Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

# Cours 7
# Test d'indépendance statistique du $\chi^2$

**1** Dans le cours précédent, nous avons étudié expérimentalement les variations dues aux fluctuations d'échantillonnage ; nous avons remarqué qu'en supposant les variables X et Y indépendantes, les effectifs observés n'étaient jamais identiques aux effectifs théoriques ; dit autrement, le $\chi^2$ d'une observation conjointe de deux variables indépendantes sur un échantillon quelconque est un nombre positif plus ou moins proche de 0, mais jamais nul. Il s'ensuit qu'en pratique, l'indépendance ne s'observe jamais sur un échantillon.

Pour sortir de cette impasse, nous allons introduire une seconde forme de l'indépendance, moins naturelle, mais plus opératoire, l'indépendance statistique, équivalente à l'indépendance si le phénomène de fluctuation d'échantillonnage n'existait pas ; et à défaut d'observer *l'indépendance* de deux variables nous chercherons à apprécier leur *indépendance statistique*.

**2** déf **Deux variables conjointes X et Y sont dites *statistiquement* indépendantes** (pour un échantillon) si les différences entre effectifs observés et effectifs théoriques sont suffisamment réduites pour être considérées comme le seul effet des fluctuations d'échantillonnage ; autrement dit, si la valeur du $\chi^2$ calculée est suffisamment proche de 0 pour être considérée comme le seul fait des fluctuations d'échantillonnage.

**3** **Valeurs théoriques du $\chi^2$.** Pour rendre cette définition utilisable, il faut déterminer ces valeurs proches de 0 ; l'idée est simple dans son principe : on compare la valeur du $\chi^2$ de l'observation aux valeurs théoriques du $\chi^2$ qu'on observerait sur des échantillons de même taille *en supposant X et Y indépendantes*, reflet des variations dues aux fluctuations d'échantillonnage ; si elle fait partie des petites valeurs théoriques vraisemblables, nous estimerons que tout se passe comme si l'échantillon observé faisait partie des échantillons observés *sous l'hypothèse d'indépendance*, et nous pourrons décider de traiter X et Y comme deux variables statistiquement indépendantes ; si par contre cette valeur est supérieure aux valeurs théoriques, ou même aux grandes valeurs théoriques très improbables, nous devrons en conclure que les fluctuations à elles seules n'expliquent pas cette valeur, et que l'hypothèse d'une liaison statistique entre X et Y s'impose.

Nous allons étudier ces valeurs théoriques de deux manières : en simulant des observations conjointes de deux variables supposées indépendantes, et en utilisant des résultats de travaux mathématiques ; nous verrons ensuite comment choisir les *petites valeurs théoriques vraisemblables* pour construire le test du $\chi^2$ d'indépendance statistique.

## Distribution du $\chi^2$ sous hypothèse d'indépendance.

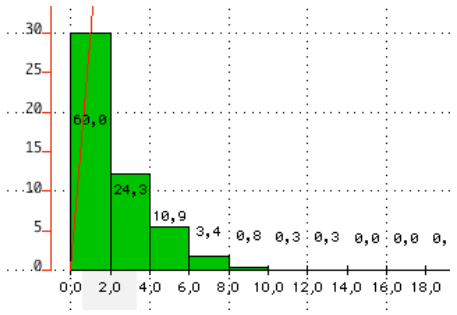
### Distribution du $\chi^2$ par simulation.

**4** Considérons donc une observation conjointe $D$ de deux variables sur un échantillon de taille n, et supposons que les distributions de X et Y dans la population sont les marges de $D$ ; pour obtenir la distribution des valeurs théoriques du $\chi^2$ sous hypothèse d'indépendance, nous simulons un grand nombre de fois l'observation conjointe de X et Y sur des échantillons de taille $n$ en supposant X et Y indépendantes, comme nous l'avons fait dans le cours précédent : en notant à chaque fois la valeur du $\chi^2$ de la distribution conjointe obtenue, on obtient une série de mesures dont on peut construire la distribution.

