
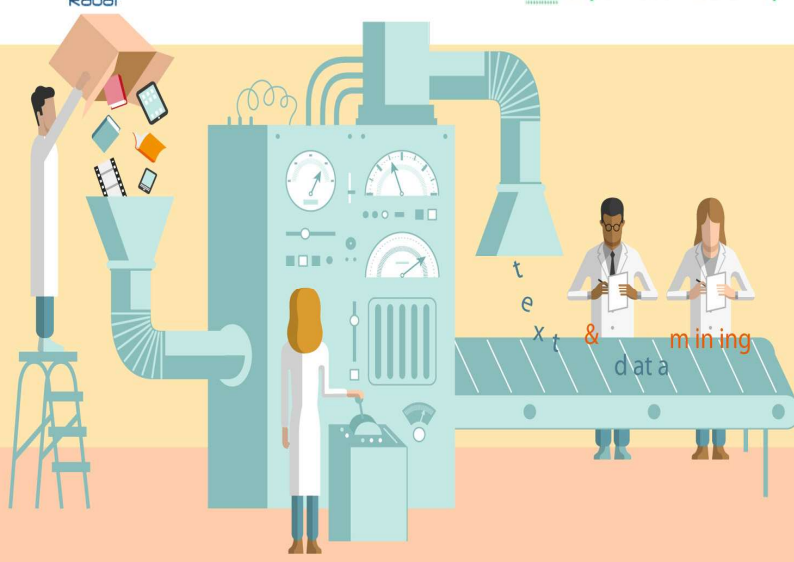


Université Mohammed V  
Faculté des Sciences  
Rabat

**Royaume du Maroc**  
Université Mohammed V de Rabat  
Faculté des Sciences  
Département d'Informatique



**IPSS**  
Intelligent Processing  
Systems & Security



# Data Mining & Machine Learning

Master IPS  
Faculté des sciences – Rabat  
Université Mohamed V

93

**Chapitre 1 (suite)**

# Data Mining : Compréhension des données

Analyse bivariable



ipes.boutyour@gmail.com

94

94

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée:

- Détecter les liaisons entre la variable cible et les variables explicatives
  - Garder les variables explicatives les plus discriminantes
  - Eliminer les variables explicatives sans aucun impact
- Détecter les liaisons entre les variables explicatives entre elles, qui sont à éviter dans certaines techniques

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée:

- Détecter liaisons entre la variable cible et les variables
- Deux variables quantitatives (représentées par le nuage de points)
- Deux variables qualitatives (représentées par les diagrammes en barres)
- Une variable qualitative et une variable quantitative (représentées par les boîtes parallèles)

## Compréhension des données

### Troisième étape: explorer les données

- Analyse bivariable: Deux variables quantitatives
  - Coefficient de corrélation linéaire (coefficient de Pearson)
    - Indicateur rendant compte numériquement de la manière dont deux variables quantitatives varient simultanément
    - Mesure leur degré de liaison linéaire
    - Nécessite la mesure de la covariance

## Compréhension des données

### Troisième étape: explorer les données

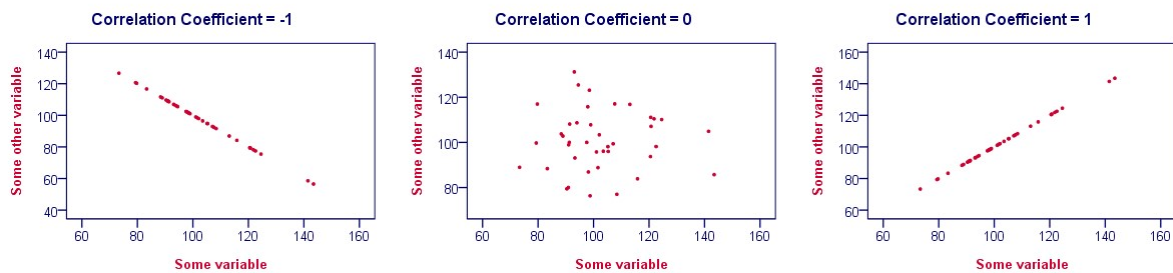
- Analyse bivariable: Deux variables quantitatives
  - Coefficient de corrélation linéaire (coefficient de Pearson)
    - Soit deux variables  $X=(x_1, \dots, x_n)$  et  $Y=(y_1, \dots, y_n)$
    - $Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{X})(y_i - \bar{Y})] \rightarrow$  Pas de signification concrète
    - Passer au coefficient de corrélation pour avoir une valeur interprétable
 
