Analyse statistique - Séance 6 : Les tests d'hypothèse et la statistique du Khi2

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2025

Quizz sur l'inférence

1. Les enquêtes par sondage et les enquêtes par questionnaires correspondent-elles à deux choses différentes ?
2. Pour une enquête visant à estimer la fréquence à laquelle les lyonnais·es mangent au restaurant, est-ce une bonne idée de sélectionner aléatoirement 1/10 des restaurants de la ville et demander à leurs clients combien de fois ils se rendent au restaurant par semaine ?
3. Un biais de sélection correspond à la situation où les personnes recrutées pour une enquêtes ne sont pas représentatives de la société dans son ensemble : vrai ou faux ?
4. Expliquez ce qu'on entend par un échantillon "représentatif"
5. Expliquez quelle est la différence entre une méthode d'échantillonnage par quota et une méthode d'échantillonnage aléatoire
6. Est-ce qu'il serait facile de faire un tirage aléatoire dans une manifestation ? Comment procéderiez-vous ?
7. Dans le cadre de la statistique inférentielle, l'intervalle de confiance permet de quantifier l'incertitude : expliquez cette phrase
8. La précision de mon estimation dépend de la taille de ma population : vrai ou faux ?

Principe

Les *tests statistiques* sont des procédures qui permettent d'éprouver nos *hypothèses*, c'est-à-dire d'estimer la probabilité de se tromper en affirmant quelque chose puis de choisir s'il faut *rejeter* ou *accepter* notre hypothèse.

C'est une démarche qui reposent sur des outils, les *statistiques de test*, et des critères de décision, les *seuils de risque* considéré comme acceptable.

> *J'ai une probabilité de 10% de me tromper en affirmant que la durée de maladie est plus courte pour mes patient·es traité·es par mon médicament que pour mes patient·es témoins (non traité·es).*
> *Ma probabilité de me tromper étant supérieure à mon seuil de 1%, je ne commercialise pas mon médicament.*

Les types d'erreur

Il existe deux types d'erreur : l'erreur de type I quand accepte l'hypothèse alors qu'elle était fausse; l'erreur de type II quand on rejette l'hypothèse alors qu'elle était vraie.

> *Il est plus grave de donner un traitement inefficace (type I) à une personne malade que de ne pas lui donner un traitement qui aurait pu la guérir (type II).*

Les types de test

Il existe de très nombreux types de tests qui s'appliquent à des paramètres pour vérifier que :

- La moyenne (par ex. d'âge) est différente entre deux groupes
- La variance (par ex. d'âge) est différente entre deux groupes
- La proportion (par ex. la part de propriétaires) est différente entre deux groupes
- Le ratio est différent (par ex. le PIB par tête) est différent entre deux groupes

Tester l'interdépendance ou la corrélation

Il existe finalement des tests statistiques pour juger, **à partir de notre échantillon**, de la *corrélation* (variables quantitatives) ou de *l'interdépendance* (variables qualitatives) entre deux dimensions sociales **dans la population**.

Quand les niveaux de corrélation ou d'interdépendance *observés* sont suffisamment importants, on conclut que ce qu'on a observé est *statistiquement significatif*, c'est-à dire que ce n'est pas le fruit du hasard.

> *Les différences dans les loisirs déclarés entre les seniors et les jeunes de notre échantillon sont assez importantes pour juger que ces deux catégories n'ont pas, dans la population, les mêmes préférences de loisir.*

Principe

Le test du Khi2 sert à tester la dépendance entre des variables catégorielles à plusieurs modalités.

Il repose sur une idée simple : si les effectifs du tri croisé de deux variables sont très différents des effectifs qu'on aurait dû observer si les deux variables étaient indépendantes, alors on peut raisonnablement conclure qu'elles n'ont pas d'effets l'une sur l'autre.

Voilà pourquoi on part d'abord du *tableau de contingence* qui présente les *effectifs observés* du croisement de deux variables.

Le tableau de contingence

Dans notre échantillon, 33 individus de 18 à 30 ans n'ont pas de frères et soeurs contre 72 de 31 à 60 ans.

Tableau de contingence : effectifs observés

	Type de famille			
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs	Ensemble
18 à 30 ans	33	242	76	351
31 à 60 ans	72	616	473	1 161
61 ans et plus	62	260	166	488
Ensemble	167	1 118	715	2 000

Il faut adopter un raisonnement conditionnel pour interpréter ces effectifs observés : si les deux variables étaient indépendantes, quels seraient nos effectifs ?

Les effectifs marginaux

Ce sont les *effectifs marginaux* (les totaux d'ensemble) qui nous permettent de calculer les valeurs obtenues en regroupant tous les individus, peu importe leur catégorie d'âge.

Tableau de contingence : effectifs observés

	Type de famille			
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs	Ensemble
18 à 30 ans	33	242	76	351
31 à 60 ans	72	616	473	1 161
61 ans et plus	62	260	166	488
Ensemble	167	1 118	715	2 000

Les effectifs théoriques

Le tableau des *effectifs théoriques* rend compte de la situation qu'on observerait **en cas d'indépendance entre les deux variables**.

- On sait que la catégorie d'âge 18 à 30 ans représente $\frac{351}{2000} = 17,55\%$ de la population.
- On sait que les familles uniques représentent $\frac{167}{2000} = 8,35\%$ de la population.
- Par déduction, parmi les 18-30 ans, soit $17,55\%$ des 2000 individus, il devrait y en avoir $8,35\%$ qui appartiennent à une famille unique. On obtient 29 individus.

