

Séance 6. Statistique bivariée

RELATION ENTRE DEUX VARIABLES QUALITATIVES

Tableau de contingence ou tri-croisé

- Tableau d'ordre 2 : croise 2 variables et permet d'étudier leur relation
- Chacune de ces 2 variables possède des modalités (catégories).
 - Les modalités de la première composent les lignes du tableau.
 - Les modalités de la seconde composent les colonnes du tableau
- Par convention : la variable **indépendante** (explicative) est présentée en **ligne** et la variable **dépendante** (à expliquer) en **colonne**

L'utilité des % (vs. effectifs)

- **Les jeunes (18-24 ans) votent-ils :**
 - plus ou moins que l'ensemble de l'électorat pour Nicolas Sarkozy ?
 - plus ou moins que les plus de 60 ans ?

- **L'électorat de Ségolène Royal est-il :**

- plus jeune que l'électorat en général ?
- plus jeune que celui de Nicolas Sarkozy ?

-> **Délicat à lire avec les seuls effectifs**

age5	votp2_8807		Total
	gauche	droite	
18-24 ans	245	140	385
25-34 ans	326	301	627
35-44 ans	326	300	626
45-59 ans	410	406	816
60 et+	366	613	979
Total	1,673	1,760	3,433

Tableaux de pourcentages en ligne

age5	votp2_8807		Total
	gauche	droite	
18-24 ans	245 63.64	140 36.36	385 100.00
25-34 ans	326 51.99	301 48.01	627 100.00
35-44 ans	326 52.08	300 47.92	626 100.00
45-59 ans	410 50.25	406 49.75	816 100.00
60 et+	366 37.39	613 62.61	979 100.00
Total	1,673 48.73	1,760 51.27	3,433 100.00

Formule ?

(effectif case / effectif marginal ligne)*100
 $(245/385)*100 = 63,64\%$

Intérêt ?

On observe que le
 vote pour Nicolas
 Sarkozy augmente
 avec l'âge

Tableaux de pourcentages en colonne

Formule : (effectif case/effectif marginal colonne)*100
(366/1673)*100 = 21,88%

age5	votp2_8807		Total
	gauche	droite	
18-24 ans	245 14.64	140 7.95	385 11.21
25-34 ans	326 19.49	301 17.10	627 18.26
35-44 ans	326 19.49	300 17.05	626 18.23
45-59 ans	410 24.51	406 23.07	816 23.77
60 et+	366 21.88	613 34.83	979 28.52
Total	1,673 100.00	1,760 100.00	3,433 100.00

Clef de lecture : sur 100 votants pour Ségolène Royal, 21,88 sont sexagénaires et plus

Intérêt : l'électorat de **Ségolène Royal** se compose à plus de **53% de moins de 44 ans** (42% pour Nicolas Sarkozy)

Tableaux de pourcentage total (= en ligne et en colonne)

age5	votp2_8807		Total
	gauche	droite	
18-24 ans	245 7.14	140 4.08	385 11.21
25-34 ans	326 9.50	301 8.77	627 18.26
35-44 ans	326 9.50	300 8.74	626 18.23
45-59 ans	410 11.94	406 11.83	816 23.77
60 et+	366 10.66	613 17.86	979 28.52
Total	1,673 48.73	1,760 51.27	3,433 100.00

Formule ?

$(\text{effectif cas} / \text{effectif total}) * 100$
 $(300/3433)*100 = 8,74\%$

Intérêt assez faible ...

Comparaison de pourcentages : les écarts à la moyenne

age5	Gauche	Droite	Total	age5	Gauche	Droite	age5	Gauche	Droite
18-24 ans	64	36	100%	18-24 ans	64-49	36-51	18-24 ans	15	-15
25-34 ans	52	48	100%	25-34 ans	52-49	48-51	25-34 ans	3	-3
35-44 ans	52	48	100%	35-44 ans	52-49	48-51	35-44 ans	3	-3
45-59 ans	50	50	100%	45-59 ans	50-49	50-51	45-59 ans	1	-1
60 et+	37	63	100%	60 et+	37-49	63-51	60 et+	-12	12
Total	49	51	100%	Total	49	51	Total	49	51

Interprétation :

Les moins de 25 ans ont voté davantage pour Ségolène Royal que la moyenne (+15 points) tandis que les personnes âgées de plus de 60 ans ont moins voté pour Ségolène Royal que la moyenne (-12 points).

