{n-1} (1-p)^{k-n} \cdot p$

§II - Continuous Random variables:

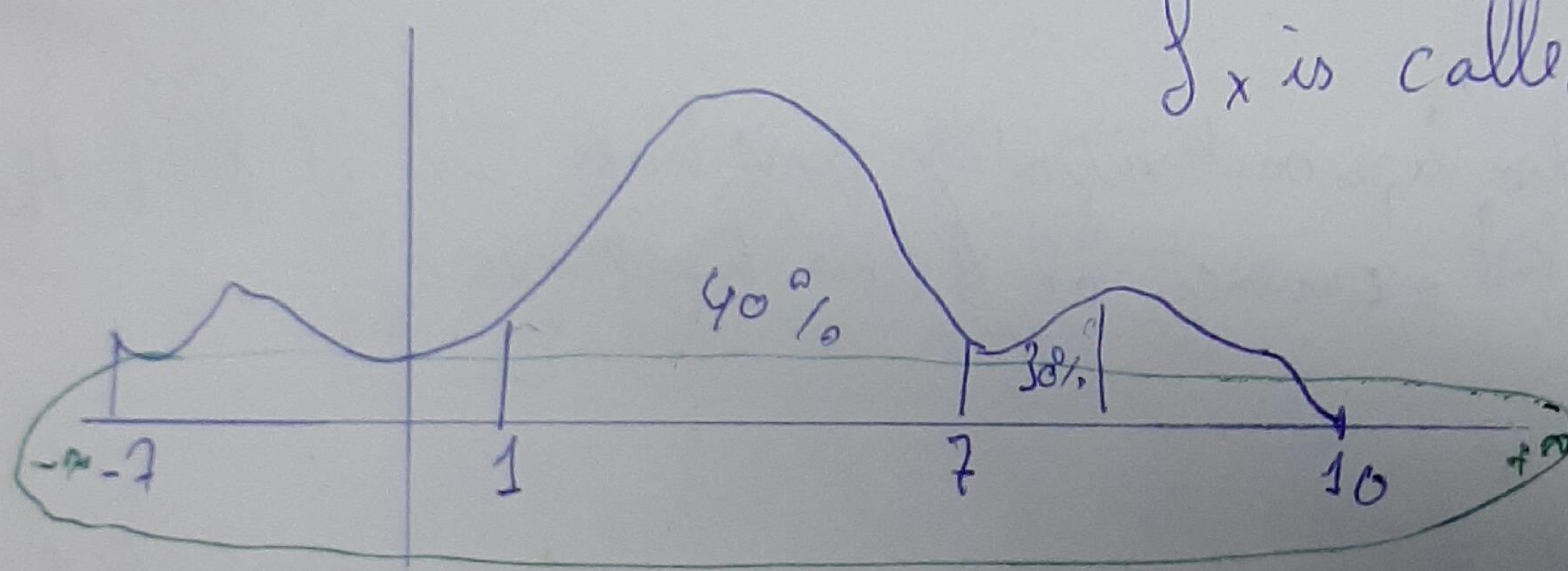
- Given X a random variable, it is said to be continuous iff (if & only if) there exists

$$f_X : \mathbb{R} \longrightarrow \mathbb{R}_+ \quad \text{such that}$$

i) $\int_{\mathbb{R}} f_X(x)\,dx = 1$

ii) $P\{a \le X \le b\} = \int_a^b f_X(x)\,dx \qquad -\infty \le a \le b \le +\infty$

f_X is called the probability density function



$X(\Omega)$

Notes:

1) - We associate to f_X (a given density) it's cumulative function

$$F_X(x) := \mathbb{P}\{X \le x\} = \int_{-\infty}^{x} f_X(t)\,dt$$

2) . $F_X(+\infty) = 1$, $F_X(-\infty) = 0$

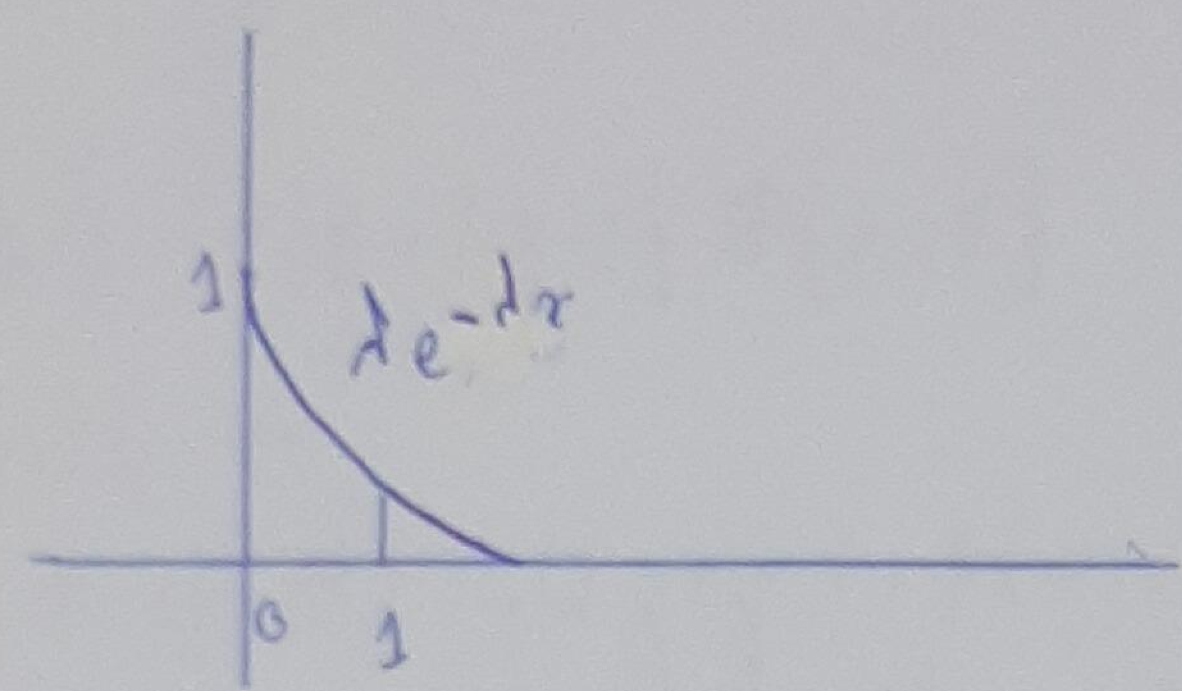
. F_X is continuous

. $f_X = \frac{d}{dx} F_X$ (outside discontinuous points of f_X)

§ Some usual densities :

§1- Exponential distribution

$\lambda > 0, \ f(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$ $\qquad \mathbb{1}_{\mathbb{R}_+} : \begin{cases} 1 & \text{if } x \in \mathbb{R}_+ \\ 0 & \text{if not} \end{cases}$



$$\mathbb{P}\{X \geq T\} = \int_T^{+\infty} e^{-\lambda x} d(-\lambda x) \underset{u = \lambda x}{=} \int_{-\lambda T}^{-\infty} e^u du = e^{-\lambda T} \underset{T \to +\infty}{\longrightarrow} 0$$

- $\lim\limits_{T \to +\infty} \mathbb{P}\{X \geq T\} = \mathbb{P}\{X = +\infty\}$
- Exponential distribution is memory-less

  $\hookrightarrow \mathbb{P}\{X \geq t + h \,/\, X \geq t\} = \mathbb{P}\{X \geq h\}$

" $e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$ "

i) $F_X(x) = (1 - e^{-\lambda x}) \mathbb{1}_{\mathbb{R}_+}(x)$

ii) $\mathbb{P}\{a \leq X \leq b\} = F_X(b) - F_X(a)$

§2- Normal distribution :

Given $m \in \mathbb{R}$ and $\sigma > 0$

X is following the normal distribution with mean m and variance σ^2 if :