$$r = Cov(X, Y) / \sigma(X) \cdot \sigma(Y)$$
      - ✓  $-1 < r < 1$
      - ✓  $r > 0 \rightarrow X$  et  $Y$  varient dans le même sens
      - ✓  $r < 0 \rightarrow X$  et  $Y$  varient en sens opposé

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables quantitatives

##### ■ Coefficient de corrélation linéaire (coefficient de Pearson)



ipes.boutyour@gmail.com

99

99

## Compréhension des données

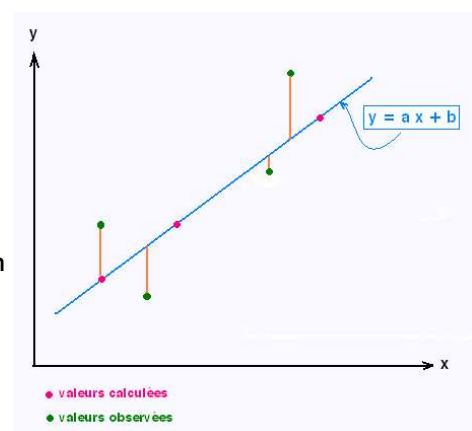
### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables

##### quantitatives

##### ■ Régression linéaire

- Si X, Y correctement corrélées ( $|r|$  proche de 1) et X est cause de Y alors on cherche une fonction linéaire  $f(X) = aX + b$  de X approchant au mieux Y
- → Régression de Y sur X



ipes.boutyour@gmail.com

100

100

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables quantitatives

##### ■ Régression linéaire

- Si X, Y correctement corrélées ( $|r|$  proche de 1) et X est cause de Y alors on cherche une fonction linéaire  $f(X) = aX + b$  de X approchant au mieux Y
- → **Régression de Y sur X**
- Chercher la droite qui passe au mieux dans le nuage de points sera obtenue à l'aide du « critère des moindres carrés » → droite de régression
- Trouver a et b qui minimisent :

$$F(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

ipes.boutyour@gmail.com

101

101

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

###### ■ Corrélation :

- Liaison entre deux variables quantitatives X et Y
- Rôle symétrique

###### ■ Régression :

- Liaison entre deux variables quantitatives X et Y
- Rôle asymétrique uniquement
  - X = variable explicative / Y = variable expliquée
  - X = variable indépendante / Y = variable dépendante
- La corrélation mesure l'intensité de la liaison entre des variables, tandis que la régression analyse la relation d'une variable par rapport à une ou plusieurs autres.

ipes.boutyour@gmail.com

102

102

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

###### ■ Corrélation :

- La corrélation mesure l'intensité de la liaison entre des variables

###### ■ Régression :

- la régression analyse la relation d'une variable par rapport à une ou plusieurs autres.

