

# **Chapitre 7**

## **Données catégorielles et khi-carré**

# Les tests khi-carré

---

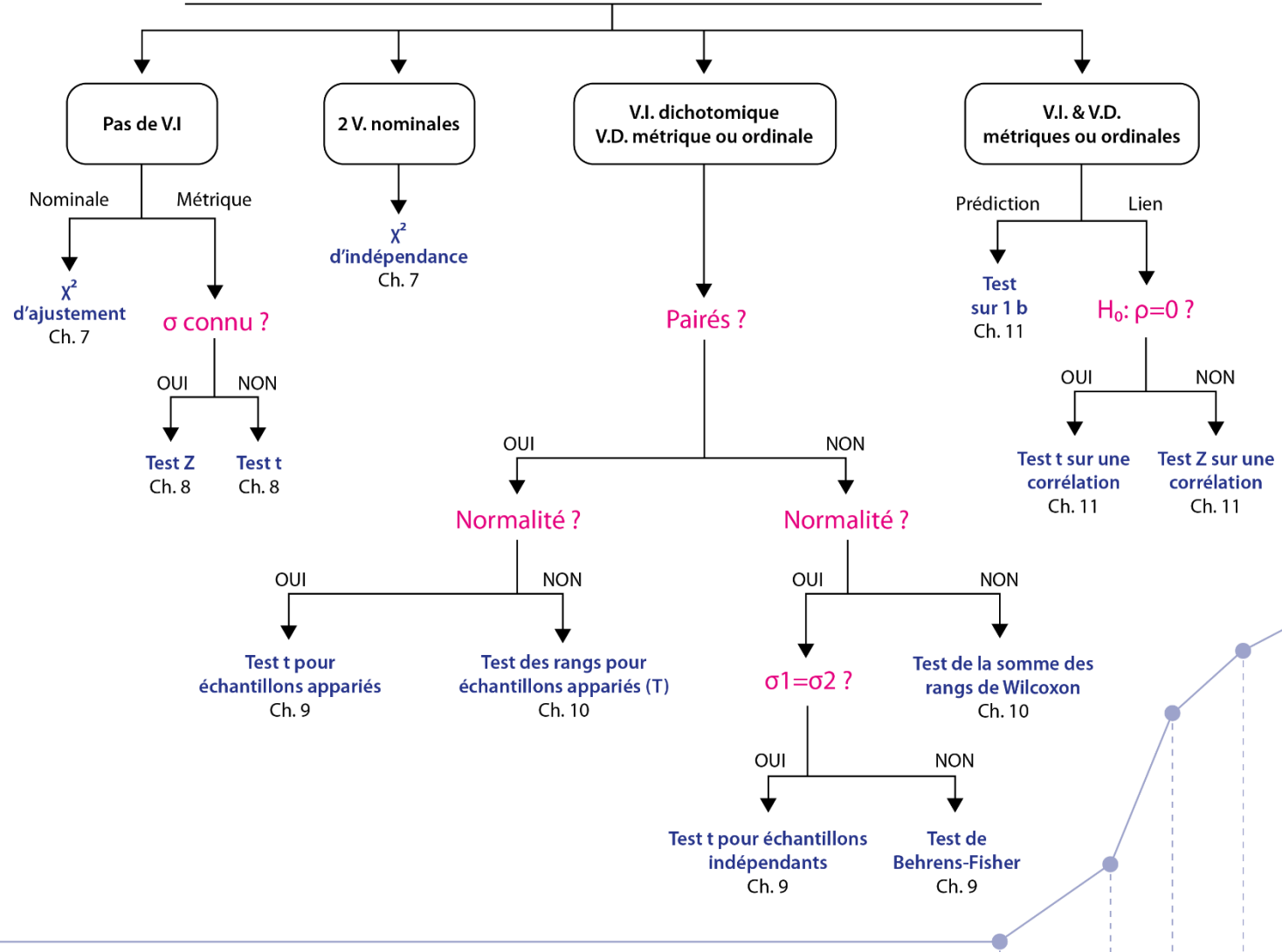
**Variables nominales**

**Fréquences**

**Khi-carré ( $\chi^2$ ) est une distribution  
mathématique**

# Choisir le bon test d'hypothèse

## Identifier les variables et déterminer leurs natures



# Le test khi-carré d'ajustement

---

**Une seule variable nominale**

**Teste si une distribution théorique s'applique**

**(càd une certaine répartition de % dans les catégories de la VD pour la population)**

# Le test khi-carré d'ajustement

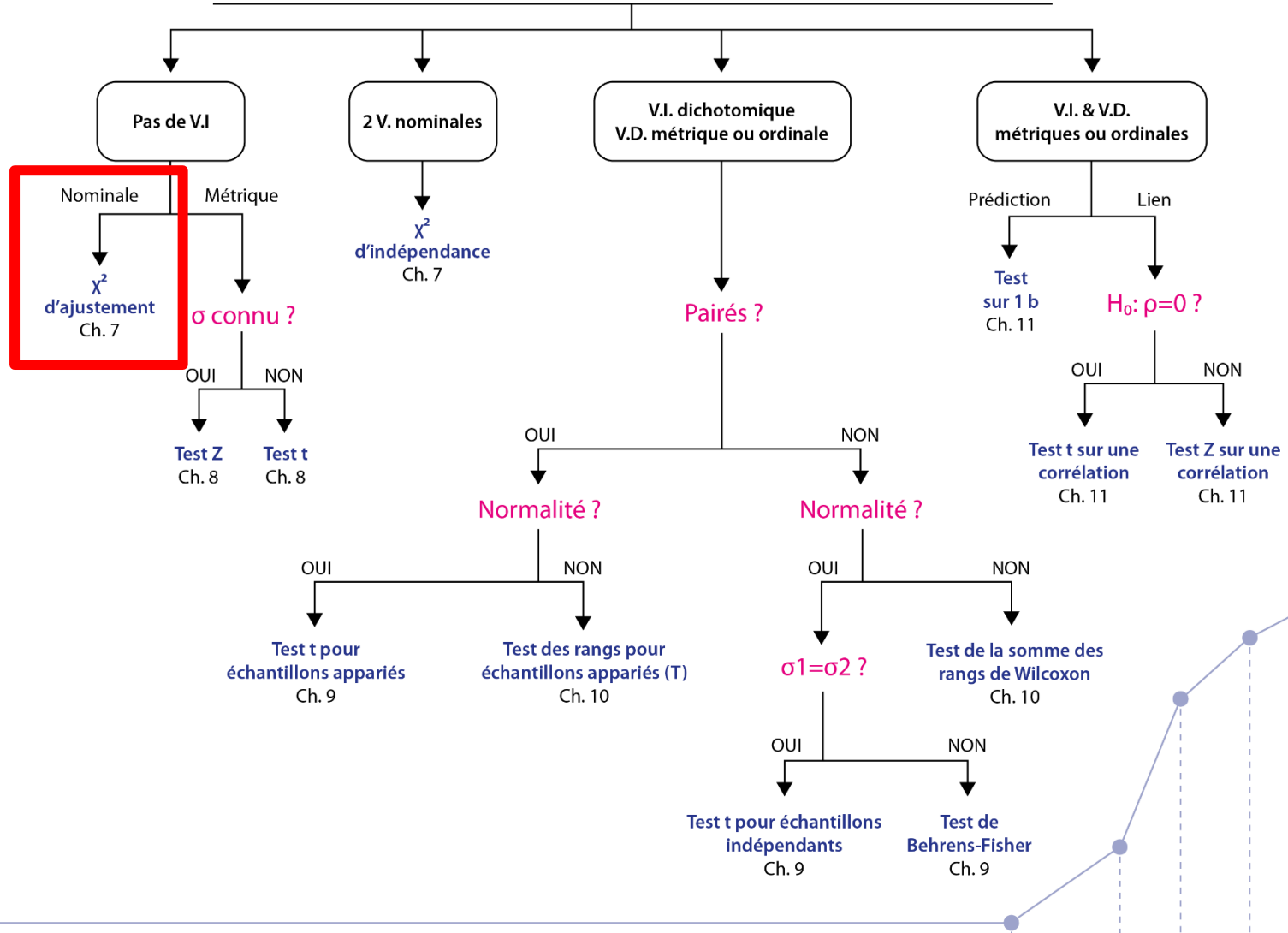
---

**Exemple: Les hommes sont-ils plus représentés dans les professions médicales?**

**= répartition différente de 50%-50%**

# Choisir le bon test d'hypothèse

## Identifier les variables et déterminer leurs natures



# Le test khi-carré d'ajustement

Les hommes sont-ils plus représentés dans les professions médicales?

	M	F	Total
Fréquences observées	65	35	100

Différence significative ou erreur d'échantillonnage ?

# Variabilité due au hasard et erreur d'échantillonnage

**Simulation de 5 échantillons de 100 personnes dans une population avec 50% d'hommes et de femmes**

	Hommes	Femmes	% d'hommes
Echantillon 1	42	58	42%