**5** **Exemple 1 : « niveau scolaire et absentéisme ».** C'est l'exemple développé dans le cours précédent ; comme l'échantillon est de taille 27, on effectue 1000 simulations d'observation conjointe

de X et Y *supposées indépendantes* sur des échantillons de taille 27, en prenant comme distributions de X et Y dans la population les marges de $D$ (§9) ; en notant à chaque fois le $\chi^2$ de la distribution conjointe obtenue, on obtient une série de 1000 valeurs théoriques du $\chi^2$ reflétant les variations dues aux fluctuations d'échantillonnage dans un échantillon de taille 27 (§11) ; sa distribution a l'allure suivante si on prend des intervalles de largeur 2 comme modalités :

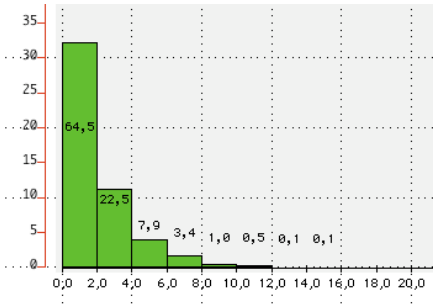
| Z | [0 2[ | [2 4[ | [4 6[ | [6 8[ | [8 10[ | [10 12[ | [12 14[ | [14 16[ | total |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 600 | 243 | 109 | 34 | 8 | 3 | 3 | 0 | 1000 |
| $f_i$ en % | 60,0 | 24,3 | 10,9 | 3,4 | 0,8 | 0,3 | 0,3 | 0,0 | 100 |




En comparant $\chi^2(D) = 1{,}05$ à la distribution, on constate que c'est une petite valeur pour un $\chi^2$ puisque 70% environ des valeurs théoriques lui sont supérieures ; tout se passe donc comme si l'échantillon observé était un des 1000 échantillons simulés, et donc comme si $\chi^2(D)$ s'expliquait entièrement par les fluctuations d'échantillonnage : l'indépendance statistique entre X et Y parait très vraisemblable.

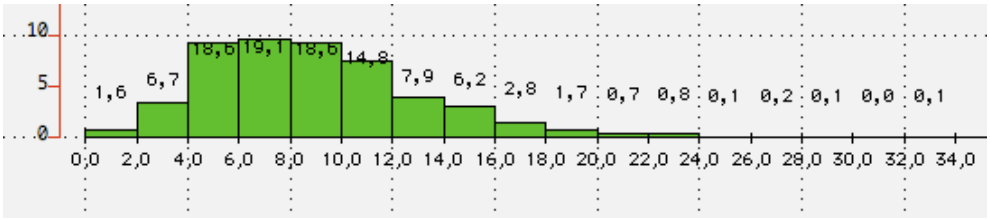
**6 Exemple 2 : « traitements anti-termites ».** On procède de la même manière à partir du tableau de contingence, dont l'effectif total est 600 ; on simule l'observation conjointe de X et Y supposées indépendantes sur des échantillons de taille 600, en prenant comme distributions de X et Y dans la population les marges du tableau ; la distribution des 1000 $\chi^2$ calculés aura l'allure suivante, qui ressemble beaucoup à la précédente :

| Z | [0 2[ | [2 4[ | [4 6[ | [6 8[ | [8 10[ | [10 12[ | [12 14[ | [14 16[ | total |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 645 | 225 | 79 | 34 | 10 | 5 | 1 | 1 | 1000 |
| $f_i$ en % | 64,5 | 22,5 | 7,9 | 3,4 | 1,0 | 0,5 | 0,1 | 0,1 | 100 |




En comparant $\chi^2(D) = 18{,}6$ à la distribution, on voit qu'il n'y a pratiquement aucune chance que l'observation de deux variables indépendantes sur un échantillon aléatoire de taille 600 donne un $\chi^2$ aussi important : les fluctuations d'échantillonnage ne paraissent pas expliquer à elles seules cette forte valeur, et il faut supposer une liaison statistique entre les variables, sans doute expliquée par l'efficacité des traitements.

