Comprendre, construire et interpréter les statistiques

Petit précis d'usage des statistiques à l'attention des non statisticiens

NOTE METHODOLOGIQUE

Note méthodologique réalisée par Léo Mignot, sociologue, Université Bordeaux 2 et Alexandre Bertin, Responsable Etudes & Diagnostic, AEC.

Version du 30 mars 2012

Sommaire

GÉNÉRALITÉS 5
Pourquoi utiliser des tests statistiques ?..... 5
Ressources..... 6

COMPARAISON DE MOYENNES : Le T de Student..... 7
Exemple : sexe et réussite scolaire..... 7
Mise en œuvre du test « t » 7

COMPARAISON DE POURCENTAGES : Test de comparaison de proportions..... 10
Exemple : sexe et utilisation d'internet..... 10
Mise en œuvre 10

DÉPENDANCE ENTRE DEUX VARIABLES : Le chi2 12
Exemple : sexe et discussions politiques..... 12
Mise en œuvre du test du chi2..... 12

NIVEAU DE CONFIANCE ET MARGE D'ERREUR 15
Définitions 15
Précisions..... 15

« ERREURS DE BASE » A EVITER 16
Evolutions (données chronologiques)..... 16
Pourcentages..... 16

GÉNÉRALITÉS

On assiste aujourd'hui à une généralisation de l'usage de données chiffrées et de l'appel aux statistiques afin de produire de l'information, d'évaluer ou de justifier la mise en place de réformes politiques, ou encore d'estimer les probables résultats d'une élection. Les statistiques sont partout ; de l'étude des caractéristiques de la population à l'affrontement de chiffres lors des débats politiques, de l'ouverture des JT à la multiplication des sondages d'opinions.

Si les bienfaits de la diffusion de l'usage des statistiques sont indéniables - meilleure connaissance et description de phénomènes, nouvelles façons de traiter l'information (datajournalisme, …), etc. –, cette propagation ne va pas sans poser quelques problèmes.

Ainsi, mésinterprétations des données, surinterprétations de résultats, conclusions hâtives, négligence des marges d'erreurs lors de l'interprétation de sondages ou encore biais d'échantillonnage et de représentativité[1] sont monnaie courante lorsque l'on utilise des « chiffres » sans le sérieux et le soin y étant nécessaires. A ce titre, si l'on dit souvent que « l'on peut faire dire ce que l'on veut aux chiffres », cela est bien moins vrai si l'on s'astreint à la rigueur intellectuelle et scientifique que suppose l'usage de telles données.

Pourquoi utiliser des tests statistiques ?

La quasi totalité des études réalisées raisonnent sur la base d'échantillons. En effet, pour des raisons de faisabilité technique et de coût financier, il est presque toujours impossible de travailler sur la population totale (la population réelle) concernée par une enquête. On doit donc se contenter de travailler sur une partie de cette population : on va alors en construire un échantillon représentatif[2].

Toutefois, même si les règles d'échantillonnage sont scrupuleusement respectées, il existe toujours une part de hasard. Autrement dit, « un échantillon, même tiré au sort, n'est pas le reflet exact de la population dont il est issu »[3]. Il existe donc un écart entre un phénomène mesuré dans un échantillon et sa vraie valeur dans la population totale : on parle de *fluctuations d'échantillonnage*. Commet savoir alors si une tendance observée dans un échantillon est bien valable pour l'ensemble de la population (et n'est donc pas le seul fruit du hasard) ?

C'est pour répondre à cette question que l'on fait appel aux tests statistiques. Ceux-ci permettent de s'assurer que les résultats observés dans un échantillon sont généralisables (ou non) à l'ensemble de la population concernée. Ainsi, sous leur contrôle et avec un risque d'erreur connu[4], on saura si l'on peut étendre à l'ensemble de la population des conclusions établies sur un échantillon de cette dernière.

Nous verrons ici comment mettre en place trois des principaux tests employés en statistique : le *t de Student* (utilisé quand on cherche à comparer des moyennes), les tests de comparaison de proportion (mis en place quand on cherche à comparer des pourcentages) et le *chi2* (mobilisé lorsque l'on étudie l'existence d'une corrélation entre deux variables).

---

[1] La multiplication des « questions du jour » (RTL, Europe 1, BFMTV, etc.) en est emblématique.

