Comparaison de fréquences et de distributions

Tests visant à mettre en évidence une corrélation entre deux variables qualitatives

M. L. Delignette-Muller
VetAgro Sup

5 octobre 2020

VetAgro Sup

UNIVERSITÉ DE LYON

Objectifs pédagogiques

- Savoir repérer dans quels cas on doit utiliser un test du χ^2 d'ajustement, un test du χ^2 d'indépendance, un test de McNemar, et un test de Cochran-Mantel-Haenszel.
- Savoir vérifier les conditions d'utilisation de ces tests et en interpréter les résultats.
- Savoir réaliser à la main les tests du χ^2 (ajustement et indépendance) et le test de McNemar.*

* *savoir faire évalué uniquement en S5*

Plan

Exemple de comparaison d'une fréquence observée à une fréquence théorique

Sur un échantillon aléatoire de 15 étudiants vétérinaires on compte 4 garçons et 11 filles.
Peut-on dire qu'il y a plus de filles que de garçons dans la population des étudiants vétérinaires ?
La proportion de filles est-elle significativement différente de 50% ?

Calcul des effectifs théoriques sous H_0

H_0 : " différence nulle c'est-à-dire $p_{filles} = 0.5$"

- **Effectifs observés** notés O_i :
  filles 11, garçons 4
- **Effectifs théoriques** (attendus sous H_0) **notés** C_i :
  filles 7.5, garçons 7.5

On veut comparer les O_i et les C_i.
Il nous faut définir une variable de décision pour faire un test.
Quelle variable ?

Le test du χ^2 d'ajustement

- **Variable de décision**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i} = \left(\sum_{i=1}^{k} \frac{O_i^2}{C_i}\right) - N$$

avec

  - k le nombre de classes de la variable qualitative (ici 2, filles et garçons),
  - N le nombre total d'observations (ici 15).

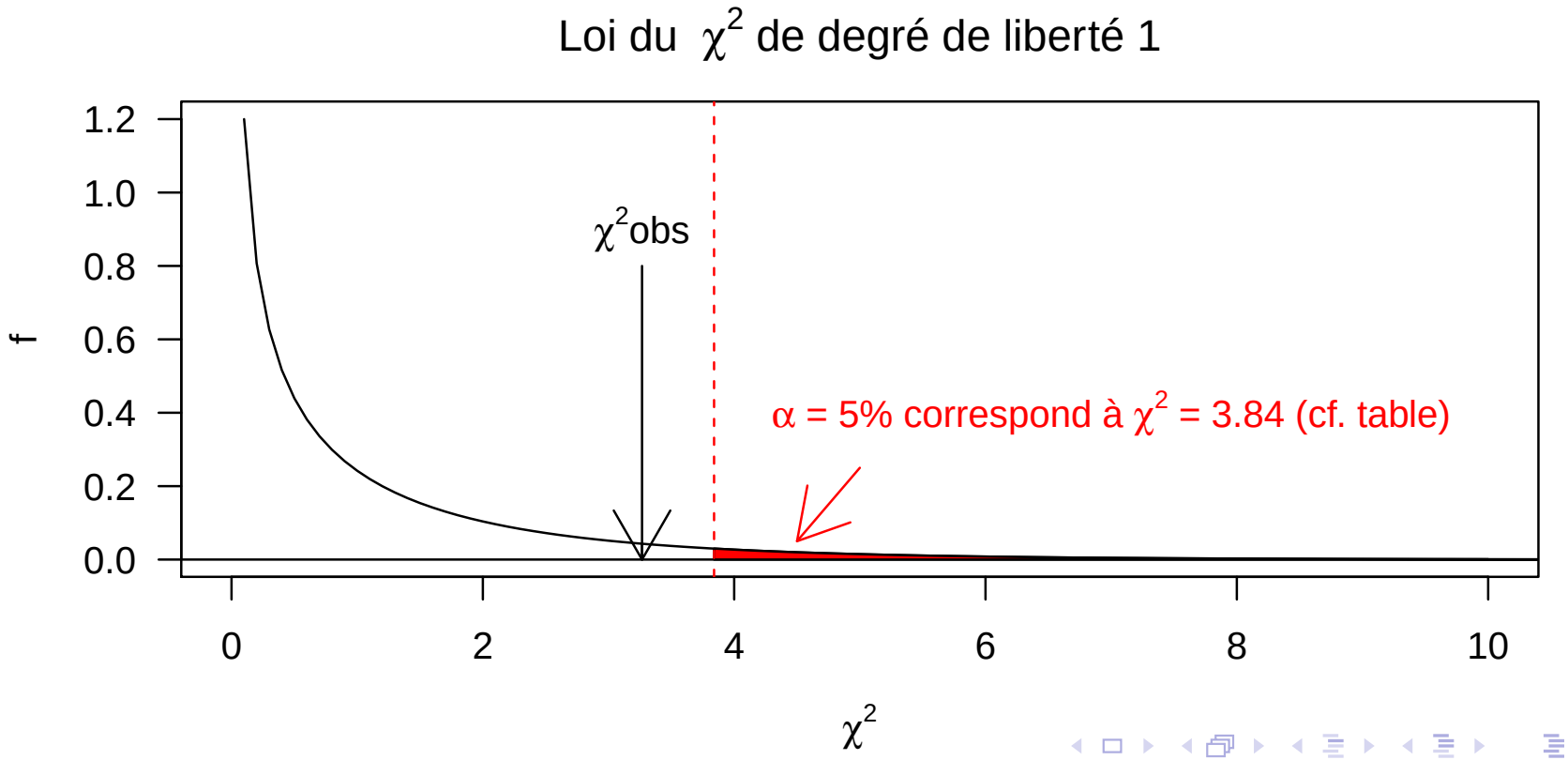
- **Conditions d'utilisation du test**
  Si tous les C_i sont supérieurs à 5 (c'est le cas ici), on peut considérer que la variable de décision suit à peu près la **loi du χ^2 de degré de liberté $k-1$.**

