Analyse de données L3 2024-2025

Mise en relation de deux caractères qualitatifs

Florian Bayer

Rappels et objectifs du cours

En partant du **tableau de contingence**, nous allons voir comment analyser les relations entre deux ou plusieurs séries de données **qualitatives.**

- Existe-t-il un lien entre le versant d'un massif (N,S,E,O) et le type de culture ?
- Les accidents aériens (oui/non) sont ils dépendant du type de piste des aéroports (tarmac/herbe/terre) ?
- La réussite à un concours (oui/non) est elle dépendante de l'enseignant ?

Mettre en évidence une relation n'est cependant pas suffisant. Il convient de savoir si le résultat peut être **validé avec certitude** ou non. Nous aborderons donc dans ce cours :

1. La démarche scientifique
2. Description d'une relation entre 2 caractères qualitatifs discrets
3. Test du Chi^2

1- La démarche scientifique

Les critères de scientificité

La théorie scientifique comporte **3 composantes** :

- Objectivité
- Reproductibilité
- Falsifiabilité

K.R. Popper pose la question : *Existe-t-il un critère permettant d'établir la nature ou le statut scientifique d'une théorie ?*

Il introduit alors le principe de **falsifiabilité** : on ne peut accepter une théorie scientifique que s'il existe un moyen de **prouver qu'elle est erronée**. L'objectif est donc de pouvoir soumettre cette théorie à des expériences (des tests) afin de vérifier si la théorie et l'observé concorde.

«Une théorie qui n'est réfutable par aucun événement qui se puisse concevoir est dépourvue de caractère scientifique. Pour les théories, l'irréfutabilité n'est pas vertu mais défaut ». **Une théorie irréfutable est alors jugée comme méta-physique.**

«Le critère de la scientificité d'une théorie réside dans la possibilité de l'invalider, de la réfuter ou encore de la tester » K.R. Popper, Conjecture et réfutation, Payot, Paris, pp. 58-65.

Tests de significativité : exemple 1

Obtenir un résultat (un indicateur d'intensité de relation, une comparaison de deux moyennes) n'est pas suffisant. Il est obligatoire de déterminer si ce résultat est lié au hasard (fluctuations d'échantillonage, pas de relation, biais de sélection, effectifs insuffisants etc.) ou s'il peut être **validé** avec un certaine **marge d'erreur**. On parle de **test de significativité**.

Exemple : un article démontre que la part estimée de diabétique dans la population d'un pays est de 20%. Un scientifique décide de vérifier si ce taux est égal à 20% dans sa région. Comme il ne peut pas faire un test sur toute la population régionale, il réalise une étude sur un groupe représentatif (**échantillon**) semblable à celle de l'étude du pays.

- Cas 1 : il obtient un taux de 5%. Dans ce cas il **rejete** son hypothèse du fait de l'écart important entre les deux.
- Cas 2 : il obtient un taux de 18 %. L'hypothèse peut être jugée raisonnable et **il n'y a pas de raison évidente de la rejeter**.

L'intuition qui conduit à cette conclusion doit être **formalisée** en une **règle précise** s'appliquant à chaque cas et en fonction de la taille de la population étudiée : **un test de significativité**.

Tests de significativité : exemple 2

Deux équipes de recherche décident de mesurer la relation entre la température et l'altitude.

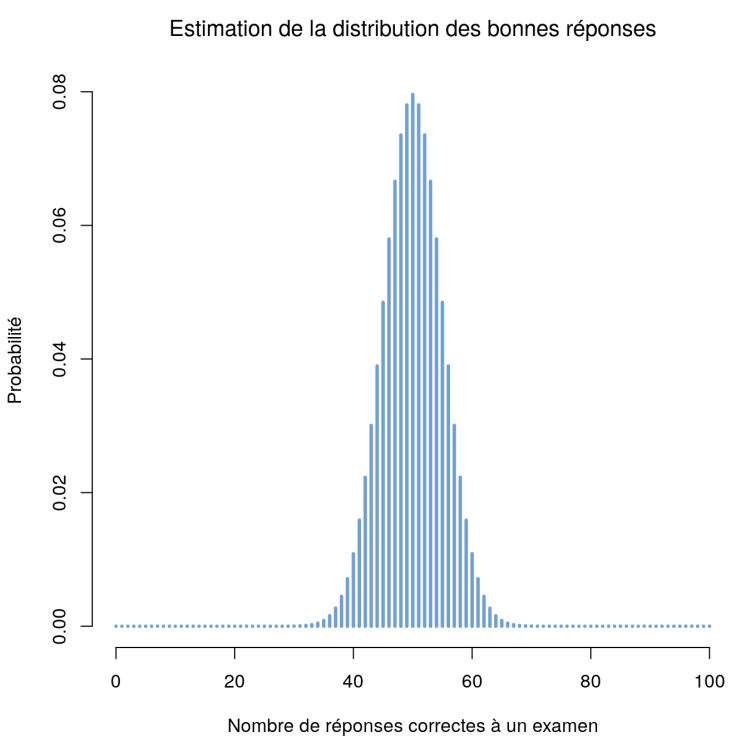
- Par manque de moyen, l'équipe A n'a pu faire **que 5 mesures**, l'équipe B **100 mesures**.
- Les deux équipes obtiennent le même résultat : tous les 100 mètres, la température baisse environ de 0,6°C (environ car il y a des fluctuation de mesures : parfois de 0,5°, 0,65° etc.)

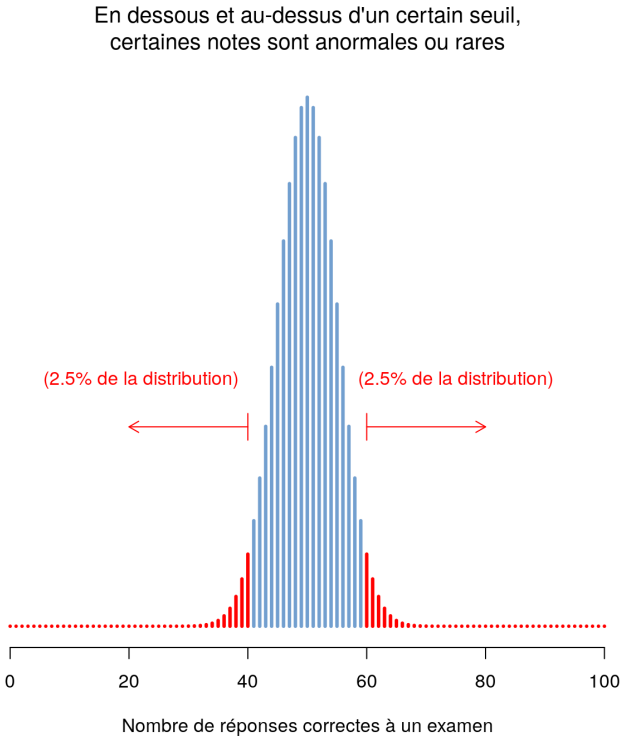
Pourtant, seul le résultat de l'équipe B peut être validé car les mesures de l'équipe A sont insuffisantes et **peuvent être liées au hasard**. Pour compenser ce manque, il faudrait que les mesures réelles de la température baissent exactement (ou quasi) de 0,6° tous les 100 mètres pour l'équipe A. A l'inverse, l'équipe B peut valider son modèle en dépit des éventuelles fluctuations des mesures de la température.