Echantillon 2	56	44	56%
Echantillon 3	51	49	51%
Echantillon 4	54	46	54%
Echantillon 5	48	52	48%



# Variabilité due au hasard et erreur d'échantillonnage

**Simulation de 5 échantillons de 10 personnes dans une population avec 50% d'hommes et de femmes**

	Hommes	Femmes	% d'hommes
Echantillon 1	8	2	80%
Echantillon 2	5	5	50%
Echantillon 3	4	6	40%
Echantillon 4	3	7	30%
Echantillon 5	4	6	40%

# Le test khi-carré d'ajustement

---

$H_0$  : La population présente un rapport hommes/femmes de 1 : 1

$H_A$  : La population présente un rapport hommes/femmes différent de 1 : 1

# Le test khi-carré d'ajustement

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

**k** = le nombre de catégories

**O<sub>i</sub>** = les fréquences observées dans chaque catégorie

**E<sub>i</sub>** = les fréquences attendues dans chaque catégorie selon l'hypothèse H<sub>0</sub>

# Comment calculer les fréquences attendues ?

	M	F	Total
Fréquences observées	65	35	100
Fréquences attendues	50	50	100

$$\text{Fréquence attendue} = N \times p$$

# Calcul du khi-carré

	M	F	Total
Fréquences observées	65	35	100
Fréquences attendues	50	50	100

$$\chi^2_{obs} = \frac{(65-50)^2}{50} + \frac{(35-50)^2}{50} = \frac{225}{50} + \frac{225}{50} = 9$$