Test du χ^2 : calcul de la valeur de p

$\chi^2_{obs} = 3.267$ correspondant à $p > 0.05$.

La différence observée n'est pas significative.

On ne peut pas conclure à une proportion plus importante de filles à partir de ce seul échantillon.



Cadre d'utilisation du test du χ^2 d'ajustement

- Dans le cas de la comparaison d'une fréquence observée à une fréquence théorique, ce test du χ^2 est strictement équivalent au test utilisant la loi normale (cf. cours d'introduction aux tests).
- Ce test a un cadre d'utilisation plus large : il permet de comparer une distribution observée d'une variable qualitative (quelle que soit le nombre k de classes) à une distribution théorique.

Exemple de comparaison de plusieurs fréquences observées sur des échantillons indépendants

Sur un échantillon de 999 chiennes d'élevage on voudrait savoir **si la fréquence d'intervention** de l'éleveur ou du vétérinaire pendant leur mise-bas **dépend de la taille des races**
(4 groupes : "races géantes (XL)", "grandes races (L)", "races moyennes (M)" et "petites races (S)")),
Autrement dit,
les **fréquences d'intervention sont-elles différentes** entre les 4 groupes de taille de race,
ou encore
la variable intervention est-elle corrélée à la variable taille de race ?

(données extraites de la thèse vétérinaire de Mathilde Poinssot, Maisons Alfort, 2011)

Les données observées

Table de contingence

Taille de race Intervention	XL	L	M	S	Total
NON	29	183	146	170	528
OUI	62	134	107	168	471
Total	91	317	253	338	999

Fréquences observées

Taille de race	XL	L	M	S
% d'interventions	68.1	42.3	42.3	49.7

Calcul des effectifs théoriques sous H_0

H_0 : " différence nulle entre les fréquences" ou encore "indépendance entre les variables (intervention et taille de race)"

Calcul des effectifs théoriques C_{ij} à partir des totaux $C_{i.}$ et $C_{.j}$

Sous H_0, les probabilités marginales et conditionnelles sont les mêmes, c'est-à-dire $\frac{C_{ij}}{C_{i.}} = \frac{C_{.j}}{N}$ donc $C_{ij} = C_{i.} \times \frac{C_{.j}}{N}$

Exemple de calcul pour une cellule :

Taille de race / **Intervention**	**XL**	**L**	**M**	**S**	Total
NON					528
OUI		471 × 317 / 999			471
Total	91	317	253	338	999

Le test du χ^2 d'indépendance

- **Variable de décision**

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - C_{ij})^2}{C_{ij}} = \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{O_{ij}^2}{C_{ij}}\right) - N$$

avec

  - k le nombre de lignes de la table de contingence (nombre de classes de la variable en ligne),
  - l le nombre de colonnes de la table de contingence (nombre de classes de la variable en colonne),
  - et N le nombre total d'observations.

- **Conditions d'utilisation du test**

Si tous les C_{ij} sont supérieurs à 5, on peut considérer que la variable de décision suit à peu près la **loi du χ^2 de degré de liberté** $(k-1)\ (l-1)$.