Dans cet exemple, c'est l'effectif qui est insuffisant (**manque de puissance**), mais l'intensité de la relation ou la rareté d'un phénomène peuvent également jouer.

Les lois en statistiques

Pour faire un test de significativité, on se base sur des lois statistiques, comme la loi normale (cf. cours sur l'univariée). Le chi² possède sa propre loi. Ces lois permettent de prédire la probabilité qu'un événement théorique aléatoire ou rare survienne.





Principes d'un test statistique

Pour valider **objectivement** le fait que des différences ou des relations ne soient pas liées au hasard, on formalise un test statistique qui suit **toujours les mêmes règles**.

1. Formuler une hypothèse nulle et l'hypothèse alternative
   - On parle généralement **d'hypothèse d'indépendance**, notée **H0** : il n'y a pas de relation entre un caractère X et un caractère Y.
   - Son alternative, **H1**, qui décrit la situation si H0 est fausse : il y a une relation entre X et Y.
2. Calculer une statistique qui **résume la situation observée**
   - On utilise des outils statistiques comme le coefficient de corrélation ou le calcul du Chi^2.
   - Mais on peut aussi comparer deux moyennes.
   - C'est le **résultat observé**.
3. Comparer le calcul de l'étape 2 et ce que prévoit **les lois de la statistique** à l'aide d'une règle de décision.
   - Si le résultat observé se trouve être un élément commun de la loi statistique théorique, on ne pourra pas rejeter H0.
   - Si le résulta observé est un évènement rare de la loi (par exemple dans les 2.5% de l'exemple précédent), on aura tendance à rejeter H0 et accepter H1.

La règle de décision qui nous permet de rejeter ou non H0 est appelée le risque d'erreur α (à partir de quelle écart entre le calculé et l'observé peut-on rejeter H0 ?)

Le risque d'erreur α

La règle de décision étant basée sur des observations, elle est sujette à des erreurs. La marge d'erreur est définie par un paramètre α et un paramètre β qui ne sera pas abordé.

α est aussi appelé erreur de première espèce : rejeter l'hypothèse d'indépendance H0 alors qu'elle est vraie (conclure que le résultat n'est pas lié au hasard alors qu'il l'est). En science sociale, il est commun d'utiliser un α égal à 5%. En rejettant H0 avec un risque $\alpha = 0.05$, il existe un risque de 5% de rejeter H0 à tort (le résultat observé est bien du hasard).

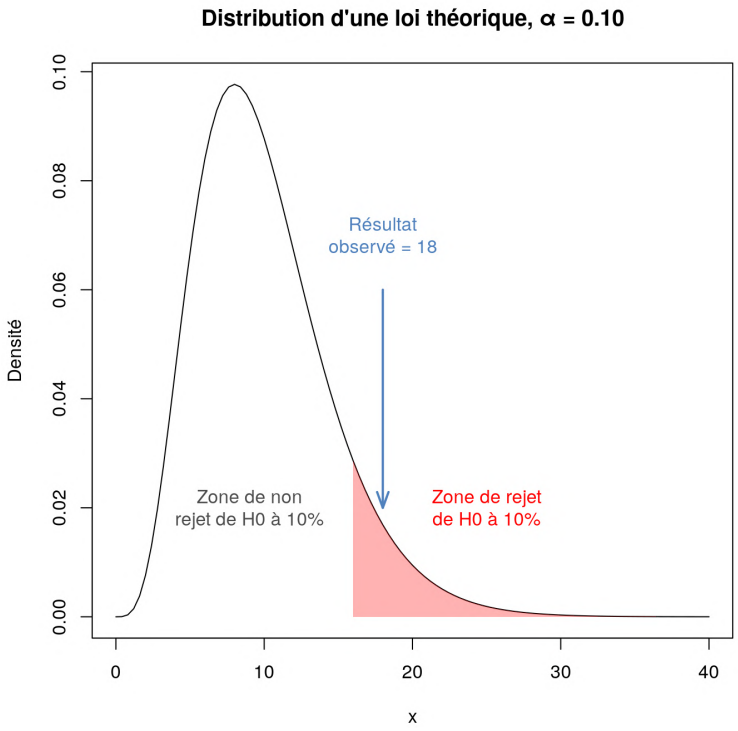
- Il s'agit d'une valeur que **vous** devrez toujours déterminer. Il est nécessaire de s'y tenir et de ne pas la modifier au cours du test.

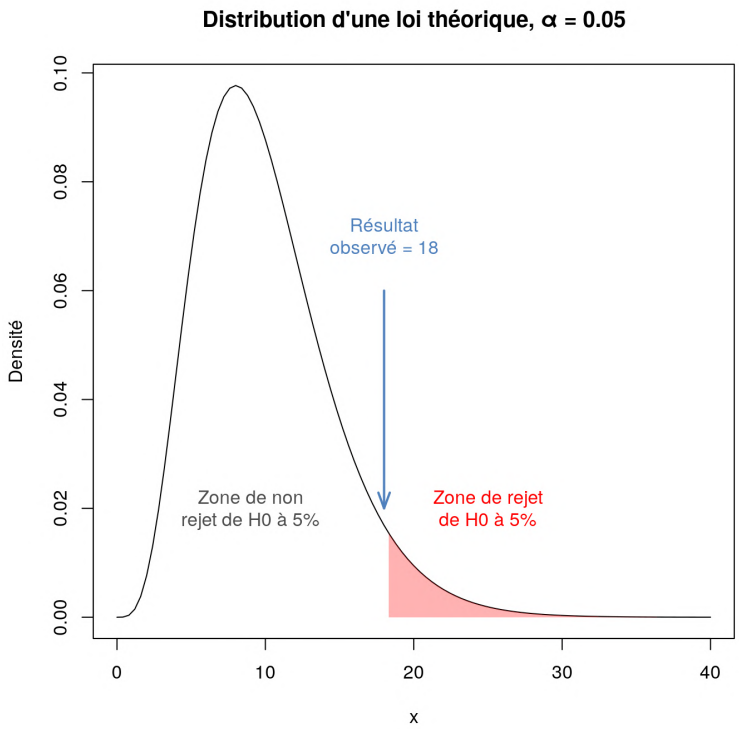
Dans l'exemple des diabétiques, on souhaite vérifier s'il y avait une différence entre entre le % de diabétiques au niveau national (20%) et dans la région d'étude.

- Si H0 est vraie (20% de diabétiques dans la population de la région) avec un risque $\alpha = 0.05$, on admet de se tromper en moyenne 5 fois sur 100 en rejetant H0 alors que H0 est vraie.

«Il faut retenir que la seule chose qu'on puisse démontrer par des observations, est qu'une hypothèse est fausse. On ne peut pas démontrer qu'elle est vraie. Tout au plus, peut-on dire qu'elle n'est pas contredite par l'expérience» (Popper 1973).