**7 Exemple 3 : « revenu mari et femme ».** On procède de manière analogue à partir du tableau de l'exemple 6 du premier cours ; en simulant 1000 observations conjointes de X et Y supposées indépendantes, sur des échantillons de taille 100, on obtiendrait pour les $\chi^2$ calculés une distribution de la forme :



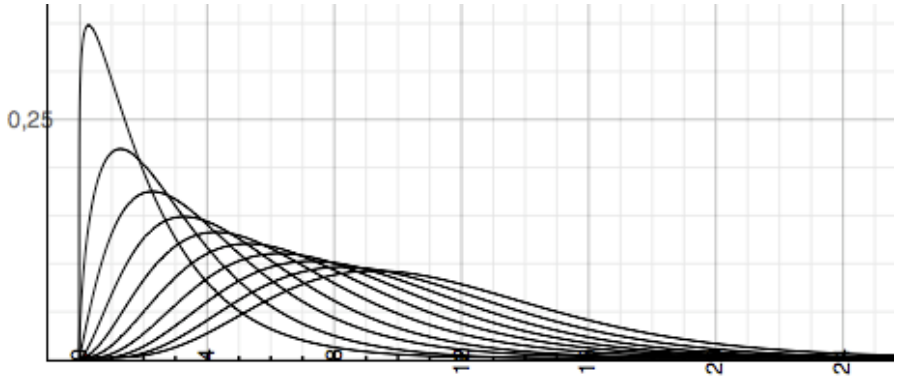

| Z | [0 2[ | [2 4[ | [4 6[ | [6 8[ | [8 10[ | [10 12[ | [12 14[ | [14 16[ | [16 18[ | [18 20[ | [20 1000[ | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 16 | 67 | 186 | 191 | 186 | 148 | 79 | 62 | 28 | 17 | 20 | 1000 |
| $f_i$ en % | 1,6 | 6,7 | 18,6 | 19,1 | 18,6 | 14,8 | 7,9 | 6,2 | 2,8 | 1,7 | 2,0 | 100 |

Le $\chi^2$ de l'échantillon effectivement observé vaut 158,78 ; comparé à la distribution, ce nombre semble impossible à obtenir pour un échantillon sous hypothèse d'indépendance : les variables sont manifestement liées (l'étude des distributions conditionnelles conduit à la même conclusion).

## Distribution théorique du $\chi^2$

**8** La seconde méthode pour déterminer la distribution du $\chi^2$ sous l'hypothèse d'indépendance de X et Y consiste à utiliser le résultat suivant de statistique mathématique : si X et Y ayant k et p modalités sont indépendantes, la distribution [théorique] des $\chi^2$ obtenus à partir d'échantillons aléatoires est une fonction connue, la loi du $\chi^2$, qui ne dépend pas de la taille de l'échantillon, mais seulement du produit $(k-1)*(p-1)$ appelé degré de liberté (ddl) de la distribution conjointe ou de la loi.

Ces distributions théoriques sont représentées ci-dessous pour des ddl compris entre 2 et 11, de gauche à droite : les modalités sont ici des intervalles de largeur infinitésimale, si bien que les histogrammes ont la forme de courbe appelées *courbes de densités* :



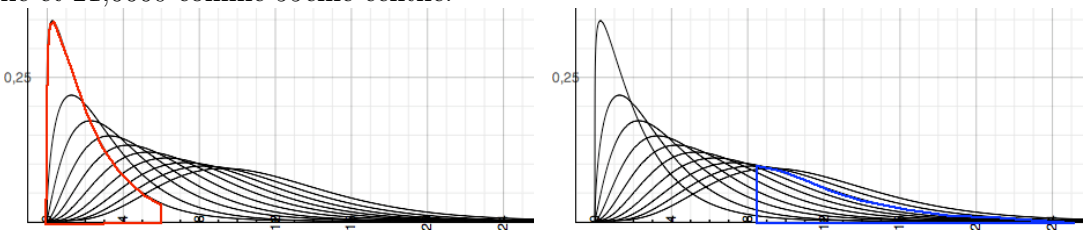

On pourra comparer les distributions des exemples 1 et 2 dont le ddl est (2-1)*(3-1)=2 avec la distribution théorique de ddl 2, et la distribution de l'exemple 3 avec la distribution théorique de ddl (4-1)*(4-1)=9.