On peut dire alors que la différence entre ces deux mesures statistiques est que la corrélation mesure le degré d'une relation entre deux variables (x et y), tandis que la régression est la façon dont une variable affecte une autre.

ipes.boutyour@gmail.com

103

103

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

###### 1. Exemple : corrélation (positive)

- X = ventes de paires de lunettes de soleil en été

- Y = ventes de crèmes glacées en été

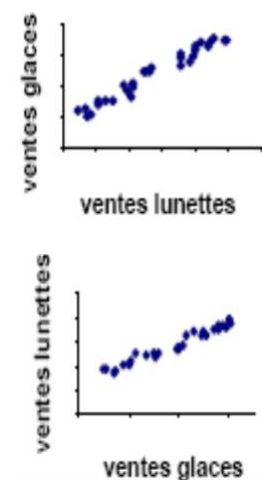
###### ■ Il existe une liaison entre X et Y :

- Quand X augmente, Y augmente (météo estivale)
- Quand X diminue, Y diminue (météo pluvieuse)

###### ■ La liaison est **symétrique** :

- X est liée à Y, et Y est liée à X
- mais X ne dépend pas de Y et Y ne dépend pas de X
- on peut permuter X et Y en abscisses et en ordonnées

###### ■ Y ne peut pas être prédite par X



ipes.boutyour@gmail.com

104

104

## Compréhension des données

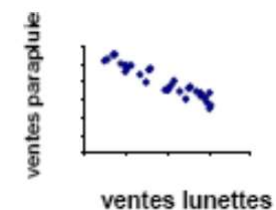
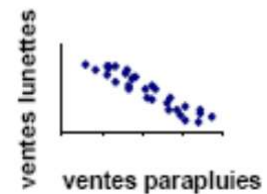
### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

##### 2. Exemple : corrélation (négative)

- X = ventes de paires de lunettes de soleil en été
- Y = ventes de parapluies en été
- **Il existe une liaison entre X et Y :**
  - Quand X augmente, Y diminue (météo estivale)
  - Quand X diminue, Y augmente (météo pluvieuse)
- **La liaison est symétrique :**
  - X est liée à Y, et Y est liée à X
  - mais X ne dépend pas de Y et Y ne dépend pas de X
  - on peut permuter X et Y en abscisses et en ordonnées
- **Y ne peut pas être prédite par X**



ipes.boutyour@gmail.com

105

105

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

##### 3. Exemple : régression

- X = âge (de 0 à 15 ans)
- Y = taille (cm)
- **Il existe une liaison entre X et Y :**
  - Quand l'âge augmente, la taille augmente
  - Quand l'âge diminue, la taille diminue
- **La liaison est asymétrique :**
  - la taille dépend de l'âge mais l'âge ne dépend pas de la taille
  - on ne peut pas permuter X et Y en abscisses et en ordonnées
- **On peut prédire la taille par l'âge à l'aide d'une équation de droite ou de courbe de régression**

ipes.boutyour@gmail.com

106

106

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables quantitatives

##### ■ Corrélation vs Régression linéaire

	Corrélation	Régression
Variables	X est quantitative Y est quantitative	X est quantitative Y est quantitative
Symétrie de la liaison	Oui/Non Y liée à X X liée à Y	Non Y dépend de X -
Exemples	Y = cons. cigarettes X = Temp. Moy. annuelle	Y = taille X = âge
Prédiction	Non	Oui

ipes.boutyour@gmail.com

107

107

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

##### ■ Présenter les données sous forme d'une table de contingence

##### ■ Soit deux variables X et Y

- X à r modalités notées  $x_1, \dots, x_r$
- Y à c modalités notées  $y_1, \dots, y_c$

	$y_1$	...	$y_h$	...	$y_c$	Total
$x_1$	$n_{11}$	...	$x_{1h}$	...	$x_{1c}$	$n_{1+}$
...		...	...	...	...	
$x_l$	$n_{l1}$	...	$x_{lh}$	...	$x_{lc}$	$n_{l+}$
...		...	...	...	...	...
$x_r$	$n_{r1}$		$n_{rh}$		$n_{rc}$	$n_{r+}$
Total	$n_{+1}$		$n_{+h}$		$n_{+c}$	n

ipes.boutyour@gmail.com

108

108



## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

- Exemple : On veut dresser le tableau de contingence de la variable sexe par rapport à la variable couleur des yeux.