Illustration graphique



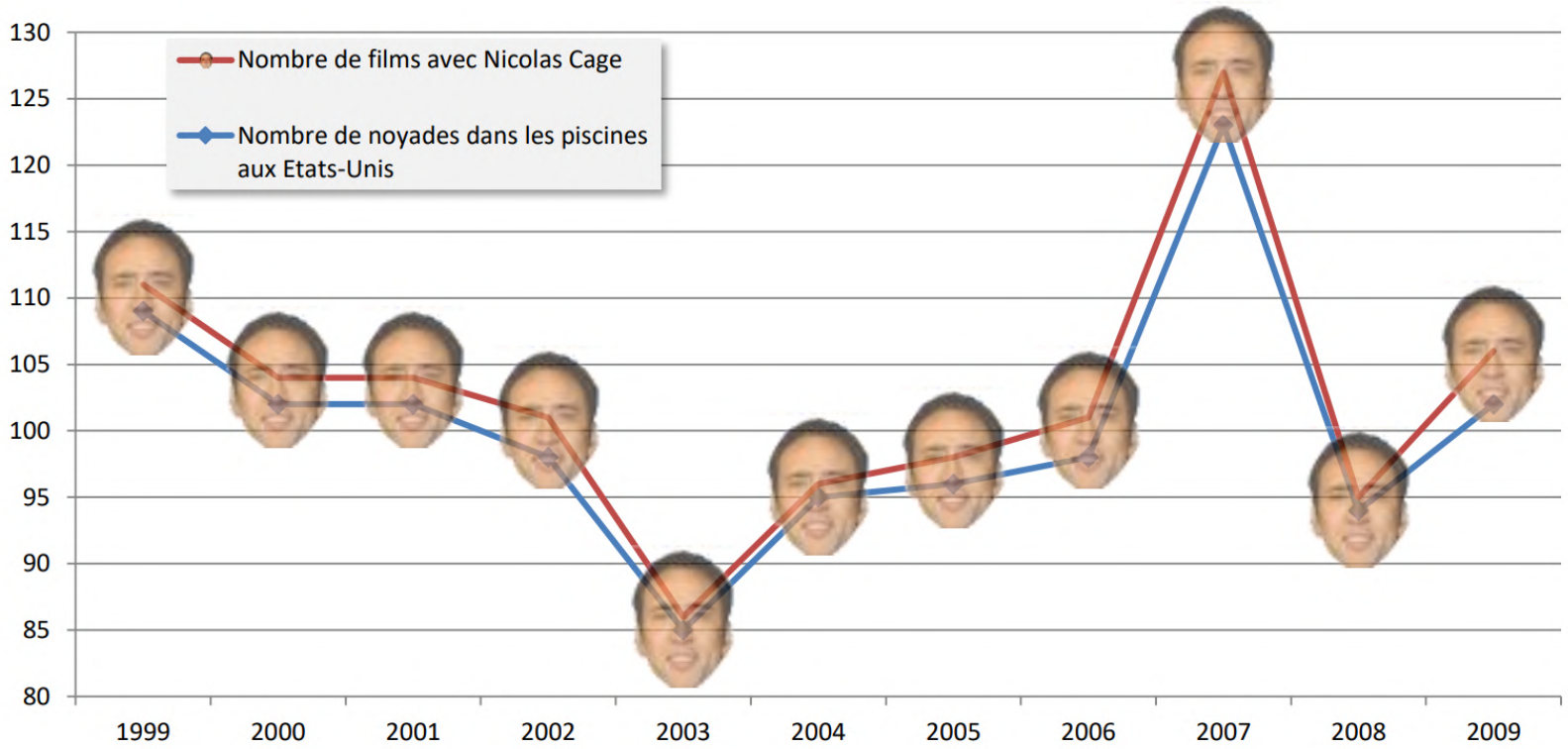


On calcule une statistique à partir d'observations, qui nous donne un résulat égal à 18 (une moyenne, un chi² ou tout autre calcul). La table théorique de la loi associée à cette statistique, indique qu'avec α = 0.1, le résultat observé est compris dans la zone de rejet de l'hypothèse H0. On accepte l'hypothèse alternative H1 : la statistique calculée est un évènement rare (en prenant le risque de rejeter à tord H0 1 fois sur 10)

Avec α = 0.05, cette statistique n'est plus comprise dans la zone de rejet de l'hypothèse H0. On ne peut donc pas rejeter H0. L'observation faite n'est pas "rare" avec ce seuil α

Causalité et corrélation

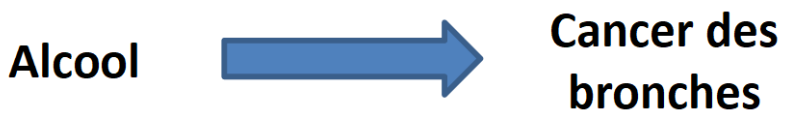
Attention, un lien statistique n'est pas forcément le signe d'une causalité (une interprétation explicative de la corrélation). Elle peut-être liée au hasard :



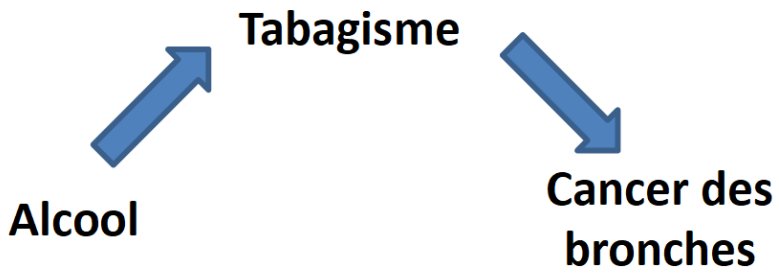
Facteur de confusion

Exemple consommation d'alcool et cancer des bronches.

- Une étude analyse le rôle de la consommation d'alcool dans le risque d'avoir un cancer des bronches. Elle trouve un lien :



En réalité, le risque de développer un cancer des bronches n'est pas directement lié à la consommation d'alcool, mais au fait que les consommateurs d'alcool ont une probabilité plus forte d'être fumeur. Le tabagisme est ici un **facteur de confusion**.



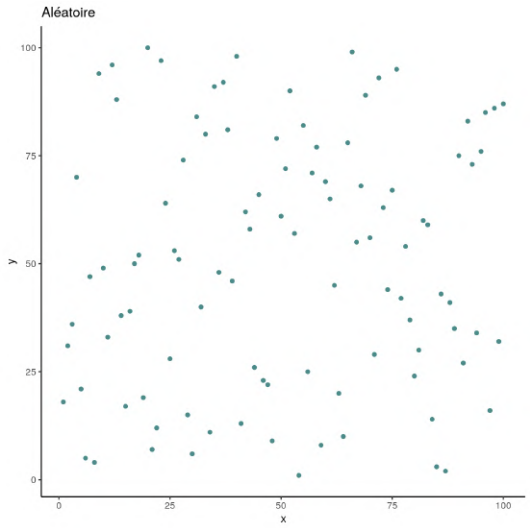
2- Description d’une relation entre des caractères qualitatifs

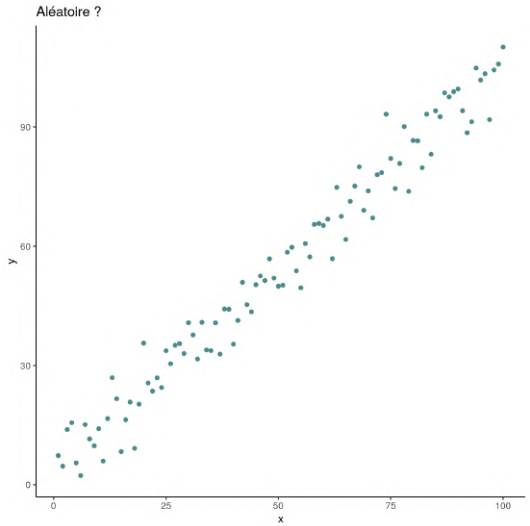
Concepts

Parler de relation en statistique revient à établir un rapport logique entre au moins deux caractères X et Y

L'intensité de la relation peut souvent être mesurée et elle doit toujours être testée statistiquement (ce résultat est-il lié au hasard ?).