# La formule khi-carré

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Plus il y a désaccord entre O et E, plus le  $\chi^2$  est élevé

On met les écarts au carré pour éviter les valeurs négatives

# La formule khi-carré

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

**La division par les fréquences attendues permet de maintenir les écarts en proportion**

# La formule khi-carré

La division par les fréquences attendues permet de maintenir les écarts en proportion

Exemple:

N =20	M	F	% M
Fréquences observées	15	5	75%
Fréquences attendues	10	10	50%



# La formule khi-carré

La division par les fréquences attendues permet de maintenir les écarts en proportion

Exemple:

N =100	M	F	% M
Fréquences observées	55	45	55%
Fréquences attendues	50	50	50%

# La formule khi-carré

---

Le  $\chi^2$  calculé est une mesure du désaccord entre l'hypothèse et les données

Ajustement parfait:  $\chi^2 = 0$

Plus le  $\chi^2$  calculé est grand et plus l'hypothèse  $H_0$  est improbable

# La formule khi-carré

---

Si  $H_0$  est fausse,  $\chi^2$  doit être élevé

Mais  $\chi^2$  peut aussi être élevé en raison de l'erreur d'échantillonnage

Calculer la probabilité d'obtenir un  $\chi^2$  aussi élevé par le simple fait du hasard