Chacune des cases contient la différence, en points de pourcentage entre la fréquence conditionnelle pour la case considérée et la fréquence moyenne
 $\text{Écart} = \text{fréquence conditionnelle} - \text{fréquence moyenne}$



Test du χ^2 (chi2, khi2, chi-deux, chi-square, etc.)

- Comment savoir si les **différences observées** dans notre tri-croisé sont **significatives** ?
- Sont-elles dues au hasard de l'échantillonnage ? ...
- ... ou peut-on **inférer** les écarts observés dans l'échantillon à la population d'intérêt ?
- Le test du χ^2 permet de trancher

Les tests statistiques

- Le principe général :
 - On formule une **hypothèse « nulle » (H0)** : il n'y a pas de lien entre les deux variables = les deux variables sont indépendantes
 - On cherche à rejeter cette hypothèse nulle (avec un risque d'erreur)
- Dans la pratique pour le χ^2 :
 - On calcule un score (un nombre)
 - On compare ce nombre calculé à un nombre théorique
 - On conclue

Spoiler alert !

- $\chi^2 = \sum \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$
- Dans la vraie vie vous n'aurez pas à le faire à la main
- Jamovi (ou autre) bossera pour vous pour la partie calcul
- Mais il est important de comprendre le principe général et la logique pour savoir comment interpréter les résultats et en conclure quelque chose de pertinent

Hypothèse nulle et test du χ^2

- Hypothèse nulle (H_0) : les deux variables X et Y sont indépendantes
 - = « absence de lien entre les deux variables »
- Comparaison entre les effectifs **observés** et les effectifs **théoriques**
 - = ce qu'on aurait observé dans le tri croisé si les variables étaient totalement indépendantes
 - comment on fait ?
- Petit point vocabulaire et synonymes
 - Effectif observé, fréquence observée (effectifs observés dans Jamovi)
 - Effectif théorique/attendu ; fréquence théorique/attendue (quantités attendues dans Jamovi)

Le test du χ^2

Effectifs observés :

	Réussite au bac	Echec au bac	Total
Parents diplômés du supérieur	160	40	200
Parents non diplômés du sup	100	100	200
Total	260	140	400

Mais quels sont les effectifs théoriques (cf. l'hypothèse nulle) ?


Le test du χ^2 : effectifs théoriques

	Réussite au bac	Echec au bac	Total
Parents diplômés du supérieur			200
Parents non diplômés du sup			200
Total	260	140	400

Effectifs théoriques :

	Réussite au bac	Echec au bac	Total
Parents diplômés du supérieur	130	70	200
Parents non diplômés du sup	130	70	200
Total	260	140	400

hypothèse nulle : il devrait y avoir autant de bacheliers chez les enfants de parents diplômés du supérieur que chez les parents non diplômés du supérieur



Pour chaque case on multiplie les marges puis on les divise par l'effectif total

= (marge colonne * marge ligne) / effectif total

= (260 * 200) / 400 = 130

Le test du χ^2 : calcul

	Réussite au bac	Echec au bac	Total
Parents diplômés du supérieur	160 130	40 70	200
Parents non diplômés du sup	100 130	100 70	200
Total	260	140	400

Effectifs observés
et théoriques

Formule du χ^2 :

$$\chi^2 = \sum \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

Dans notre cas :

$$\begin{aligned} & \frac{(160-130)^2}{130} \oplus \frac{(40-70)^2}{70} \oplus \frac{(100-130)^2}{130} \oplus \frac{(100-70)^2}{70} \\ &= 30^2/130 + (-30^2)/70 + (-30^2)/130 + 30^2/70 \\ &= 6,9 + 12,8 + 6,9 + 12,8 = \mathbf{39,2} \end{aligned}$$

La table du χ^2

- On compare le χ^2 calculé à un χ^2 critique lu dans la table de distribution du Chi2
 - <https://www.chisquaretable.net/>
- Valeur critique **dépend de la taille du tableau** croisé
 - « degrés de libertés » : ddl = (nb lignes - 1)*(nb colonnes - 1)
 - dans notre exemple, ddl = (2-1)*(2-1) = 1
- La table nous donne le X^2 minimum pour que l'on puisse rejeter l'hypothèse nulle d'indépendance avec un seuil de risque défini (généralement $\leq 5\%$ en shs)
- Si X^2 calculé > X^2 théorique = rejet de l'hypothèse nulle / existence d'un lien statistique