L'idée est de **comparer** une situation **théorique aléatoire et l'observé**. Si l'observé est très proche de la situation aléatoire, on ne peut pas conclure qu'il existe une relation. A l'inverse, si observé et aléatoire sont très différents, on peut se poser la question d'une relation entre X et Y





Relation entre caractères qualitatifs discrets

L'un des outils les plus courant pour mesurer la relation entre des caractères qualitatifs discrets est le test du **Chi²**. Il s'applique sur les **tableaux de contingence**.

Le tableau de contingence permet :

- d'étudier la répartition des valeurs de X par rapport à Y et des valeurs de Y par rapport à X (profils lignes et profils colonnes).
- de comparer une **distribution observée** à une **distribution théorique** (indépendante), qui correspond à ce que serait la réalité s'il n'y avait aucune relation entre X et Y (distribution aléatoire entre X et Y).

Si l'écart entre l'observé et l'indépendance est faible, il est fort probable qu'il n'y ait pas de relation. En effet, cela revient à considérer que les deux caractères X et Y se comportent indépendamment. A l'inverse, **un fort écart** à l'indépendance peut laisser supposer une relation entre les individus mesurés.

Exemple

36 étudiants d'une promotion sont décrits par un ensemble de variables relatives au sexe, à l'âge et au groupe (il existe deux groupes). On souhaiterait :

- connaître la répartition des groupes dans l'ensemble de la promotion et par sexe
- connaître la répartition des sexes dans l'ensemble de la promotion et par groupe
- savoir si les hommes et les femmes sont répartis de façon aléatoire entre les deux groupes

Etudiant	Groupe	Sexe
A	1	H
B	1	H
C	2	F
D	2	H
E	1	F
...	...	...

La variable Groupe (X) possède deux modalités et la variable Sexe (Y) possède deux modalités . Il existe donc 4 possibilités de croisement :

- femme du groupe 1
- femme du groupe 2
- homme du groupe 1
- homme du groupe 2

Du tableau élémentaire au tableau de contingence

Pour répondre à ces questions, le **tableau de contingence** (un tableau dénombrant les modalités croisées des deux caractères X et Y) est un outil adapté. Ce tableau aura :

- k lignes (nombre de modalités de X)
- p colonnes (nombres de modalités de Y)

Des marges seront ajoutées avec :

- les totaux en lignes (effectif de chaque modalité de X)
- les totaux en colonnes (effectif de chaque modalité de Y)
- le total général (nombre n d'individus étudiés)

Construction du tableau de contingence

Nij	Femme	Homme	Total
Groupe 1	3	14	**17**
Groupe 2	9	10	**19**
Total	**12**	**24**	**36**

Ce tableau de contingence permet de dénombrer tous les cas possibles de modalités.

On peut ainsi dire :

- qu'il y a 14 hommes dans le groupe 1
- qu'il y a 19 étudiants dans le groupe 2
- qu'il y a 12 étudiantes
- ou encore qu'il y a en tout 36 étudiants

En revanche ce tableau n'indique que des effectifs bruts.

- Il ne permet pas de comparer les proportions des modalités
- Il ne permet pas de répondre directement à des questions du type **la proportion d'hommes est-elle plus élevée dans le groupe 1 que dans le groupe 2 ?**

On construit donc généralement deux tableaux de profils indiquant les pourcentages en lignes ou les pourcentages en colonnes.

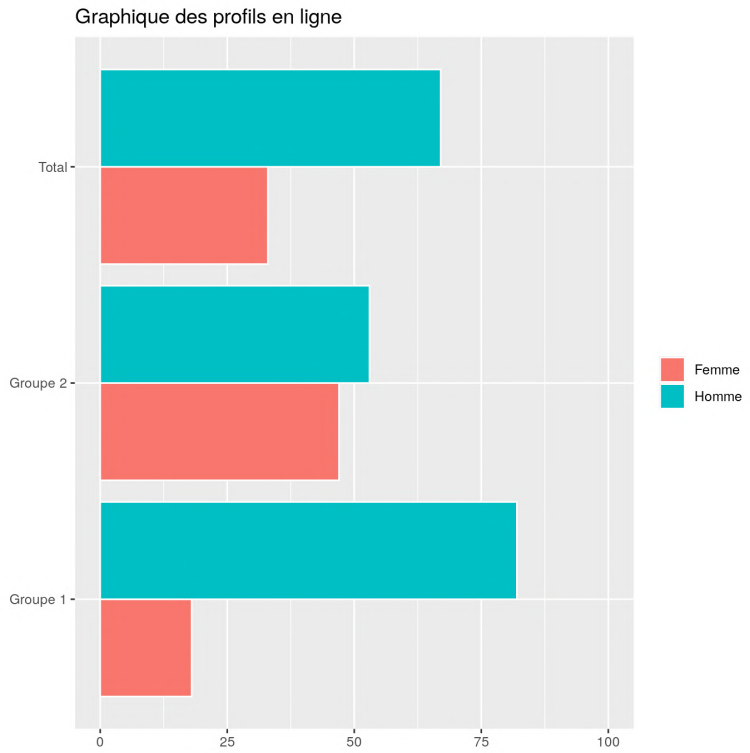
Analyse des profils en ligne

Le tableau des **profils en ligne** est construit en divisant l'effectif de chaque case par le total de la ligne correspondante :

Nij/Ni	Femme	Homme	Total
Groupe 1	18% (3/17x100)	82% (14/17x100)	**100% (17/17x100)**
Groupe 2	47%	53%	**100%**
Total	**33%**	**67%**	**100%**

Nij	Femme	Homme	Total
Groupe 1	3	14	**17**
Groupe 2	9	10	**19**
Total	**12**	**24**	**36**

La part de femmes dans l'ensemble de la promotion est de 33% mais elle est sensiblement plus élevée dans le groupe 2 (47%) que dans le groupe 1 (18%)

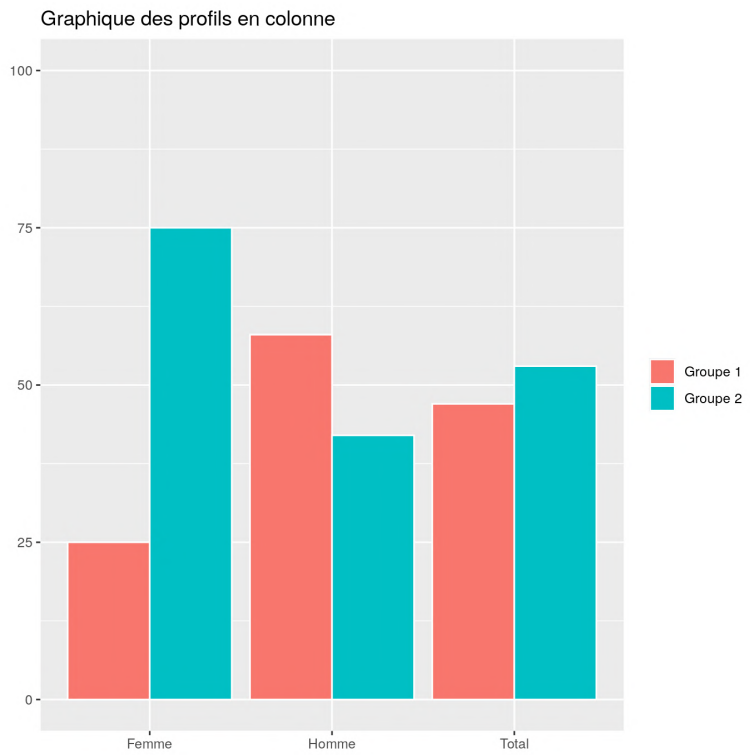


Analyse des profils en colonne

Le tableau des **profils en colonnes** est construit en divisant l'effectif de chaque case par le total de la colonne correspondante :

Nij/Ni	Femme	Homme	Total
Groupe 1	25% (3/12x100)	58%	**47%**
Groupe 2	75% (9/12x100)	42%	**53%**
Total	**100% (12/12x100)**	**100%**	**100%**

Nij	Femme	Homme	Total
Groupe 1	3	14	**17**
Groupe 2	9	10	**19**
Total	**12**	**24**	**36**



Le groupe 1 ne totalise que 47% des étudiants de la promotion. Mais on y trouve 58 % de l'ensemble des hommes et seulement 25 % des femmes de l'ensemble de la promotion.