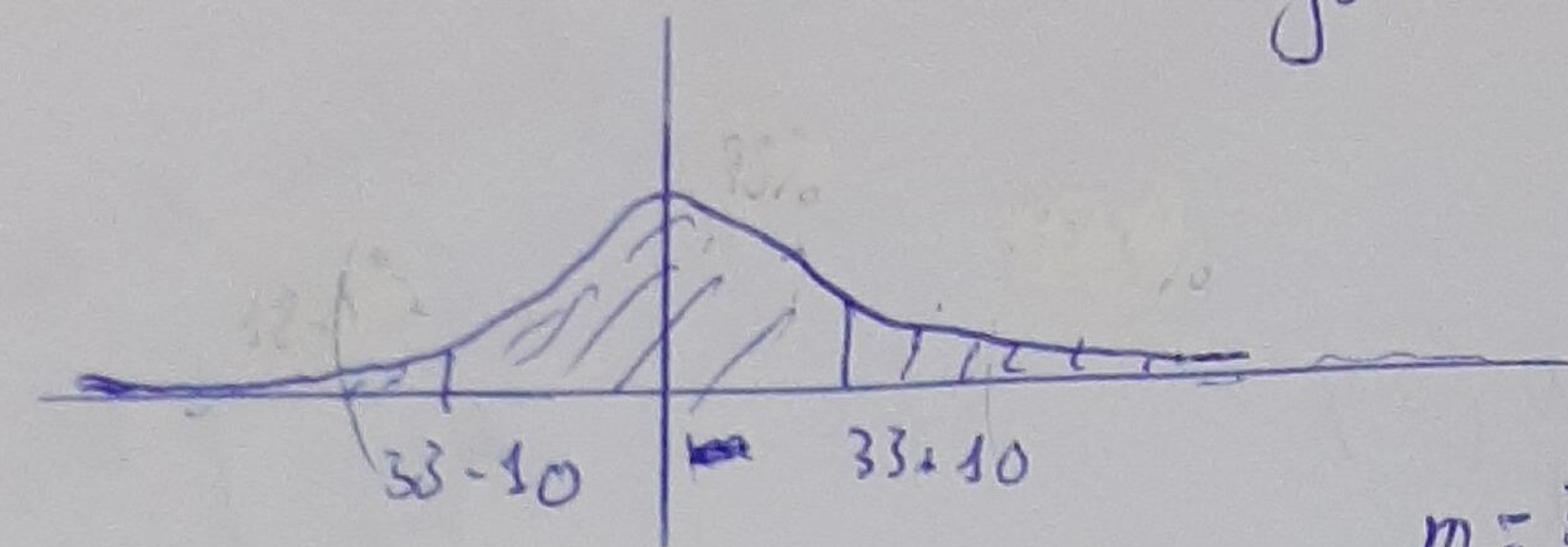
$$f_X = f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-m)^2}{2\sigma^2} \right)$$

We denote $X \sim N(m, \sigma^2)$

forme cloche / bell shape

Notes

i)



$m = 33$ (for example)

ii) $E(X) = m$

$V(X) = \sigma^2$

. Recall that

.) $A, B \in \mathcal{P}(U)$, $A \perp B$ (are "independent")

$\iff \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(B)\,\mathbb{P}(A)$

..) X, Y two random variables, $X \perp Y$ "are independent"

$\iff \mathbb{P}\{X = x_i, Y = y_j\} = \mathbb{P}\{X = x_i\}\,\mathbb{P}\{Y = y_j\}$

...) $\mathbb{P}(A/B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

::) $E(Y|X) = \varphi(X) \sim Y$

Def: Joint probability distribution of a given random couple (X, Y)

$\mathbb{P}_{XY} \in \mathcal{M}(\mathbb{R}^{(\#X)\times(\#Y)})$, $\#X = \#X(\Omega)$

$\mathbb{P}_{XY} = p_{ij} = \mathbb{P}\{X = x_i, Y = y_j\}$

$$\begin{array}{c|ccc|} & y_1 & y_2 & y_3 \\ \hline x_1 & p_{11} = \mathbb{P}\{X = x_1, Y = y_1\} & & \\ x_2 & & & \\ & & & \end{array}$$

- Two successive draws X uniformly from $\{1, \dots, 4\}$
- The second one Y uniformly from $\{1, \dots \cancel{}\ X\}$

.) Determine $\mathbb{P}_{XY}$?