Données triées par sexe, puis par couleur

	Prénom	Sexe	Couleur d'yeux
1	Bernadette	F	Bleus
6	Sophie	F	Bleus
4	Marie	F	Noirs
2	Jean-Pierre	M	Bleus
3	Marc	M	Noirs
5	Pierre	M	Noirs

ipes.boutyour@gmail.com

109

109

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

- Exemple : On veut dresser le tableau de contingence de la variable sexe par rapport à la variable couleur des yeux.

Données triées par sexe, puis par couleur

	Sexe		Total
Couleur des yeux	F	M	
Bleus			
Noirs			
Total			

	Prénom	Sexe	Couleur d'yeux
1	Bernadette	F	Bleus
6	Sophie	F	Bleus
4	Marie	F	Noirs
2	Jean-Pierre	M	Bleus
3	Marc	M	Noirs
5	Pierre	M	Noirs

ipes.boutyour@gmail.com

110

110

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

- Exemple : On veut dresser le tableau de contingence de la variable sexe par rapport à la variable couleur des yeux.

Données triées par sexe, puis par couleur

Couleur des yeux	Sexe		Total
	F	M	
Bleus	2	1	3
Noirs	1	2	3
Total	3	3	6

	Prénom	Sexe	Couleur d'yeux
1	Bernadette	F	Bleus
6	Sophie	F	Bleus
4	Marie	F	Noirs
2	Jean-Pierre	M	Bleus
3	Marc	M	Noirs
5	Pierre	M	Noirs

ipes.boutyour@gmail.com

111

111

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives : test du khi-deux

- Un test du Khi deux est un test d'hypothèse qui compare la loi de distribution observée de vos données à une loi attendue.  $\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$  avec  $o_i$  = effectifs observés et  $e_i$  = effectifs théoriques
- **Test d'ajustement du Khi deux** : Cette analyse permet de vérifier à quel point un échantillon de données de catégorie est ajusté à une loi théorique.
- **Tests d'association et d'indépendance** du Khi deux :
  - **Test d'association** : on peut utiliser un test d'association afin de déterminer si une variable est associée à une autre.
  - **Test d'indépendance** : on utilise un test d'indépendance afin de déterminer si la valeur observée d'une variable dépend de la valeur observée d'une autre variable.

ipes.boutyour@gmail.com

112

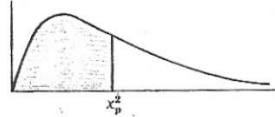
112

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables qualitatives : test du khi-deux

VALEURS DES CENTILES ( $\chi_p^2$ )  
pour la  
DISTRIBUTION du KHI-DEUX  
en fonction du nombre  $\nu$  de degrés de liberté  
(aire en grisé =  $p$ )



$\nu$	$\chi_{0,995}^2$	$\chi_{0,99}^2$	$\chi_{0,975}^2$	$\chi_{0,95}^2$	$\chi_{0,90}^2$	$\chi_{0,75}^2$	$\chi_{0,50}^2$	$\chi_{0,25}^2$	$\chi_{0,10}^2$	$\chi_{0,05}^2$	$\chi_{0,025}^2$	$\chi_{0,01}^2$	$\chi_{0,005}^2$
1	7,88	6,63	5,02	3,84	2,71	1,32	0,455	0,102	0,0158	0,0039	0,0010	0,0002	0,0000
2	10,6	9,21	7,38	5,99	4,61	2,77	1,39	0,575	0,211	0,103	0,0506	0,0201	0,0100
3	12,8	11,3	9,35	7,81	6,25	4,11	2,37	1,21	0,584	0,352	0,216	0,115	0,072
4	14,9	13,3	11,1	9,49	7,78	5,39	3,36	1,92	1,06	0,711	0,484	0,297	0,207
5	16,7	15,1	12,8	11,1	9,24	6,63	4,35	2,57	1,61	1,15	0,831	0,554	0,412
6	18,5	16,8	14,4	12,6	10,6	7,84	5,35	3,45	2,20	1,64	1,24	0,872	0,676
7	20,3	18,5	16,0	14,1	12,0	9,04	6,35	4,25	2,83	2,17	1,69	1,24	0,989
8	22,0	20,1	17,5	15,5	13,4	10,2	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,6	21,7	19,0	16,9	14,7	11,4	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,2	23,2	20,5	18,3	16,0	12,5	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,8	24,7	21,9	19,7	17,3	13,7	10,3	7,58	5,58	4,57	3,82	3,05	2,60
12	28,3	26,2	23,3	21,0	18,5	14,8	11,3	8,44	6,30	5,23	4,40	3,57	3,07
13	29,8	27,7	24,7	22,4	19,8	16,0	12,3	9,30	7,04	5,89	5,01	4,11	3,57
14	31,3	29,1	26,1	23,7	21,1	17,1	13,3	10,2	7,79	6,57	5,63	4,66	4,07