Les effectifs théoriques

Le tableau de contingence nous permet de **décrire objectivement la répartition** par sexe et par groupe. En revanche, il ne permet pas de déterminer s'il y a une quelconque **relation** entre groupe et sexe (la formation des groupes s'est elle faite indépendamment du sexe ?)

L'étude du tableau de contingence peut aussi se faire en comparant les effectifs observés de chacune des cases N_{ij} aux effectifs théoriques des cases N^*_{ij} :

- Les effectifs qui seraient obtenus s'il n'y avait aucun lien entre les deux modalités X et Y. Autrement dit, si l'attribution de chaque modalité se faisait de façon aléatoire entre X et Y.

Pour reconstituer la distribution théorique des k*p cases du tableau de contingence, on va se servir des **marges** du tableau qui définissent les probabilités conditionnelles qu'un individu reçoive telle modalité de X ou de Y.

Cela signifie que les effectifs théoriques ne correspondent pas à une répartition aléatoire des groupes et des sexes, mais que l'on conserve **l'état** des groupes (17 étudiants dans le groupe 1, 19 dans le 2) et des sexes (12 femmes, 24 hommes). On se base donc sur la réalité (on ne modifie pas aléatoirement le sexe des étudiants) afin de calculer les 4 états possibles s'ils étaient aléatoires (g1♂, g1♀, g2♂, g2♀)

Calcul des effectifs théoriques

On note :

- $Ni.$: somme de la $i^{ème}$ ligne, c'est-à-dire nombre d'individus ayant comme attribut la $i^{ème}$ modalité de X
- $Nj.$: somme de la $j^{ème}$ colonne, c'est-à-dire nombre d'individus ayant comme attribut la $j^{ème}$ modalité de Y
- $N..$: somme générale du tableau, c'est-à-dire nombre total d'individus étudiés

1. La probabilité qu'un individu reçoive la modalité i de X est égale à $\frac{Ni.}{N..}$
2. La probabilité qu'un individu reçoive la modalité j de Y est égale à $\frac{Nj.}{N..}$
3. L'effectif théorique de la case Nij (noté $N * ij$) est obtenu en multipliant la probabilité qu'un individu reçoive cette modalité par le nombre d'individu ($N..$).

On aboutit donc à la formule générale du valcul des effectifs théoriques : $N^*_{ij} = \frac{Ni.*Nj.}{N..}$

Interprétation des effectifs théoriques

Ces effectifs théoriques sont ceux qui seraient obtenus s'il existait une **indépendance parfaite** entre l'attribution des modalités de X et de Y. Autrement dit, si en **gardant la même situation** (effectif des groupes et mêmes étudiants), la formation des groupes avait été faite **au hasard**.

Il peut évidemment exister des **écarts** entre la distribution théorique et la distribution observée. Soit en raison :

- de fluctuations aléatoires
- soit en raison de l'existence d'une dépendance entre les deux caractères X et Y (les groupes n'ont pas été formé au hasard, mais selon une logique qui pour le moment nous est inconnue : on peut montrer la corrélation avec les statistiques, mais la causalité devra être recherchée auprès de la personne qui a fait les groupes)

Effectifs théoriques : exemple

N*ij	Femme	Homme	Total
Groupe 1	5.7 (17*12/36)	11.3 (17*24/36)	**17**
Groupe 2	6.3	12.7	**19**
Total	**12**	**24**	**36**

Nij	Femme	Homme	Total
Groupe 1	3	14	**17**
Groupe 2	9	10	**19**
Total	**12**	**24**	**36**

Ce tableau nous indique, que si l'affectation des étudiants à un groupe s'était effectuée indépendamment de leur sexe, il aurait du y avoir 5 ou 6 filles dans le groupe 1 (valeur théorique = 5.7) et non pas 3 comme on l'observe dans la distribution réelle.

On observe bien que pour calculer les effectifs théoriques, on garde **l'état** des groupes et des sexes (les marges) : on reste dans la même situation marginale (mêmes effectifs totaux pour chaque situation possible entre Nij et N^*ij), mais on redistribue aléatoirement à l'intérieur des cases de N^*ij.

Ecarts à l'indépendance

L'intérêt de cette approche et de comparer la répartition des groupes **observée** à la répartition **théorique ou indépendante**. On parle d'**écart à l'indépendance**.

Pour cela, on soustrait pour chaque case l'effectif observé - la valeur théorique (le réel par rapport à l'aléatoire).

Nij - N*ij	Femme	Homme	Total
Groupe 1	-2.7 (3-5.7)	+2.7 (11.3-14)	**0**
Groupe 2	+2.7	-2.7	**0**
Total	**0**	**0**	**0**

Ce tableau nous indique qu'il existe une **surreprésentation** des femmes dans le groupe 2 (+2.7) et donc des hommes dans le groupe 1 (+2.7). Inversement, les femmes sont **sous-représentées** dans le groupe 1 et les hommes sous-représentés dans le groupe 2 (-2.7).

Tests des écarts à l'indépendance

Sachant qu'une distribution empirique ne peut jamais coïncider exactement avec une distribution théorique aléatoire, la question qui se pose est de savoir si les écarts observés sont :

- l'effet du hasard
- s'ils sont les révélateurs d'une relation significative entre les deux variables X et Y

Le test du Chi² est l'outil idéal pour répondre à cette question.

3- Test du Chi²

Concepts

Les calculs précédents permettent de mettre en **évidence des différences**, mais ils ne permettent pas de déterminer si ces différences sont liées au hasard ou si elles sont le signe d'une relation entre X et Y. Dans notre exemple, les groupes ont-ils été formés aux hasard ou suivent-ils une certaine logique ?

Pour cela, on utilise le test du Chi-2 (Khi², Chi², KHI-2, χ^2), dont les objectifs sont :

- de quantifier la somme des déviations (écarts) entre effectifs observées et effectifs théoriques qui sont présentes à l'intérieur d'un tableau de contingence à l'aide **d'une quantité unique** (et non pas case par case).
- De comparer la valeur de cette statistique à sa probabilité d'apparition dans le cas d'une série de tirages aux sorts effectués de façon aléatoire.
- De tenir compte de la taille du tableau (nombre de degrés de liberté).

Détermination du χ^2 observé

Pour éliminer le signe des écarts à l'indépendance et prendre en compte la taille des effectifs des lignes et colonnes, on calcule pour chaque cellule une mesure d'écart à l'indépendance qui est une quantité toujours positive.