	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697

Conclure

- Dans notre exemple : $39,2 > 3,841$, on rejette donc l'hypothèse nulle au seuil de risque de 5%
- Il existe un lien statistiquement significatif entre nos variables
- = Il existe un lien entre le niveau de diplôme des parents et la réussite au baccalauréat

Les limites du χ^2

- Plus l'échantillon est large, plus on a de chances d'avoir des relations significatives
- Significativité **substantielle** et la significativité **statistique** :
 - Dépend de la taille d'échantillon et du nombre de cases du tableau
 - Renvoie à la question de toute recherche sociologique : il faut une théorie derrière un tableau
 - Applicable à tous les travaux et toutes les méthodes statistiques
- Les variables « cachées » (aussi appelées spurious correlation)
- Le test du χ^2 ne permet pas de conclure sur **l'intensité** du lien entre les variables

Le V de Cramer

- Le **V de Cramer** permet de pallier ces limites :
 - Ne dépend pas des effectifs et du nombre de cases d'un tableau

$$\sqrt{\frac{\chi^2}{n \times [\min(l, c) - 1]}}$$

m = le m minimum (nombre de modalités var1 et var2) - 1
ou l, c - 1
n = l'effectif total

- Varie de 0 à 1
 - 0 = pas de lien
 - 1 = lien parfait
 - Devient intéressant vers 0,15
- Permet donc de hiérarchiser des liens statistiques

Le V de Cramer

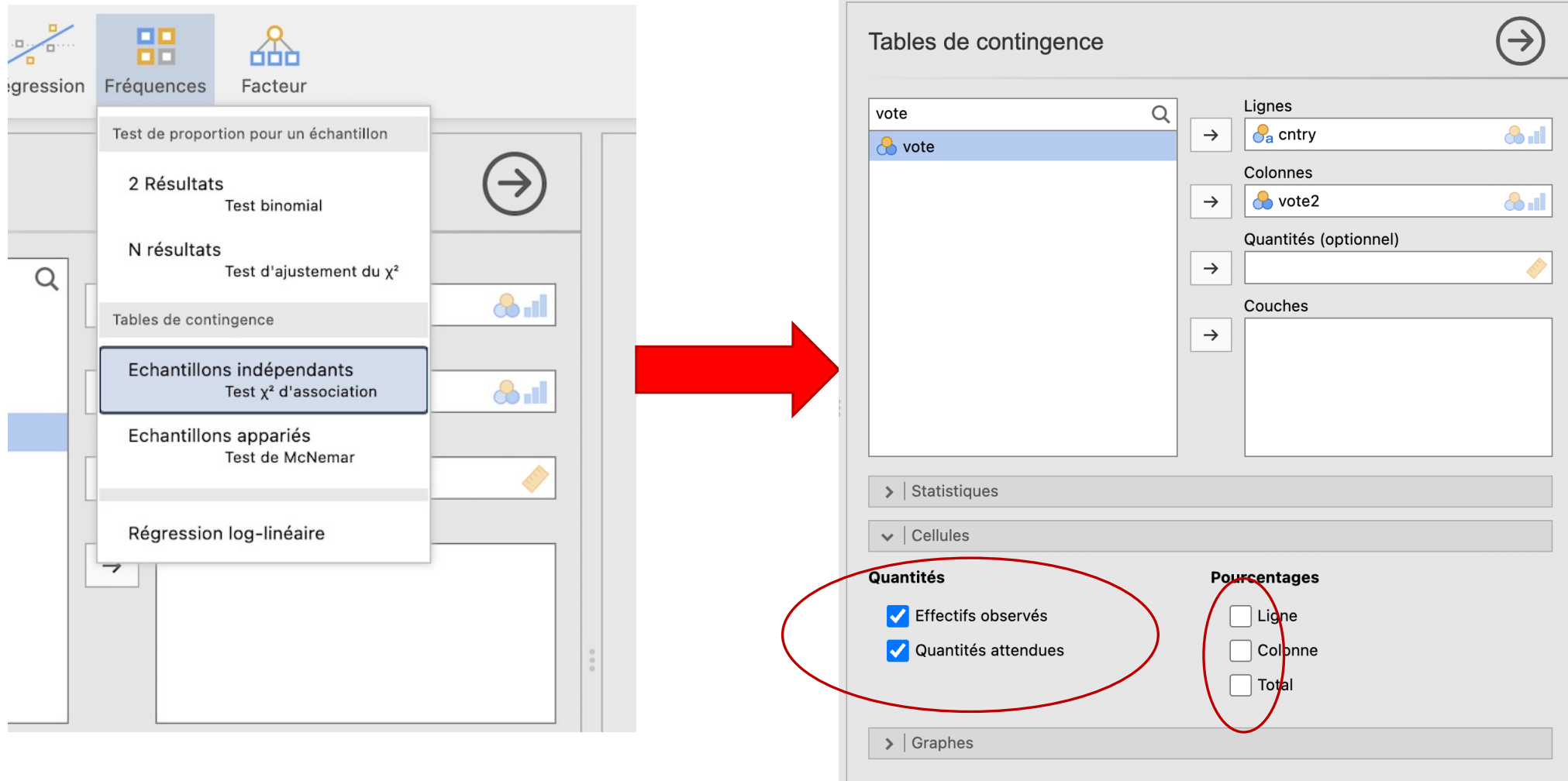
Retour à l'exemple des résultats au bac :