# Comment calculer la probabilité liée à un $\chi^2$

---

Utilisation de la table  $\chi^2$

Les degrés de liberté (dl):

$$dl = k - 1$$

<b>dl</b>	<b>0,100</b>	<b>0,050</b>	<b>0,025</b>	<b>0,010</b>	<b>0,005</b>
<b>1</b>	<b>2,71</b>	<b>3,84</b>	<b>5,02</b>	<b>6,63</b>	<b>7,88</b>
<b>2</b>	<b>4,61</b>	<b>5,99</b>	<b>7,38</b>	<b>9,21</b>	<b>10,60</b>
<b>3</b>	<b>6,25</b>	<b>7,81</b>	<b>9,35</b>	<b>11,34</b>	<b>12,84</b>
<b>4</b>	<b>7,78</b>	<b>9,49</b>	<b>11,14</b>	<b>13,28</b>	<b>14,86</b>
<b>5</b>	<b>9,24</b>	<b>11,07</b>	<b>12,83</b>	<b>15,09</b>	<b>16,75</b>
<b>6</b>	<b>10,64</b>	<b>12,59</b>	<b>14,45</b>	<b>16,81</b>	<b>18,55</b>
<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>

# Exemple

---

$H_0$  : La population présente un rapport hommes/femmes de 1 : 1

$H_A$  : La population présente un rapport hommes/femmes différent de 1 : 1

# Exemple

	M	F	Total
Fréquences observées	65	35	100
Fréquences attendues	50	50	100

$$\chi^2_{obs} = \frac{(65-50)^2}{50} + \frac{(35-50)^2}{50} = \frac{225}{50} + \frac{225}{50} = 9$$

# Exemple

---

$$dl = k - 1 = 2 - 1 = 1$$

$$\chi_{obs}^2 = 9$$

Quelle est la probabilité d'avoir un  $\chi^2$  aussi élevé que 9 avec 1 dl ?



dl	0,100	0,050	0,025	0,010	0,005
1	2,71	3,84	5,02	6,63	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,81	9,35	11,34	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,64	12,59	14,45	16,81	18,55
...	...	...	...	...	...

# Exemple

$$dl = k - 1 = 2 - 1 = 1$$

$$\chi_{obs}^2 = 9$$

$$\chi_{0,05}^2 = 3,84$$

**Puisque  $9 > 3,84$ , alors  $p < 0,05$ , rejeter  $H_0$**

**NB:  $p = 0,0027$**

# Exemple

---

**Conclusion:**

**Nous avons rejeté l'hypothèse  $H_0$  selon laquelle la proportion Hô/Fê est de 1:1**

**Il y a plus d'hommes que de femmes qui travaillent dans le monde médical**

# Test khi-carré d'ajustement avec plus de deux catégories

---

Test et formule identiques

Seuls changent les  $dl = k - 1$

# La formule khi-carré

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

On doit nécessairement s'attendre à un  $\chi^2$  de plus en plus élevé en fonction du nombre de catégories

<b>dl</b>	<b>0,100</b>	<b>0,050</b>	<b>0,025</b>	<b>0,010</b>	<b>0,005</b>
<b>1</b>	<b>2,71</b>	<b>3,84</b>	<b>5,02</b>	<b>6,63</b>	<b>7,88</b>
<b>2</b>	<b>4,61</b>	<b>5,99</b>	<b>7,38</b>	<b>9,21</b>	<b>10,60</b>
<b>3</b>	<b>6,25</b>	<b>7,81</b>	<b>9,35</b>	<b>11,34</b>	<b>12,84</b>
<b>4</b>	<b>7,78</b>	<b>9,49</b>	<b>11,14</b>	<b>13,28</b>	<b>14,86</b>
<b>5</b>	<b>9,24</b>	<b>11,07</b>	<b>12,83</b>	<b>15,09</b>	<b>16,75</b>
<b>6</b>	<b>10,64</b>	<b>12,59</b>	<b>14,45</b>	<b>16,81</b>	<b>18,55</b>
<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>

# Les fréquences attendues

---

Comment déterminer les fréquences attendues ?