**9** Dans la pratique, on utilise les distributions théoriques à partir de tables qui donnent en ligne pour un ddl donné, les quantiles les plus utiles :

| ddl / P | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 | 0,95 | 0,975 | 0,99 | 0,995 | 0,999 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,4549 | 0,7083 | 1,0742 | 1,6424 | 2,7055 | 3,8415 | 5,0239 | 6,6349 | 7,8794 | 10,8276 |
| 2 | 1,3863 | 1,8326 | 2,4079 | 3,2189 | 4,6052 | 5,9915 | 7,3778 | 9,2103 | 10,5966 | 13,8155 |
| 3 | 2,366 | 2,9462 | 3,6649 | 4,6416 | 6,2514 | 7,8147 | 9,3484 | 11,3449 | 12,8382 | 16,2662 |
| 4 | 3,3567 | 4,0446 | 4,8784 | 5,9886 | 7,7794 | 9,4877 | 11,1433 | 13,2767 | 14,8603 | 18,4668 |
| 5 | 4,3515 | 5,1319 | 6,0644 | 7,2893 | 9,2364 | 11,0705 | 12,8325 | 15,0863 | 16,7496 | 20,515 |
| 6 | 5,3481 | 6,2108 | 7,2311 | 8,5581 | 10,6446 | 12,5916 | 14,4494 | 16,8119 | 18,5476 | 22,4577 |
| 7 | 6,3458 | 7,2832 | 8,3834 | 9,8032 | 12,017 | 14,0671 | 16,0128 | 18,4753 | 20,2777 | 24,3219 |
| 8 | 7,3441 | 8,3505 | 9,5245 | 11,0301 | 13,3616 | 15,5073 | 17,5345 | 20,0902 | 21,955 | 26,1245 |
| 9 | 8,3428 | 9,4136 | 10,6564 | 12,2421 | 14,6837 | 16,919 | 19,0228 | 21,666 | 23,5894 | 27,8772 |
| 10 | 9,3418 | 10,4732 | 11,7807 | 13,442 | 15,9872 | 18,307 | 20,4832 | 23,2093 | 25,1882 | 29,5883 |

Par exemple, pour un ddl de 2 (seconde ligne) on trouve comme médiane 1,3863 (la proportion des $\chi^2$ inférieurs à 1,3863 est 0,5), comme 9ème décile 4,6052 (9 dixième des $\chi^2$ sont inférieurs ou égaux à 4,6052), comme 95ème centile 5,9915 (95 % des $\chi^2$ sont inférieurs ou égaux à 5,9915), ou encore comme 99ème centile 9,2103 (1% seulement des $\chi^2$ sont supérieurs à 9,2103).

Pour un ddl de 9 on trouve 8,3428 comme médiane (figure de droite), 14,6837 comme 9ème décile et 21,6660 comme 99ème centile.




## Test du $\chi^2$

**10** La comparaison de la valeur $\chi^2(D)$ avec la distribution des $\chi^2$ théoriques (qui ne dépend que du nombre de lignes et de colonnes de $D$) donne l'idée assez simple d'une règle de décision concernant l'indépendance statistique de X et Y : si la valeur $\chi^2(D)$ est supérieur à la plus grande valeur théorique $s$ on décide de rejeter l'hypothèse d'indépendance statistique puisque les fluctuations d'échantillonnage ne peuvent à elles seules expliquer cette valeur, contrairement au cas où $\chi^2(D) \leq s$, qui nous incite donc à conclure à l'indépendance statistique.