$$\begin{aligned} V &= \sqrt{X^2 / (m \text{ min} - 1) n \text{ total}} \\ &= \sqrt{39,2 / (1 \times 400)} \\ &= \sqrt{39,2 / 400} \\ &= \sqrt{0,098} \\ &= 0,31 \end{aligned}$$

→ La relation entre le niveau de diplôme des parents et la réussite au bac est **significative**

- Retrouver tout ça dans Jamovi
- L'appliquer à vos données

Et dans Jamovi ?



Regression Fréquences Facteur

Test de proportion pour un échantillon

2 Résultats
Test binomial

N résultats
Test d'ajustement du χ^2

Tables de contingence

Echantillons indépendants
Test χ^2 d'association

Echantillons appariés
Test de McNemar

Régression log-linéaire

Tables de contingence

vote

vote

Lignes
→ cntry

Colonnes
→ vote2

Quantités (optionnel)
→

Couches
→

> | Statistiques

▼ | Cellules

Quantités

☒ Effectifs observés

☒ Quantités attendues

Pourcentages

☐ Ligne

☐ Colonne

☐ Total

> | Graphes

Statistiques

Tests

☒ χ^2
☐ Correction de continuité du χ^2
☐ Ratio de vraisemblance
 ☐ Test exact de Fisher
 ☐ test z pour la différence entre deux proportions

Hypothèse

☒ Groupe 1 \neq Groupe 2
 ☐ Groupe 1 > Groupe 2
 ☐ Groupe 1 < Groupe 2

Nominal

☐ Coefficient de contingence
 ☒ V de Phi et Cramer

Ordinal

☐ Gamma
 ☐ Tau b de Kendall
 ☐ Mantel-Haenszel

Quantités

☒ Effectifs observés

Mesures comparatives (2x2 seulement)

☐ Rapport des cotes (odds ratio)
 ☐ Log du rapport des cotes (odds ratio)
 ☐ Risque relatif
 ☐ Différence entre les proportions
 ☒ Intervalles de confiance

Intervalle

95 %

Comparer

lignes

Cellules

Pourcentages

☐ Ligne

PI	Observe	142	220	970
	Attendu	761	209	970
RS	Observé	1454	448	1902
	Attendu	1493	409	1902
SE	Observé	1398	73	1471
	Attendu	1155	316	1471
SI	Observé	830	417	1247
	Attendu	979	268	1247
SK	Observé	714	332	1046
	Attendu	821	225	1046
Total	Observé	35505	9728	45233
	Attendu	35505	9728	45233

Tests χ^2

	Valeur	ddl	p
χ^2	2016	28	<.001
N	45233		

Nominal

	Valeur
Coefficient Phi	NaN
V de Cramer	0.211