1. Parfois évident Ex: rapport hô/fê
2. Théorie
3. Etude antérieure

# Exemple

---

**Une grande marque de soda a créé une nouvelle boisson gazeuse qui existe en trois formules (A, B et C)**

**On décide de tester les trois formules sur 120 sujets**

**30 sujets ont choisi la formule A, 54 ont choisi la B et 36 la C**



# Exemple

---

$H_0$  : La préférence se répartit selon un rapport de 1 : 1 : 1

$H_A$  : La préférence se répartit selon un rapport différent de 1 : 1 : 1

# Exemple

	A	B	C	Total
Fréquences observées	30	54	36	120
Fréquences attendues	40	40	40	120

$$\chi^2_{obs} = \frac{(30-40)^2}{40} + \frac{(54-40)^2}{40} + \frac{(36-40)^2}{40} = 7,8$$

---

dl	.500	.250	.100	.050	.025	.010	.005
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75
...	...	...	...	...	...	...	...

---

# Exemple

$$\chi_{obs}^2 = 7,8$$

$$dl = 2$$

$$\chi_{0,05}^2 = 5,99$$

**Puisque  $7,8 > 5,99$ , alors  $p < 0,05$ , rejeter  $H_0$**

**NB:  $p = 0,0202$**

# Exemple

---

## Conclusion:

**On peut rejeter l'hypothèse selon laquelle les trois formules sont également appréciées.**

**Selon les résultats, la formule B semble plus appréciée que les autres.**

# Résumé des étapes du test khi-carré

---

1. Poser une hypothèse sur le rapport entre les différentes catégories
2. Récolter des observations et calculer le khi-carré d'ajustement sur ces observations

# Résumé des étapes du test khi-carré

---

3. Calculer la probabilité d'obtenir ces observations si l'hypothèse posée est correcte : comparer le khi-carré avec la valeur critique de la table