ipes.boutyour@gmail.com

113

113

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Deux variables qualitatives : test du khi-deux

##### ■ Exemple de calcul du Khi deux d'ajustement :

- Pour tester si un dé n'est pas truqué, on le jette 150 fois et on note les résultats obtenus :

1	2	3	4	5	6
17	26	38	22	25	22

- En posant comme hypothèse nulle « le dé n'est pas truqué », on s'attend à ce que les effectifs observés ne diffèrent pas des effectifs théoriques, qui sont 25, 25, 25, ..., 25 (150 divisé par 6)

$$\chi^2_{\text{observé}} = \frac{(17-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(38-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(25-25)^2}{25} + \frac{(22-25)^2}{25} = 10,08$$

On fixe le seuil de significativité à 10% par exemple, le nombre de degrés de liberté est égal à  $6-1=5$ .

On lit dans la table :  $\chi^2_{0,90}$  à  $\nu = 5 \rightarrow \chi^2_{\text{théorique}} = 9,24$

Dans notre cas,  $\chi^2_{\text{observé}} > \chi^2_{\text{théorique}}$ , on rejette  $H_0$  (et on conclut que le dé est truqué) avec 10 chances sur 100 de se tromper.

ipes.boutyour@gmail.com

114

114

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives : test du khi-deux

##### ■ Exemple de calcul du Khi deux de croisement :

- On travaille à partir du tableau de contingence qui sert à calculer les effectifs théoriques :
- En général,

	A	B	totaux marginaux totaux de ligne
X	e <sub>11</sub>	e <sub>12</sub>	L <sub>1</sub>
Y	e <sub>21</sub>	e <sub>22</sub>	L <sub>2</sub>
totaux marginaux totaux de colonne	C <sub>1</sub>	C <sub>2</sub>	N

$$e_{11} = (L_1 \times C_1) / N$$

$$e_{21} = (L_2 \times C_1) / N$$

1) relation entre tabagisme et sexe

	non-fumeurs	fumeurs	
hommes	350	150	500
femmes	400	100	500
	750	250	1000

$$\chi^2 = \frac{1000 \times (350 \times 100 - 400 \times 150)^2}{500 \times 500 \times 750 \times 250} = 13.33$$

ipes.boutyour@gmail.com 115

115

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

##### ❖ Autres tests liés au khi-deux : mesures d'association

- Le coefficient de contingence :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

**Exemple 1 :**  $C = \sqrt{\frac{13.33}{13.33 + 10}} = 0.11$

- Le coefficient Phi-deux  $\Phi^2$  :