Cette quantité appelée Chi-2 local est égale, pour chaque case, au carré de l'écart entre valeur observée et valeur théorique, divisé par la valeur théorique. $\chi^2_{ij} = \frac{(N_{ij} - N^*_{ij})^2}{N^*_{ij}}$

Il s'agit donc d'un **écart relatif** qui prend en compte le fait qu'un écart de +3 n'a pas le même sens entre 2 et 5 qu'entre 102 et 105 : le χ^2 donne plus de poids aux écarts entre les petites valeurs qu'entre les plus grandes valeurs.

Plus le Chi-2 local d'une case est **élevé**, plus la **déviation** entre valeurs observées et valeurs estimées est **significative** sur le plan statistique (c'est-à-dire plus elle correspond à un événement rare ayant peu de chance de se produire si les variables X et Y étaient indépendantes).

Calcul du χ^2 observé

On résume ensuite la quantité globale de déviation présente à l'intérieur du tableau en calculant la valeur χ^2 observé qui est la somme de tous les χ^2 locaux des cases du tableau (1.255+0.628+1.129+0.561 = 3.567).

Chi² ij	Femme	Homme	Total
Groupe 1	1.255 (-2.7²/5.7)	0.628 (2.7²/11.3)	.
Groupe 2	1.129	0.561	.
Total	.	.	**3.567**

Nij - N*ij	Femme	Homme	Total
Groupe 1	-2.7 (3-5.7)	+2.7 (11.3-14)	**0**
Groupe 2	+2.7	-2.7	**0**
Total	**0**	**0**	**0**

Interprétations

Le χ^2 local le plus important concerne la sous-représentation des femmes dans le groupe 1 (1.255). Cela signifie que **l'écart entre l'observé et le théorique aléatoire**, en **prenant en compte les effectifs**, est plus important pour les femmes du groupe 1. En regardant cet écart non pondéré (observé-théorique), on constate qu'on attendait 5.7 femmes dans ce groupe, alors que "seules" 3 sont présentes.

Un autre χ^2 local important est celui des femmes du groupes 2, qui sont plus nombreuses que ce que prévoit la distribution théorique (6.3 attendues contre 9 observées).

Cela illustre bien l'un des intérêts du χ^2 : même si les écarts entre l'observé et le théorique sont symétriques (+/- 2.7), le fait d'avoir des ordres de grandeurs différents (groupes et sexes) est pris en compte dans le calcul.

Interprétations

Les χ^2 locaux sont donc très pratiques pour déterminer où se trouvent les écarts à l'indépendance les plus "anormaux". Pour vous aider à les identifier, vous pouvez calculer la part qu'ils représentent dans le χ^2 total : les part les plus importantes sont les plus intéressantes à observer (g1♀ = 1.255/3.567 *100 = 35%).

Il convient maintenant de déterminer si ses observations sont dues au hasard ou si elles traduisent une relation dans la formation des groupes. Pour cela, on utilise la somme des χ^2 locaux, le χ^2 **total**, afin de réaliser un test statistique du χ^2

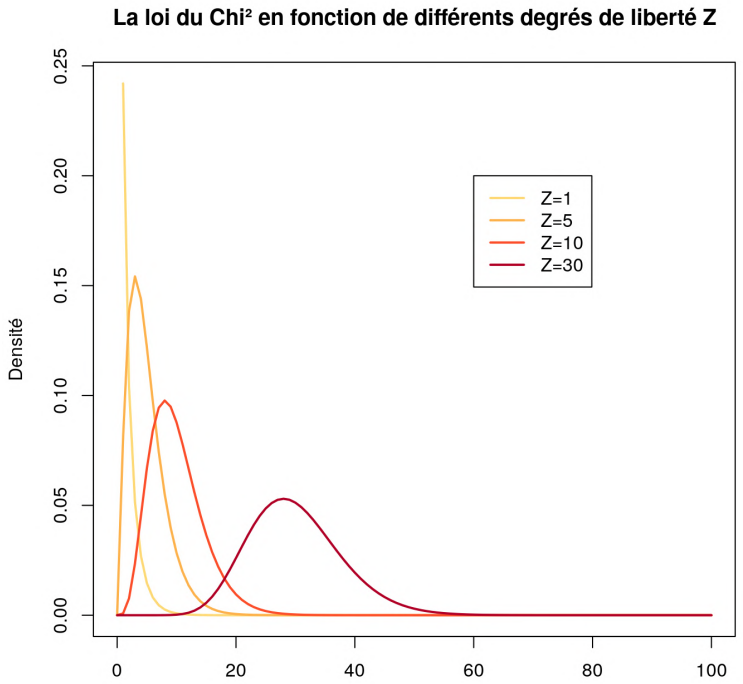
Test du χ^2 : degré de liberté

La première chose à faire avant de réaliser le test est de déterminer le nombre Z de degrés de liberté qui dépend du nombre de lignes et de colonnes du tableau de contingence. Ce nombre Z n'a rien en commun avec le test Z présenté en début de cours.

Il correspond au nombre de cases pouvant produire des déviations indépendantes les unes des autres.

Dans le cas d'un tableau de contingence à 2 lignes et 2 colonnes, ce nombre de degrés de liberté est égal à 1 puisque, la somme des déviations marginales devant être égale à zéro, il suffit de connaître la déviation d'une case pour trouver par différence les déviations de toutes les autres.

D'une manière plus générale, le **nombre de degrés de liberté** pour le Chi² est égal aux **nombre de colonnes moins une multiplié par le nombre de lignes moins une**, soit : $Z = (k-1) * (p-1)$



Déroulement du test d'indépendance du χ^2

Le but du test est de déterminer si la valeur observée du χ^2 correspond à :

- un événement fréquent ou lié au hasard (en quel cas on ne peut rejeter l'hypothèse d'indépendance)
- un événement rare (en quel cas on peut rejeter l'hypothèse d'indépendance).

1. On pose l'hypothèse H0 : "Il n'y a pas de relation entre les caractères X et Y".
2. On détermine la valeur Chi-2Obs du tableau étudié.
3. On détermine le nombre de degrés de liberté z du tableau étudié.
4. On fixe le risque d'erreur alpha de rejeter H0 à tort (ex. alpha=5%).
5. On détermine la valeur Chi-2(z,alpha) qui est la valeur de Chi-2 d'un tableau de contingence à z degrés de liberté qui ne serait dépassé que dans alpha % des cas si les variables X et Y étaient indépendantes. Cette valeur est lue dans une table du test du Chi-2 que l'on peut trouver en annexe de tous les manuels de statistique
6. On procède au test : H0 est vraie si : Chi-2Obs est inférieur ou égal à Chi-2(z,alpha)
7. Suivant le résultat du test, on accepte H0 sans pouvoir accepter H1 ou bien l'on rejette H0 et l'on accepte l'hypothèse inverse H1 ("il y a une relation de dépendance entre X et Y") avec un risque d'erreur de alpha%

Table du χ^2 et explications

Pour réaliser un test du χ^2, il faut une table de la distribution théorique du χ^2. Elle se trouve facilement sur internet ou dans les ouvrages de statistiques. Toutes les tables théorique du χ^2 sont les mêmes.