4. Tirer une conclusion :

Si le khi-carré dépasse la valeur critique, alors on rejette l'hypothèse  $H_0$

# Le test khi-carré d'indépendance

---

**2 variables nominales (ou plus)**

**Ces variables sont-elles indépendantes?**

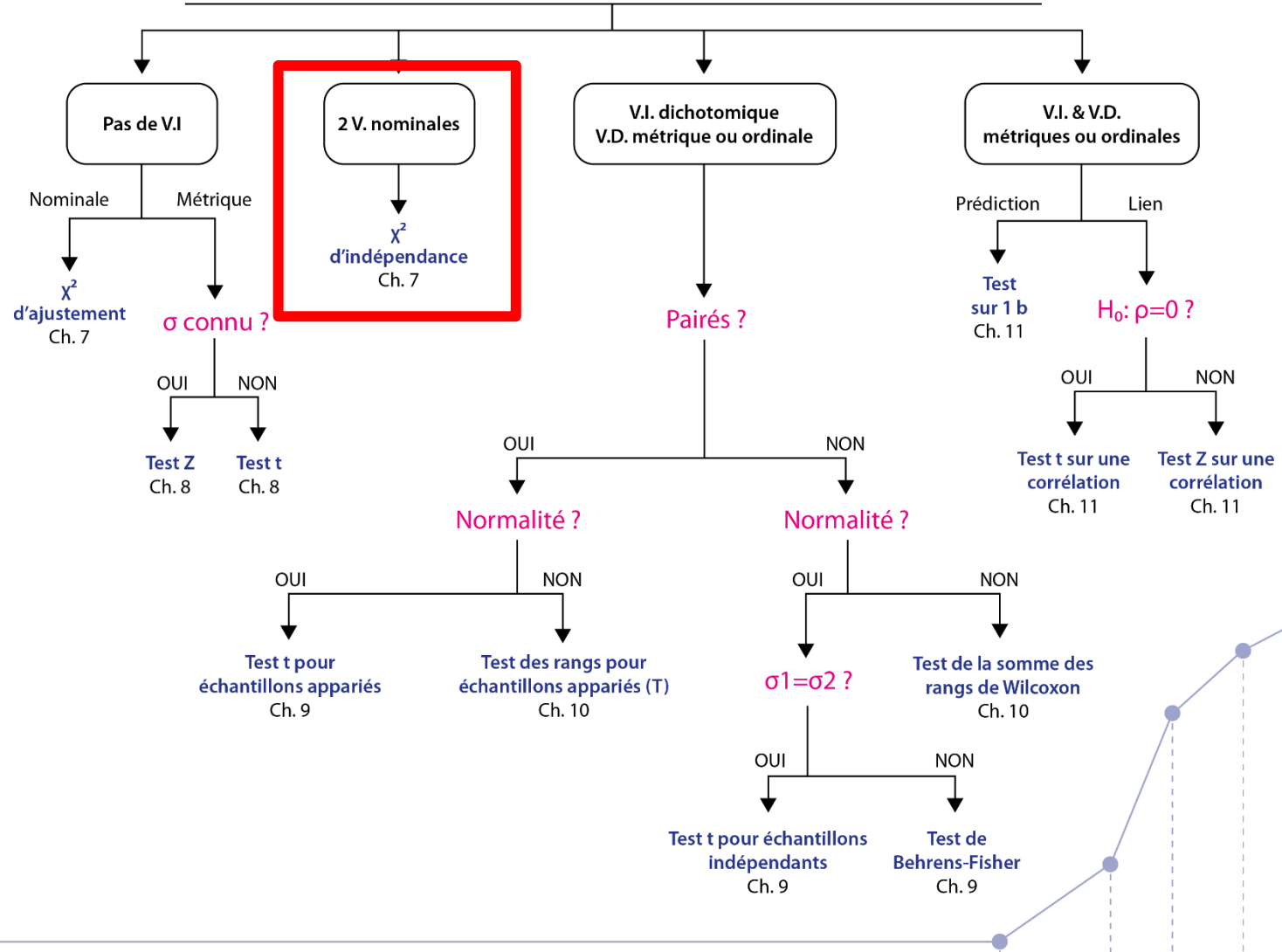
**Exemples:**

- **Le genre influence le fait de travailler à temps partiel ?**
- **Le fait d'être croyant influence le fait de se marier ?**



# Choisir le bon test d'hypothèse

## Identifier les variables et déterminer leurs natures



# Les tables de contingence

---

**Exemple de l'étude de Pugh (1983):**

**Le fait de souligner la faute de la victime  
influence-t-il le verdict de culpabilité?**

# Les tables de contingence

Faute	Verdict		Total
	Coupable	Non coupable	
Légère	153	24	177
Grave	105	76	181
Total	258	100	358

# Le test khi-carré d'indépendance

---

$H_0$  : Le verdict est indépendant de la faute attribuée à la victime

$H_A$  : Le verdict n'est pas indépendant de la faute attribuée à la victime

# Le test khi-carré d'indépendance

---

Pour calculer le khi-carré, il faut calculer les fréquences attendues si  $H_0$  est vraie

Càd si les variables sont indépendantes

---

## Verdict

---

Faute	Coupable	Non coupable	Total
Légère	153	24	177
Grave	105	76	181
Total	258	100	358

---

**Probabilité simple : total marginal / total général**

$$P(\text{coupable}) = T_C / N = 258/358 = 0,72$$

Faute	Verdict		Total
	Coupable	Non coupable	
Légère	153	24	177
Grave	105	76	181
Total	258	100	358

**Probabilité conjointe : Cellule / total général**

**$P(\text{Légère, non coupable}) = 24/358 = 0,067$**

# Le test khi-carré d'indépendance

---

**Si deux événements sont indépendants:**

**Probabilité conjointe se calcule en  
multipliant les probabilité simples**

**(loi multiplicative)**



---

## Verdict

---

Faute	Coupable	Non coupable	Total
Légère	153	24	177
Grave	105	76	181
Total	258	100	358