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

- Le coefficient V de Cramer :

$$V = \sqrt{\frac{\Phi^2}{\min(K_1 - 1, K_2 - 1)}}$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(K_1 - 1, K_2 - 1)}}$$

ipes.boutyour@gmail.com 116

116

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

#### ❖ Autres tests liés au khi-deux : mesures d'association

- Le coefficient **T** de **Tshuprow** :

$$T = \sqrt{\frac{\phi^2}{\sqrt{(r-1)(c-1)}}} \quad 0 \leq T \leq 1$$

**Plus T est grand, plus la liaison est forte**

- Les coefficients de **Cramer** et de **Tshuprow** très utilisés dans la pratique.
- Plus souvent compris entre 0,1 et 0,3, rarement supérieurs à 0,5
- Pour plus d'infos consulter les références suivantes:
  - ✓ [Revue de statistique appliquée](#)
  - ✓ [Tests du khi-deux](#)

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

#### ❖ Autres tests liés au khi-deux : mesures d'association

- Interprétation du coefficient de Cramer

Valeur de Cramer	Intensité de la relation entre les variables
< 0.10	Relation nulle ou très faible
>=0.10 et <0.20	Relation faible
>=0.20 et <0.30	Relation moyenne
>=0.3	Relation forte

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

#### ❖ Autres tests liés au khi-deux : mesures d'association

- Exemple d'application : Étant donné khi-deux = 12,85, calculer C, phi-deux, T et V

	Y1	Y2	Y3	Total
X1	10	15	15	40
X2	20	5	35	60
Total	30	20	50	100

ipes.boutyour@gmail.com 119

119

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariée: Deux variables qualitatives

#### ❖ Autres tests liés au khi-deux : mesures d'association

- Exemple d'application : Étant donné khi-deux = 12.85, calculer C, phi-deux, T et V

	Y1	Y2	Y3	Total
X1	10	15	15	40
X2	20	5	35	60
Total	30	20	50	100

$$C = \sqrt{\frac{12.85}{12.85 + 100}} \quad \phi = \sqrt{\frac{12.85}{100}} \quad V = \sqrt{\frac{\chi^2}{100 \cdot \min((2-1), (3-1))}} \quad T = \sqrt{\frac{\phi^2}{\sqrt{(2-1)(3-1)}}}$$

ipes.boutyour@gmail.com 120

120

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Une variable quantitative et une variable qualitative

- ❖ Soient X variable qualitative à r modalités notées  $X_1, \dots, X_L, \dots, X_r$  et Y variable quantitative
- ❖ La classe courante, notée  $C_L$ , contient les individus ayant la modalité  $X_L$  de X
- ❖  $n_L$  effectif de la classe  $C_L \rightarrow \sum n_L = n$  effectif total
- ❖ On peut définir pour chaque classe  $C_L$ :

- Moyenne partielle de Y :

$$\hat{y}_L = \frac{1}{n_L} \sum_{i \in C_L} y_i$$

- Variance partielle de Y :

$$s^2_{L} = \frac{1}{n_L} \sum_{i \in C_L} (y_i - \hat{y}_L)^2$$

ipes.boutyour@gmail.com

121

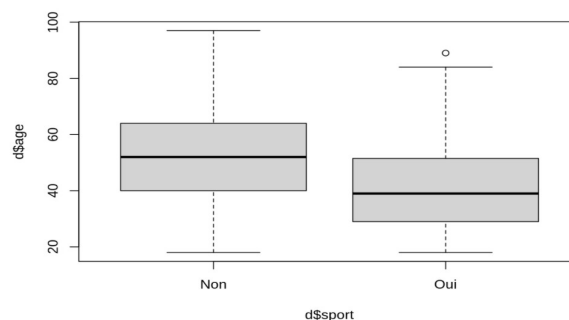
121

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyse bivariable: Une variable quantitative et une variable qualitative