On détermine :

- Z qui est égale au degré de liberté
- α qui est égale au risque d'erreur alpha

On lit la table de la façon suivante :

- Pour un degré de liberté Z=9 et un risque d'erreur alpha égale à 5%, le χ^2 théorique est égale à 16,92

Cela signifie que pour 9 degrés de liberté, le χ^2 observé sera supérieur à 16,92 dans 5% des cas, à 21,67 dans 1% des cas

Z	α = 0.05	α = 0.01	α = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	**16.92**	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12

Table du χ^2 - exemple

Test d'indépendance des variables groupe et sexe des étudiants

Si l'on se fixe un risque d'erreur $\alpha = 0.1$ (10%), la valeur théorique du χ^2 correspondant à 1 degré de liberté est $\chi^2(1, 0.1) = 2.7055$.

La valeur du χ^2 observée est **supérieure** à cette valeur théorique (3.57>2,7055), on peut rejeter H0 et affirmer avec un risque d'erreur de 10% que les groupes et les sexes ne sont pas distribués au hasard l'un par rapport à l'autre (OBS > THEO)

ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,4549	1,0742	1,6424	2,7055	3,8415	5,4119	6,6349	10,8274
2	0,2107	1,3863	2,4079	3,2189	4,6052	5,9915	7,8241	9,2104	13,8150
3	0,5844	2,3660	3,6649	4,6416	6,2514	7,8147	9,8374	11,3449	16,2660
4	1,0636	3,3567	4,8784	5,9886	7,7794	9,4877	11,6678	13,2767	18,4662
5	1,6103	4,3515	6,0644	7,2893	9,2363	11,0705	13,3882	15,0863	20,5147
6	2,2041	5,3481	7,2311	8,5581	10,6446	12,5916	15,0332	16,8119	22,4575
7	2,8331	6,3458	8,3834	9,8032	12,0170	14,0671	16,6224	18,4753	24,3213
8	3,4895	7,3441	9,5245	11,0301	13,3616	15,5073	18,1682	20,0902	26,1239
9	4,1682	8,3428	10,6564	12,2421	14,6837	16,9190	19,6790	21,6660	27,8767
10	4,8652	9,3418	11,7807	13,4420	15,9872	18,3070	21,1608	23,2093	29,5879
11	5,5778	10,3410	12,8987	14,6314	17,2750	19,6752	22,6179	24,7250	31,2635
12	6,3038	11,3403	14,0111	15,8120	18,5493	21,0261	24,0539	26,2170	32,9092
13	7,0415	12,3398	15,1187	16,9848	19,8119	22,3620	25,4715	27,6882	34,5274
14	7,7895	13,3393	16,2221	18,1508	21,0641	23,6848	26,8727	29,1412	36,1239
15	8,5468	14,3389	17,3217	19,3107	22,3071	24,9958	28,2595	30,5780	37,6978
16	9,3122	15,3385	18,4179	20,4651	23,5418	26,2962	29,6332	31,9999	39,2518
17	10,0852	16,3382	19,5110	21,6146	24,7690	27,5871	30,9950	33,4087	40,7911
18	10,8649	17,3379	20,6014	22,7595	25,9894	28,8693	32,3462	34,8052	42,3119
19	11,6509	18,3376	21,6891	23,9004	27,2036	30,1435	33,6874	36,1908	43,8194
20	12,4426	19,3374	22,7745	25,0375	28,4120	31,4104	35,0196	37,5663	45,3142
21	13,2396	20,3372	23,8578	26,1711	29,6151	32,6706	36,3434	38,9322	46,7963
22	14,0415	21,3370	24,9390	27,3015	30,8133	33,9245	37,6595	40,2894	48,2676
23	14,8480	22,3369	26,0184	28,4288	32,0069	35,1725	38,9683	41,6383	49,7276
24	15,6587	23,3367	27,0960	29,5533	33,1962	36,4150	40,2703	42,9798	51,1790
25	16,4734	24,3366	28,1719	30,6752	34,3816	37,6525	41,5660	44,3140	52,6187
26	17,2919	25,3365	29,2463	31,7946	35,5632	38,8851	42,8558	45,6416	54,0511
27	18,1139	26,3363	30,3193	32,9117	36,7412	40,1133	44,1399	46,9628	55,4751
28	18,9392	27,3362	31,3909	34,0266	37,9159	41,3372	45,4188	48,2782	56,8918
29	19,7677	28,3361	32,4612	35,1394	39,0875	42,5569	46,6926	49,5878	58,3006
30	20,5992	29,3360	33,5302	36,2502	40,2560	43,7730	47,9618	50,8922	59,7022

Table du χ^2 - exemple

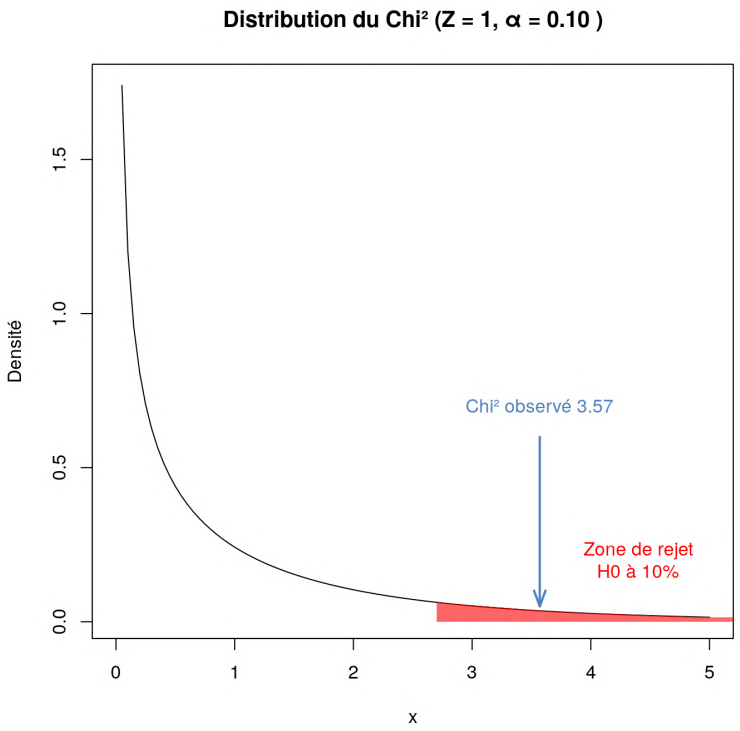
Test d'indépendance des variables groupe et sexe des étudiants

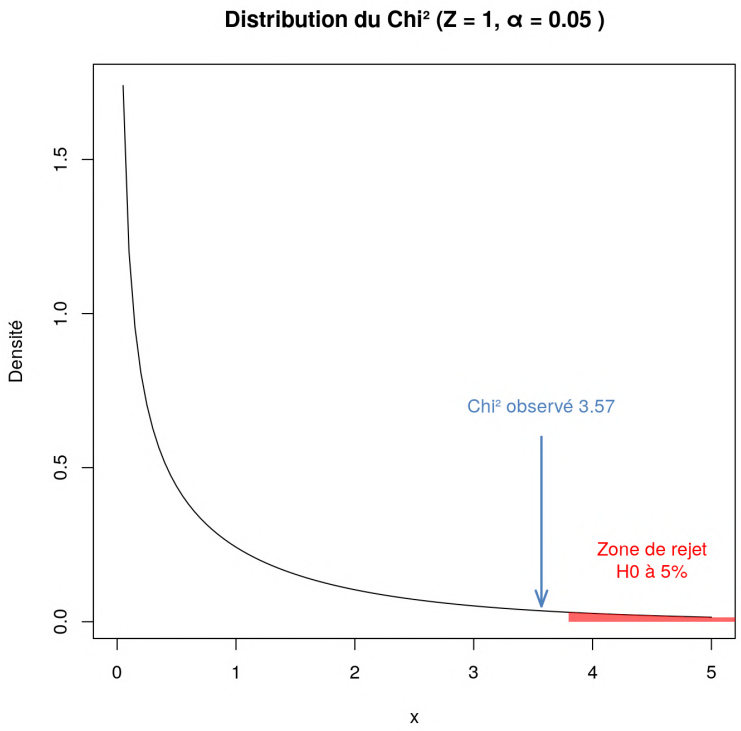
Si l'on se fixe un risque d'erreur plus faible, $\alpha = 0.05$ (5%), la valeur théorique du χ^2 correspondant à 1 degré de liberté est $\chi^2(1, 0.1) = 3.8145$.