Tableau de contingence : effectifs théoriques

	Type de famille			
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs	Ensemble
18 à 30 ans	29	196	125	351
31 à 60 ans	97	649	415	1 161
61 ans et plus	41	273	174	488
Ensemble	167	1 118	714	2 000

Les écarts au carré rapportés à l'effectif théorique

Il faut maintenant mesurer à quel point ces deux tableaux diffèrent : on calculera simplement des écarts.
Ainsi, la valeur du Khi2 est donnée par la formule :

$$\chi^2 = \sum_{i,j} \frac{(Eff_{i,j}^{theo} - Eff_{i,j}^{obs})^2}{Eff_{i,j}^{theo}}$$

Tableau de contingence : écarts au carré rapportés à l'effectif théorique

	Type de famille			
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs	Ensemble
18 à 30 ans	$(29 - 33)^2/29$	$(196 - 242)^2/196$	$(125 - 76)^2/125$	351
31 à 60 ans	$(97 - 72)^2/97$	$(649 - 616)^2/649$	$(415 - 473)^2/415$	1 161
61 ans et plus	$(41 - 62)^2/41$	$(273 - 260)^2/273$	$(174 - 166)^2/174$	488
Ensemble	167	1 118	715	2 000

Les éléments du test

- La *statistique de test* χ^2 donne donc une mesure de l'écart à l'indépendance : jusqu'à quel point ce qu'on observe est éloigné d'une situation où il n'y aurait pas de lien entre les variables ?
- Le *test statistique* part de la valeur du χ^2 et la compare avec les écarts qu'on considérerait comme étant le fruit du hasard. Grâce à la loi Normale, on sait modéliser l'aléa de sous forme de probabilités, donc le risque de se tromper en concluant quelque chose à partir des écarts observés.
- L'*hypothèse nulle* du test est celle qu'on essaie souvent de rejeter : la catégorie d'âge et le type de famille n'ont pas de lien entre eux.
- L'*hypothèse alternative* est celle qu'on accepte quand l'hypothèse nulle est rejetée : la catégorie d'âge et le type de famille sont bien interdépendants.

La notion de p-value par rapport au seuil de risque

La *p-value*, c'est le plus petit *niveau de risque* tel qu'on peut encore rejeter l'hypothèse nulle.

Au vu de la grandeur ou de la petitesse des écarts à l'indépendance mesurés par le χ^2, est-ce qu'en affirmant qu'il y a un lien entre les variables je risque de me tromper dans 50%, dans 10% ou dans seulement 1% des cas ?

En général, on prend comme niveau de risque acceptable 10% pour affirmer que la relation est significative. C'est un seuil arbitraire qui s'est cristallisé dans la pratique mais il n'y a aucune justification mathématique à ce niveau.

Ce que n'est pas la p-value

- Quand **p** est supérieur au seuil, c'est soit qu'il y a interdépendance, **soit que votre échantillon était trop petit pour conclure quelque chose sur l'interdépendance des variables.**
- **p** ne mesure pas l'intensité de la relation entre les deux variables, seulement la significativité du lien.
- **p** n'est pas la probabilité que les deux variables soient interdépendantes. C'est seulement la probabilité de se tromper en affirmant cela à partir de ce que vous avez observé !

L'analogie du procès

Le test du khi2 fonctionne comme un procès. Le prévenu est présumé innocent (variables indépendantes) jusqu'à preuve du contraire, c'est-à-dire si assez d'éléments convaincants ont été réunis par l'accusation (χ^2 élevée) pour le déclarer coupable (rejet de l'hypothèse d'indépendance).

Si la personne n'est pas déclarée coupable, c'est soit car elle était réellement innocente (χ^2 faible car les variables sont indépendantes dans la population), soit que l'enquête n'a pas permis de réunir assez de preuves (χ^2 faible car l'échantillon est trop petit).

P-value inférieure à notre seuil de risque de 10%

- Si la p-value est comprise entre $]0; 0,10[$.
- La p-valeur est extrêmement élevée. Il y a plus de 95% de chances de se tromper en affirmant que le sexe et la structure familiale sont interdépendants. On est dans une sorte d'impasse : **on ne peut pas rejeter l'hypothèse nulle, mais on ne peut pas dire qu'elle est vérifiée !**
- Un plus grand échantillon pourrait être capable d'appréhender des effets subtiles du sexe des enfants sur les comportements de fécondité des parents, par exemple si les parents cherchent à avoir un nouvel enfant quand le premier est une fille. Cela établirait in fine un lien entre ces deux variables qui n'est pas perceptible avec seulement 2000 enquêté·es. En l'état, on ne peut donc rien dire.

Guide d'interprétation

Les conclusions tirées des tests d'hypothèse ne concernent donc que les possibilités d'inférence de nos données. Il faut toujours se demander :

- La (non)-significativité de mes résultats est-elle surtout le fruit des catégories choisies, de la taille de mon échantillon et de la façon dont il a été constitué, ou est-elle vraisemblable d'un point de vue sociologique ?
  Exemple : une enquête réalisée sur internet interroge le lien entre l'âge et le fait de posséder un portable et un ordinateur. Elle trouve que les senior ne détiennent pas significativement moins d'ordinateurs et de portables que les plus jeunes.
- Est-ce que le lien statistique mis en lumière suggère un lien de causalité entre ces variables, qui a un vrai sens sociologique, ou ai-je seulement identifié une corrélation peu intéressante, qui cache une troisième variable explicative, voire fallacieuse, le lien n'étant que le pur fruit du hasard ?
  Exemple : une enquête par questionnaire cherche à savoir si les parisien·nes ont plus souvent des accidents de voiture que les habitant·es du reste du territoire français. Elle trouve au contraire que les parisien·nes sont sous-représenté·es parmi les accidents mineurs et graves.
- Quelles sont les connaissances sociologiques et les résultats empiriques d'autres recherches qui peuvent expliquer le lien d'interdépendance indiqué par les tests ?