Réalisation du test sur l'exemple

Effectifs théoriques

Taille de race Intervention	XL	L	M	S	Total
NON	48.1	167.5	133.7	178.6	528
OUI	42.9	149.4	119.3	159.4	471
Total	91	317	253	338	999

On peut vérifier ici les conditions d'utilisation du test du χ^2 d'indépendance (effectifs théoriques tous > 5)

Statistique du χ^2

$\chi^2_{obs} = 22.38$ correspondant à $p < 0.001$.

La différence observée entre les fréquences d'intervention est significative, autrement dit il y a une corrélation significative entre la taille des races et l'intervention lors de la mise bas.

Cadre d'utilisation du test du χ^2 d'indépendance

- Comparaison de deux fréquences observées sur des échantillons indépendants
  (dans ce cas le test du χ^2 est strictement équivalent au test utilisant la loi normale)
- Comparaison de plusieurs fréquences observées sur des échantillons indépendants
  (cf. exemple précédent)
- Corrélation entre deux variables qualitatives observées sur les individus d'un échantillon
  (exemple historique exposé par Karl Pearson : corrélation entre la couleur des cheveux et la couleur des yeux)

Plan

Exemple de comparaison de 2 fréquences observées sur de échantillons appariés

On dispose de 2 tests A et B pour détecter la présence d'une maladie donnée chez des souris.
Les 2 tests **sont utilisés en parallèle sur 100 souris** que l'on sait malades de façon certaine.
On souhaite comparer les sensibilités (probabilité de réponse positive chez un malade) des 2 tests.

Les données observées

Table de concordance $\neq$ table de contingence

Résultat du test B **Résultat du test A**	**positif**	**négatif**
positif	70	6
négatif	18	6

Fréquences observées

Test	**A**	**B**
Sensibilité en %	76	88

Principe du test de McNemar

Le test de McNemar **se base uniquement sur les nombres de résultats discordants**, et **compare les deux types de discordances**,
c'est-à-dire les nombres de résultats A+B- (ici 6) et A-B+ (ici 18) (cf. fiche technique pour sa réalisation).
Comparer ces 2 nombres revient bien à comparer les sensibilités des tests.
ATTENTION,
ce test permet uniquement de comparer 2 fréquences
et en aucun cas de juger de la concordance entre les tests, ce pour quoi il faudrait utiliser aussi le nombre de résultats concordants

Exemple de comparaison de plusieurs fréquences observées sur des échantillons dépendants

On dispose de 3 tests A, B et C pour détecter la présence d'une maladie donnée chez des souris.
Les 3 tests **sont utilisés en parallèle sur 100 souris** que l'on sait malades de façon certaine.
On souhaite comparer les sensibilités (probabilité de réponse positive chez un malade) des 3 tests.

Test de Cochran-Mantel-Haenszel

Les données sont plus difficiles à résumer que dans le cas de 2 fréquences.

Codage des données brutes (1 si détecté, 0 si non détecté) :

Identifiant de la souris	**test A**	**test B**	**test C**
S1	0	0	0
S2	1	0	1
S3	0	0	1
S4	1	1	1
S5	0	0	0
...	...	...	...

Test de Cochran (facilement réalisable avec **R**) :
extension du test de McNemar.

Comment choisir le bon test ?

- **Un seul échantillon**
  test du χ^2 d'ajustement de comparaison d'une fréquence observée à une fréquence théorique ou d'une distribution observée à une distribution théorique
- **Deux ou plusieurs échantillons indépendants**
  test du χ^2 d'indépendance de comparaison de plusieurs fréquences observées ou plusieurs distributions observées
- **Deux échantillons dépendants (appariés)**
  test de McNemar de comparaison de deux fréquences observées
- **Plusieurs échantillons dépendants**
  test de Cochran-Mantel-Haenszel de comparaison de plusieurs fréquences observées

Les conditions d'utilisation des tests

Bien vérifier les conditions d'utilisation des tests, notamment dans le cadre des tests du χ^2 (effectifs théoriques supérieurs à 5).
Lorsque ce n'est pas le cas, il est parfois possible d'utiliser une **statistique corrigée** (χ^2 avec correction de Yates dans le cadre de la comparaison de 2 fréquences faite systématiquement par **R**), ou de faire un **calcul exact de la valeur de** p (facilement réalisable avec **R** dans le cas de la comparaison de deux fréquences).
Sur les grandes tables de contingence, on procède parfois à des **regroupements de classes** pour satisfaire les conditions d'utilisation.