---

**Pour chaque cellule:**

$$P(E_{ij}) = \frac{L_i}{N} \times \frac{C_j}{N}$$

---

## Verdict

---

Faute	Coupable	Non coupable	Total
Légère	p = 0,356	p = 0,138	177
Grave	p = 0,364	p = 0,141	181
Total	258	100	358

---

**Pour chaque cellule:**

$$P(E_{ij}) = \frac{L_i}{N} \times \frac{C_j}{N}$$

**Exemple :  $(177/358) \times (258/358) = 0,356$**

---

## Verdict

---

Faute	Coupable	Non coupable	Total
Légère	127,56	49,44	177
Grave	130,44	50,56	181
Total	258	100	358

---

**Pour chaque cellule:**

$$E_{ij} = P(E_{ij}) \times N$$

**Exemple :  $0,356 \times 358 = 127,56$**

# Le test khi-carré d'indépendance

$$\left. \begin{array}{l} P(E_{ij}) = \frac{L_i}{N} \times \frac{C_j}{N} \\ E_{ij} = P(E_{ij}) \times N \end{array} \right\} \begin{array}{l} E_{ij} = \frac{L_i}{N} \times \frac{C_j}{N} \times N \\ \downarrow \\ E_{ij} = \frac{L_i \times C_j}{N} \end{array}$$

---

## Verdict

---

Faute	Coupable	Non coupable	Total
Légère	153	24	177
Grave	105	76	181
Total	258	100	358

---

$$E_{11} = \frac{177 \times 258}{358} = 127,559$$

## Verdict

**Faute**

**Coupable**

**Non coupable**

**Légère**

153 (127,56)

24 (49,441)

**Grave**

105 (130,44)

76 (50,559)

$$\chi^2_{obs} = \sum \frac{(O - E)^2}{E}$$

## Verdict

**Faute**

**Coupable**

**Non coupable**

**Légère**

153 (127,559)

24 (49,44)

**Grave**

105 (130,441)

76 (50,56)

$$= \frac{(153-127,559)^2}{127,559} + \frac{(24-49,441)^2}{49,441} + \frac{(105-130,441)^2}{130,441} + \frac{(76-50,559)^2}{50,559}$$

$$= 35,93$$

# Degrés de liberté

---

$$dl = (L - 1)(C - 1)$$

**L = le nombre de lignes de la table**

**C = le nombre de colonnes de la table**



# Le test khi-carré d'indépendance

$$\chi_{obs}^2 = 35,93$$

$$dl = (2 - 1)(2 - 1) = 1$$

$$\chi_{0,05}^2 = 3,84$$

**Puisque  $35,93 > 3,84$ , alors  $p < 0,05$  rejeter  $H_0$**

**NB:  $p = 0,0000000002$**

# Le test khi-carré d'indépendance

---

**Conclusion:**

**Les deux variables ne sont pas  
indépendantes**

**Le jugement de culpabilité est influencé par  
le fait que l'avocat du prévenu présente ou  
non la victime comme fautive**

# Les conditions d'application du test khi-carré

---

## 1. L'indépendance des observations

Le résultat d'un sujet ne doit pas influencer le résultat d'un autre

Violation la plus fréquente: les mêmes sujets sont utilisés plusieurs fois dans l'étude

# Les conditions d'application du test khi-carré

---

## 2. Les petites fréquences attendues


Si la fréquence attendue de l'une des cellules est petite, le test khi-carré risque de devenir inexact

Convention: toutes les fréquences attendues au moins égales à 5

# Les conditions d'application du test khi-carré

## 2. Les petites fréquences attendues

Que faire en cas de petite fréquence ?