La valeur du χ^2 observée est **inférieure** à cette valeur théorique (3.57<3,8415), et l'on ne peut plus rejeter H0. On conclue alors qu'il n'est pas possible d'affirmer qu'il existe une relation entre les deux caractères X et Y, sauf à admettre un risque d'erreur supérieur à 5%.

ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,4549	1,0742	1,6424	2,7055	3,8415	5,4119	6,6349	10,8274
2	0,2107	1,3863	2,4079	3,2189	4,6052	5,9915	7,8241	9,2104	13,8150
3	0,5844	2,3660	3,6649	4,6416	6,2514	7,8147	9,8374	11,3449	16,2660
4	1,0636	3,3567	4,8784	5,9886	7,7794	9,4877	11,6678	13,2767	18,4662
5	1,6103	4,3515	6,0644	7,2893	9,2363	11,0705	13,3882	15,0863	20,5147
6	2,2041	5,3481	7,2311	8,5581	10,6446	12,5916	15,0332	16,8119	22,4575
7	2,8331	6,3458	8,3834	9,8032	12,0170	14,0671	16,6224	18,4753	24,3213
8	3,4895	7,3441	9,5245	11,0301	13,3616	15,5073	18,1682	20,0902	26,1239
9	4,1682	8,3428	10,6564	12,2421	14,6837	16,9190	19,6790	21,6660	27,8767
10	4,8652	9,3418	11,7807	13,4420	15,9872	18,3070	21,1608	23,2093	29,5879
11	5,5778	10,3410	12,8987	14,6314	17,2750	19,6752	22,6179	24,7250	31,2635
12	6,3038	11,3403	14,0111	15,8120	18,5493	21,0261	24,0539	26,2170	32,9092
13	7,0415	12,3398	15,1187	16,9848	19,8119	22,3620	25,4715	27,6882	34,5274
14	7,7895	13,3393	16,2221	18,1508	21,0641	23,6848	26,8727	29,1412	36,1239
15	8,5468	14,3389	17,3217	19,3107	22,3071	24,9958	28,2595	30,5780	37,6978
16	9,3122	15,3385	18,4179	20,4651	23,5418	26,2962	29,6332	31,9999	39,2518
17	10,0852	16,3382	19,5110	21,6146	24,7690	27,5871	30,9950	33,4087	40,7911
18	10,8649	17,3379	20,6014	22,7595	25,9894	28,8693	32,3462	34,8052	42,3119
19	11,6509	18,3376	21,6891	23,9004	27,2036	30,1435	33,6874	36,1908	43,8194
20	12,4426	19,3374	22,7745	25,0375	28,4120	31,4104	35,0196	37,5663	45,3142
21	13,2396	20,3372	23,8578	26,1711	29,6151	32,6706	36,3434	38,9322	46,7963
22	14,0415	21,3370	24,9390	27,3015	30,8133	33,9245	37,6595	40,2894	48,2676
23	14,8480	22,3369	26,0184	28,4288	32,0069	35,1725	38,9683	41,6383	49,7276
24	15,6587	23,3367	27,0960	29,5533	33,1962	36,4150	40,2703	42,9798	51,1790
25	16,4734	24,3366	28,1719	30,6752	34,3816	37,6525	41,5660	44,3140	52,6187
26	17,2919	25,3365	29,2463	31,7946	35,5632	38,8851	42,8558	45,6416	54,0511
27	18,1139	26,3363	30,3193	32,9117	36,7412	40,1133	44,1399	46,9628	55,4751
28	18,9392	27,3362	31,3909	34,0266	37,9159	41,3372	45,4188	48,2782	56,8918
29	19,7677	28,3361	32,4612	35,1394	39,0875	42,5569	46,6926	49,5878	58,3006
30	20,5992	29,3360	33,5302	36,2502	40,2560	43,7730	47,9618	50,8922	59,7022

Illustration graphique





Pour $Z = 1$, la valeur du χ^2 théorique sera supérieure à 2.7055 dans 10% des cas. Le χ^2 observé (3.57) est bien compris dans la zone puisque supérieur à 2.7055

Pour $Z = 1$, l la valeur du χ^2 sera supérieure à 3.8415 dans 5% des cas. Le χ^2 observé est en dehors de cette zone.

Conclusions

Règles d'applications

Relativement simple à mettre en œuvre, le test du χ^2 ne peut cependant être utilisé pour tester l'indépendance de deux caractères X et Y que si certaines conditions très précises sont remplies. Les trois principales sont les suivantes :

- l'effectif total du tableau de contingence $N..$ doit être supérieur ou égal à 20
- l'effectif marginal du tableau de contingence $Ni.$ ou $N.j$ doit toujours être supérieur ou égal à 5.
- l'effectif théorique N^*_{ij} des cases du tableau de contingence doit être supérieur à 5 dans 80% des cases du tableau de contingence.

Si une des conditions n'est pas remplie, il faut faire des regroupement ou faire un test du Fischer exact.

Concepts clés

- La démarche scientifique repose sur l'émission d'hypothèse, qu'il faut tester statistiquement.
- Une théorie irréfutable n'est pas scientifique.
- Pour évaluer la relation entre deux variables discrêtes, on utilise le tableau de contingence.
- Evaluer par l'observation une relation n'est pas suffisant, il faut déterminer si elle est "réelle" (avec une certaine marge d'erreur) ou si elle est liée au hasard.
- De manière générale, on teste en comparant l'observé par rapport à une situation aléatoire. Si les deux sont proches, on en conclut qu'il n'y a pas de relation. Si les deux sont différents, on peut en conclure, avec une marge d'erreur α qu'il existe une relation.
- α détermine l'écart entre l'observé et le théorique. Il est défini par le scientifique.
- Le test du χ^2 est utile si vous voulez tester l'association entre deux variables sur un tableau de contingence.
- Ce test permet de prendre en compte qu'un écart n'a pas la même signification en fonction du contexte (écart de 5 sur un effectif de 10 vs. écart de 5 sur un effectif de 1000).
- A noter qu'il existe des statistiques, comme le V de Cramer, qui permettent de mesurer l'intensité de l'association entre les deux variables discrêtes.

Conclusion

Conclure qu'une relation existe (ou non) n'est jamais une fin en soit. Il faut évidemment comprendre et tenter d'expliquer cette relation :

- Les accidents aériens (oui/non) sont ils dépendant du type de piste des aéroports (tarmac/herbe/terre) ? Il peut également y avoir un lien avec la formation des pilotes, le type d'appareil etc.
- La réussite à un concours (oui/non) est elle dépendante de l'enseignant ? Les groupes n'ont peut-être pas été formés au hasard.