$\mathbb{P}\{X=1, Y=1\} = \mathbb{P}\{Y=1 \mid X=1\}\, \mathbb{P}\{X=1\}$

$\mathbb{P}\{X=1, Y=2\} = 0$

$\mathbb{P}\{X=2, Y=1\} = \mathbb{P}\{Y=1 \mid X=2\} \cdot \mathbb{P}\{X=2\}$

$\mathbb{P}\{X < Y\} = 0$

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & \frac{1}{4} & 0 & 0 & 0 \\ 2 & \frac{1}{8} & \frac{1}{8} & 0 & 0 \\ 3 & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & 0 \\ 4 & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \end{array}$$

§ - Marginal Distributions

$$\begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & 0 & 0 \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & 0 \\ \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \end{bmatrix} \begin{array}{c} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{array}$$

$\cdot \quad \cdot \quad \cdot \quad \frac{1}{16}$

$$\boxed{\begin{aligned} \mathbb{P}_X(x_i) &= \mathbb{P}\{X = x_i\} \\ &= \sum_{j=1}^{\#Y} \mathbb{P}\{X = x_i, Y = y_j\} \end{aligned}}$$

Notes:

- If we know $\mathbb{P}_X$ and $\mathbb{P}_Y$ we couldn't form $\mathbb{P}_{XY}$!!

§?.1: The conditional Distributions:

Given x_i, $\mathbb{P}_{Y \mid X = x_i}$

Y	$\mathbb{P}_{Y \mid X = x_i}$
$y_1 = 1$	$\mathbb{P}\{Y=1 \mid X = x_i\}$
2	
3	
4	

$E(Y|X) = \ell(X)$ ℓ ?

Examples

Y	$\mathbb{P}_{Y|X=1}$
1	1

Y	$\mathbb{P}_{Y|X=2}$
1	1/2
2	1/2

$$\mathbb{P}_{Y|X=k} \sim \mathcal{U}\{1, \dots k\}$$

§ 2.2: The conditional Expectation

The conditional expectation of Y given X is the new r.v denoted $E(Y|X)$ and defined as follows.

$E(Y|X)$	$\mathbb{P}_X$
$E(Y|X=x_1)$	$\mathbb{P}_X(x_1)$
$E(Y|X=x_2)$	$\mathbb{P}_X(x_2)$
$\vdots$	$\vdots$

Example

$E(Y|X)$	$\mathbb{P}_X$
1	$\frac{1}{4}$
$\frac{3}{2}$	$\frac{1}{4}$
2	1/4
5/2	$\frac{1}{4}$

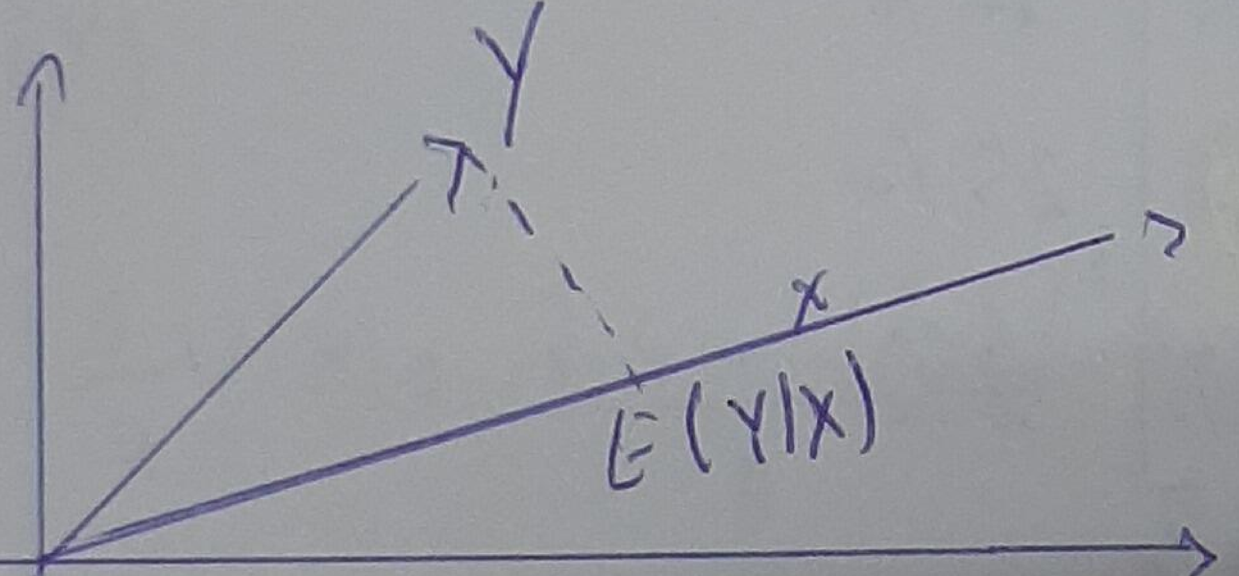
$$E(Y|X=\tfrac{1}{2}) = 1 \times \frac{1}{2} + 2 \times \frac{1}{2}$$