**11** L'inconvénient de cette règle est que $s$ est infini : il n'y a pas de valeur maximale à l'effet des fluctuations d'échantillonnage, du moins en théorie ; on peut alors choisir un seuil $s$ qui ne soit pas la plus grande valeur, mais une valeur suffisamment grande pour que les échantillons sous hypothèse d'indépendance donnant un $\chi^2$ supérieur à $s$ soient rarissimes ; dans le cas d'une distribution à 9 ddl par exemple, on pourrait prendre $s = 28$ puisque la table et la simulation donnent une proportion de 2 pour 1000 pour de tels échantillons ; il faut remarquer que ce choix conduit inévitablement à une décision avec erreur : dans l'exemple précédent où l'on rejette l'indépendance quand le $\chi^2$ est supérieur à 28, nous avons 2 chances sur 1000 de faire une erreur puisque 2 millièmes des échantillons sous hypothèse d'indépendance donnent un $\chi^2$ supérieur à 28 et conduiraient donc à une décision erronée.

**12** Si on prend $s$ trop grand, on risque de ne pas identifier les faibles liaisons, qui ont une valeur du $\chi^2$ certes éloignée de 0, mais inférieure à $s$ : la règle nous conduit en effet à considérer X et Y indépendantes dans ces conditions ; or ces cas méritent une attention particulière : les distributions conditionnelles sont suffisamment différentes pour qu'on puisse suspecter une liaison intéressante à interpréter, mais pas assez pour que ce soit manifeste. On peut alors améliorer la règle ou prenant un nombre à partir duquel la densité (la hauteur dans la représentation graphique) devient très petite : par exemple 24 (seulement 5 échantillons sous hypothèse d'indépendance sur 1000 ont un $\chi^2$ supérieur) ou même 18 (environ 3,7% des échantillons). D'un autre coté, diminuer $s$ augmente les chances de faire une erreur quand on rejette l'indépendance, et il faudrait connaître le prix de cette décision erronée pour savoir où s'arrêter ; mais cette question dépasse l'objet de ce cours, et dans la suite nous prendrons pour les exemples et exercices, le 99ème centile de la distribution des valeurs théoriques (21,7 pour 9 ddl) avec 1 chance sur 100 de faire une erreur, ou même le 95ème (16,9 pour 9 ddl) avec 5 chances sur 100 de faire une erreur.

**13** **Test du $\chi^2$.** Les réflexions précédentes conduisent naturellement au test du $\chi^2$ : le test du $\chi^2$ d'indépendance statistique est une procédure pour rejeter ou non l'hypothèse d'indépendance statistique de deux variables X et Y dans une population, à partir de leur mesure conjointe $D$ sur un échantillon ; il fournit une règle pour apprécier si $\chi^2(D)$ peut ou non être considéré comme le seul effet des fluctuations d'échantillonnage.

## 14 Pratique du test.

1. Avant l'observation :
   a) on choisit le risque d'erreur $\alpha$ que l'on consent à prendre en rejetant l'hypothèse d'indépendance ; dans les exercices on prendra $\alpha = 0{,}05$ (5%) ou $\alpha = 0{,}01$ (1%) quand il ne sera pas indiqué ;
   b) on calcule le degré de liberté $ddl = (k-1) * (p-1)$ correspondant à la modélisation de la situation statistique ;
   c) à l'aide d'une table des valeurs théorique $\chi^2$ et du $ddl$, on détermine le seuil de décision $s = q_{1-\alpha}$, valeur à partir de laquelle il y a $\alpha$% des valeurs théoriques ;
2. Après l'observation conjointe $D$ :
   a) On calcule $\chi^2(D)$ : calcul des effectifs théoriques $\widetilde{n_{ij}}$, puis des contributions $\chi^2_{ij}$ ;
   b) si $\chi^2(D) \geq s$ on rejette l'hypothèse d'indépendance statistique des variables X et Y avec $\alpha$% d'erreur de se tromper, en considérant que les fluctuations d'échantillonnage ne peuvent pas expliquer à elles seules cette valeur ;
   c) si $\chi^2(D) < S$ on ne rejette pas l'hypothèse d'indépendance, en la considérant comme plausible ;

## 15 Exemples.

1 – Niveau scolaire et absentéisme. $ddl = (2-1) * (3-1) = 2$ ; en prenant $\alpha = 0{,}01$ (1%) on trouve $s = 9{,}21$ sur la 2ème ligne de la colonne 0,99 (1-0,01) ; $\chi^2(D) = 1{,}05 < s$ : on ne rejette pas l'hypothèse d'indépendance, en considérant plausible que X et Y le soient.