- 1) Regrouper les catégories
- 2) Test exact de Fisher (pour tables 2x2 ou 2x3)
- 3) Test khi-carré d'ajustement (2 catégories)  
     test avec binomiale

# Les conditions d'application du test khi-carré

---

## 3. L'inclusion des non-occurrences

Erreur classique: oublier d'inclure les non-occurrences

Risque d'erreur grave dans la conclusion

# L'inclusion des non-occurrences

**Exemple:**

**Etude de l'influence de l'alcool sur les accidents de la route**

	Sobre	Ivre	Total
Fréquences observées	48	10	58

# L'inclusion des non-occurrences

	Sobre	Ivre	Total
Fréquences observées	48	10	58
Fréquences attendues	29	29	58

$$\chi^2_{obs} = \frac{(48 - 29)^2}{29} + \frac{(10 - 29)^2}{29} = 24,9$$



# L'inclusion des non-occurrences

$$\chi_{obs}^2 = 24,9$$

$$dl = 1$$

$$\chi_{0,05}^2 = 3,84$$

**Puisque  $24,9 > 3,84$ , alors rejeter  $H_0$**

# L'inclusion des non-occurrences

---

## Conclusion:

**Les personnes sobres commettent  
significativement plus d'accidents que les  
personnes en état d'ivresse**

# L'inclusion des non-occurrences

Exemple corrigé par l'ajout des non-occurrences

	Sobre	Ivre	Total
Accident	48	10	58
Pas d'accident	1408	42	1450
Total	1456	52	1508

# L'inclusion des non-occurrences

	Sobre	Ivre	Total
Accident	48	10	58
Pas d'accident	1408	42	1450
Total	1456	52	1508

$$P(\text{Accident} \mid \text{Sobre}) = 48 / 1456 = 0,03$$

$$P(\text{Accident} \mid \text{Ivre}) = 10 / 52 = 0,19$$

# **L'inclusion des non-occurrences**

---

**Presque tous les héroïnomanes ont  
commencé par du cannabis**

**Donc la consommation de cannabis  
conduit à consommer des drogues plus  
fortes**

# L'inclusion des non-occurrences

	Héroïne	Non héroïne	Total
Cannabis	40		
Pas cannabis	15		
Total	55		

**73% des héroïnomanes ont consommé du Cannabis**

# L'inclusion des non-occurrences

	Héroïne	Non héroïne	Total
Coca	55		
Pas coca	0		
Total	55		

**100% des héroïnomanes ont consommé du coca**

# Les mesures d'association

---

**Le test khi-carré d'indépendance ne renseigne pas sur le degré d'association de deux variables**

**Une variable peut en influencer une autre plus ou moins fortement**



# Effet du genre sur le tabagisme

	Non-fumeur	Fumeur	Total
Hommes	350	150	500
Femmes	400	100	500
Total	750	250	1000

$$\chi^2_{obs} = 13,333$$

**Hommes: 30% fumeurs**  
**Femmes: 20% fumeuses**

# Effet du genre sur le fait de faire des courses

	Courses	Non-courses	Total
Hommes	4	15	19
Femmes	15	4	19
Total	19	19	38

$$\chi^2_{obs} = 12,737$$

**Hommes: 21% courses**  
**Femmes: 79% courses**

# Le coefficient phi ( $\Phi$ )

= corrélation entre deux variables  
dichotomiques

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

Valeur entre 0 et 1

# Le coefficient phi ( $\Phi$ )

Coefficient phi pour les deux exemples:

## 1. Etude sur le tabagisme

$$\phi = \sqrt{\frac{13,333}{1000}} = 0,12$$

## 2. Etude sur les courses

$$\phi = \sqrt{\frac{12,737}{38}} = 0,58$$

# Le Phi de Cramér

= extension du phi pour des tables plus grande

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

k = nombre de catégories de la plus petite variable

# Les rapports de chance

	Crise	Pas de crise	Total
Aspirine	104	10933	11037
Placebo	189	10845	11034
Total	293	21778	22071

# Les rapports de chance

---

## Risque de faire une crise cardiaque

Groupe Aspirine:

$$104/10933 = 0,0095124$$

Groupe placebo:

$$189/10845 = 0,0174274$$

# Les rapports de chance

---

Calculer un rapport de ces deux chances

$$0,0174274/0,0095125 = 1,83$$

Une personne du groupe placebo a 1,83 fois plus de risque de faire une crise cardiaque qu'une personne du groupe aspirine