- ❖ Représentation graphique:
  - boîtes parallèles  $\rightarrow$  boîte à moustache de Y dans chaque classe courante  $C_L$



ipes.boutyour@gmail.com

122

122

## Compréhension des données

### Troisième étape: explorer les données

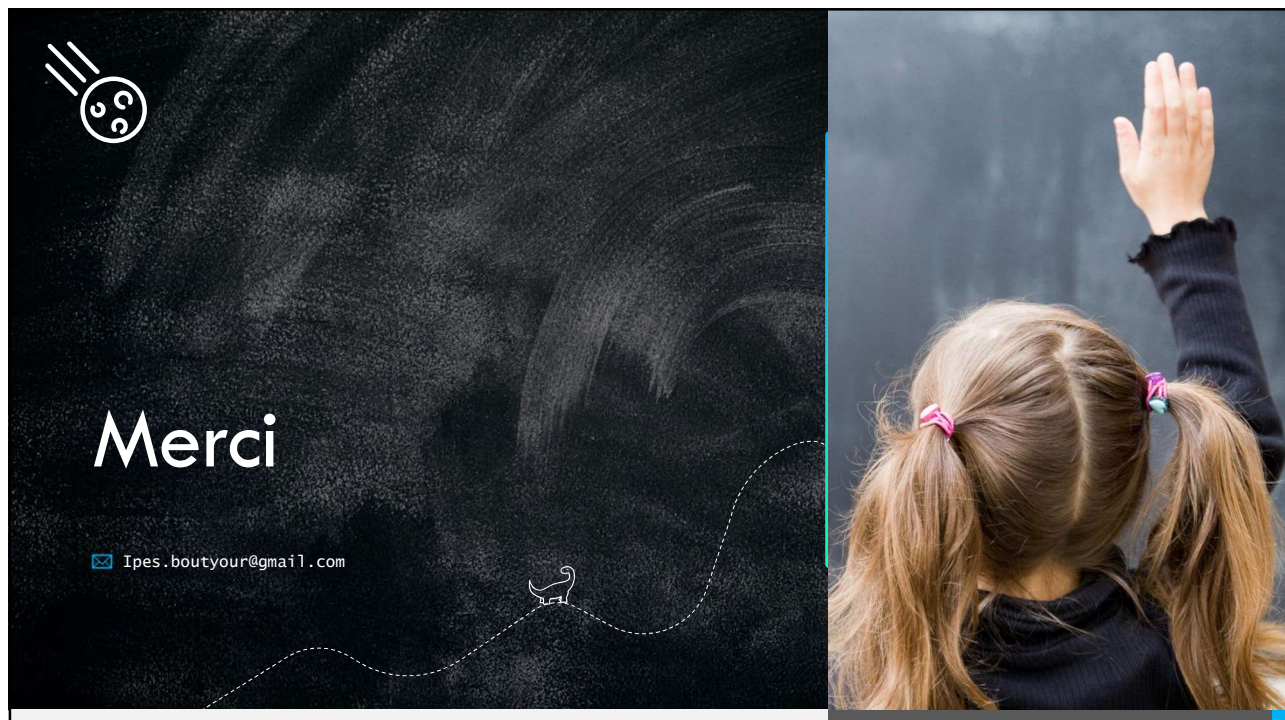
○ Analyse bivariable: Une variable quantitative et une variable qualitative

#### ❖ Rapport de corrélation :

- $S^2_Y$  est décomposée telle que  $S^2_Y = S^2_E + S^2_R$
- $S^2_E$  variance expliquée par X (variance **inter-classes**)
  - Mesure de l'influence des valeurs de X sur Y
  - Ce que serait  $S^2_Y$  si la valeur de Y était constante ( $= \hat{y}_L$ ) dans chaque classe  $C_L$
- $S^2_R$  variance résiduelle ou variance **intra-classes**

ipes.boutyour@gmail.com 123

123



124



# TP1

## Manipulations de base avec le langage R

Objectifs :

1. **Découverte de l'environnement Rstudio**
2. **Utiliser quelques fonctions de base du langage R**
3. **Manipuler et transformer les données avec les packages dplyr, tidyr et d'autres.**
4. **Créer des représentations graphiques pour les données avec les fonctions natives et le package ggplot2**

Par: Youness BOUTYOUR

ipes.boutyour@gmail.com
125

125



ipes.boutyour@gmail.com
126

126