2 – Traitements anti-termites. $ddl = 2$ ; en prenant le même risque d'erreur on trouve le même seuil $s = 9{,}21$ ; comme $\chi^2(D) = 18{,}6 > s$ on rejette l'hypothèse d'indépendance, avec un risque d'erreur de 1%.

3 – Revenu mari et femme. $ddl = (4-1) * (4-1) = 9$ ; un risque d'erreur de 5% ($\alpha = 0{,}05$) donne un seuil $s$ égal à 16,9 (colonne 0,95) ; comme $\chi^2(D) = 158$, on rejette l'hypothèse d'indépendance, avec un risque d'erreur inférieur de 5%.

## 16 Remarques.

1 – Un test du $\chi^2$ ne prouve ni que les variables sont indépendantes statistiquement, ni qu'elles ne le sont pas ; il permet seulement de choisir l'option la plus « raisonnable » au regard de l'observation d'un échantillon, en l'absence d'autres procédures plus précises.

2 – Il n'est pas toujours nécessaire de calculer $\chi^2(D)$ pour rejeter l'hypothèse d'indépendance : si la somme des contributions déjà calculées dépasse $s$, il est inutile d'en calculer d'autres puisque $\chi^2(D)$ qui sera au moins égal à cette somme partielle dépassera également $s$.

3 – Il y a deux façons de faire une erreur dans un test : rejeter l'hypothèse d'indépendance quand X et Y sont effectivement indépendantes, et ne pas rejeter cette hypothèse en admettant qu'elles sont indépendantes quand elles ne le sont pas. Quand on fait la première erreur, la distribution du $\chi^2$ est la distribution théorique puisque X et Y sont dans ce cas indépendantes ; le risque d'erreur est la proportion d'échantillons qui donnent un $\chi^2$ supérieur à $s$, c'est à dire $\alpha$. Quand on fait la seconde erreur, la distribution du $\chi^2$ est inconnue puisque on sait seulement que X et Y ne sont pas indépendantes : on ne peut donc pas quantifier cette sorte d'erreur.

## Questions de cours

1. $D$ est la distribution conjointe de X et Y : à quelle condition sur $\chi^2(D)$ peut-on affirmer que X et Y sont indépendantes ?
2. Pourquoi l'indépendance entre X et Y n'est pas une définition opératoire ?
3. À quelle conditions peut-on dire que X et Y sont statistiquement indépendantes ?
4. Différence entre indépendance et indépendance statistique.
5. $D$ est la distribution conjointe de X et Y : les distributions conditionnelles ne sont pas très différentes, mais pas identiques : peut-on estimer que X et Y sont indépendantes ? statistiquement indépendantes ?
6. Que signifie *ddl* ?
7. Quelle est la valeur du *ddl* ?
8. X a trois modalités, Y quatre : que vaut le *ddl* ?
9. Qu'est-ce que $\alpha$ ?
10. Comment s'appelle $s$ ?
11. Quels sont les deux décisions possibles dans un test du $\chi^2$ ?
12. Quelle est la proportion des observations sous hypothèse d'indépendance qui donne un $\chi^2$ supérieur à $s$ ?

## Questions sur le cours

Quel est le seuil de décision dans les conditions suivantes :

1. $k = 2, p = 2, \alpha = 0{,}05$
2. $k = 2, p = 4, \alpha = 5\%$
3. $k = 3, p = 4, \alpha = 0{,}01$
4. $k = 5, p = 7, \alpha = 1\%$