$E(Y|X) = \ell(X)$, ℓ?

x_i	$\ell(x_i) = E(Y\|X=x_i)$
1	$E(Y\|X=1) = 1$
2	$3/2$
3	2
4	$5/2$

Notes:

It can be shown that the cond. Expectation of "Y given X" $E(Y|X)$ is the "closest" function of X to Y



- $E(Y/X)$ represents the part of Y that can be described as a function of X

How much Y is dependent of X?

How Y is dependent of X?

Thm:

- Let X, Y two discrete r.v, then

$$E(E(Y|X)) = E(Y)$$

Proof:

$$E(E(Y|X)) = \sum_i E(Y|X=x_i)\,\mathbb{P}\{X=x_i\} = \sum_i \sum_j y_j \mathbb{P}\{Y=y_j|X=x_i\}\,\mathbb{P}\{X=x_i\}$$

$= \sum_{ij} y_j \, \mathbb{P}\{Y = y_j ; X = x_i\} = E(Y)$

Note :

$Y = E(Y|X) + \underbrace{Y - E(Y|X)}_{E(\text{residual}) = 0}$

§ Variance conditionnelle

$V(Y|X)$?

$V(Y|X = x_i) = E(Y^2 | X = x_i) - E(Y|X = x_i)^2$

Exemple :

$$
\begin{array}{cc|ccc|c}
 & & 1 & 2 & 3 & \mathbb{P}_X \\
 & 1 & 2\% & 5\% & 10\% & 17\% \\
\mathbb{P}_{XY} = & 2 & 5\% & 10\% & 10\% & 25\% \\
 & 3 & 30\% & 14\% & 14\% & 58\% \\
\mathbb{P}_Y & & 37\% & 29\% & 34\% &
\end{array}
$$

Remarque

$E(Y|X) = \text{const} = E(Y)$ qd $X \perp Y$
independent of

Y	$\mathbb{P}_{Y\|X=1}$
1	$\frac{2}{17}$
2	$5/17$
3	$\frac{10}{17}$

$E(Y|X = 1)$

Continuous Random pair couple

Thm :

let X, Y be two discrete r.v

ANOVA: fitted/explained variance

$$V(Y) = V(E(Y|X)) + E(V(Y|X))$$

Residual variance

$$R^2_{XY} = \frac{V(E(Y|X))}{V(Y)} \in [0,1]$$

if R^2_{XY} ↗ ⇝ X explains well Y.

Exemple:

*) $E(Y|X)$?

$E(Y|X=k)$ or $(Y|X=k) \sim U\{1, - k\}$

$$E(Y|X=k) = \frac{k+1}{2} \Rightarrow E(Y|X) = \frac{X+1}{2}$$

$$\ell(t) = \frac{t+1}{2}$$

$$R_{XY} = \frac{V(E(Y|X))}{V(Y)}$$

$$V\left(\frac{X+1}{2}\right) = \frac{1}{4}V(X) = \frac{7{,}5 - (2{,}5)^2}{4} = \frac{1{,}25}{4}$$

X^2	P_X
1	1/4
4	1/4
9	1/4
16	1/4

$E(X^2) = 7{,}2$

Y	P_Y
1	
2	
3	
4	