[2] On va tirer au sort un certain nombre d'individus parmi la population concernée. Différents types d'échantillonnages peuvent être mis en place (échantillonnage aléatoire, systématique, stratifié, par quotas, etc.).

[3] R. Michel, L. Ollivier-Gay, A. Spiegel, J-P. Boutin, « Les test statistiques : intérêt, principe et interprétations », *Med Trop 2002*, n° 62, 2002, pp. 561-563.

[4] On fixe généralement le degré (ou niveau) de confiance à 95%, soit 5% de risque de se tromper.

Ressources

La maîtrise de ces outils serait toutefois peu utile sans matériau sur lequel les appliquer. Aussi, on trouvera ci-dessous un aperçu des principaux organismes fournisseurs de données quantitatives (bases de données, notes statistiques, rapports) :

INSEE (Institut National de la Statistique et des Etudes Economiques)
http://www.insee.fr/fr/

INED (Institut National d'Etudes Démographiques)
http://www.ined.fr/

PISA (Programme International pour le Suivi des Acquis des élèves)
http://www.oecd.org/document/24/0,3746,en_32252351_32235731_38378840_1_1_1_1,00.html
http://www.oecd.org/department/0,3355,fr_2649_35845621_1_1_1_1_1,00.html

ESS (European Social Survey)
www.europeansocialsurvey.org

OCDE (Organisation de Coopération et de Développement Économiques)
http://www.oecd-ilibrary.org/fr
http://www.oecd.org/document/8/0,3746,fr_21571361_44315115_45487048_1_1_1_1,00.html

Portails :

Plus généralement, on trouvera à l'adresse suivante une liste d'organismes fournissant une grande quantité de statistiques et bases de données sur des sujets variés :
http://www.cyberjournalisme.net/ressources/chiffres-cles-statistiques

De même, plusieurs sites donnant accès à des données d'enquêtes et de sondages sont recensés aux adresses suivantes :
http://doc.sciencespo-lyon.fr/Ressources/Liens/liens.html?th=13#93
http://www.cyberjournalisme.net/ressources/instituts-de-sondage

COMPARAISON DE MOYENNES : Le T de Student

Le « t de Student » est un test de différence des moyennes utilisé lorsque l'on cherche à comparer deux moyennes entre elles. Plus précisément, il permet de mesurer la significativité statistique des différences entre deux moyennes.

Nb : Contrairement à ce que l'on pourrait croire, le test t n'a pas été créé par monsieur « Student ». Il est en réalité l'œuvre de William Sealy Gosset, salarié de Guinness. Celui-ci, ne pouvant divulguer ce test sous son propre nom (pour des raisons de secret industriel) prit le pseudonyme de « Student ». Ainsi, le monde de la statistique doit l'un de ses tests les plus utilisés à un employé de brasserie chargé du contrôle qualité de la production de bière stout.

Exemple : sexe et réussite scolaire

Clarifions la chose en prenant un exemple concret : l'étude de la réussite scolaire selon le sexe. Imaginons que l'on souhaite comparer la réussite scolaire des hommes et des femmes en étudiant la différence des notes qu'ils obtiennent lors de leur scolarité.
Il semble délicat de questionner l'ensemble des 15 millions d'élèves et étudiants français : on va donc se contenter d'interroger un échantillon de ces individus.

Prenons ici les moyennes générales d'une classe composée de 8 garçons et 9 filles :

Moyenne générale des garçons	Moyenne générale des filles
08	15
15	16
14	09
07	14
13	13
12	08
11	16
13	13
	16

Les garçons obtiennent ici une moyenne de 11,625 contre 13,333 pour les filles. Les garçons ont donc une moins bonne moyenne que les filles et l'on a tôt fait de conclure que les filles réussissent mieux que les garçons à l'école.

Toutefois, peut-on réellement faire de telles conclusions ? Autrement dit, de l'écart observé dans notre (modeste) échantillon, peut-on conclure que, dans l'ensemble de la population, les filles réussissent mieux que les garçons[5] ?

C'est ici que le *t* de Student fait son entrée : **il va nous indiquer si l'on peut généraliser nos résultats à l'ensemble de la population sans avoir trop de risque de se tromper**. Autrement dit, il évalue si la différence observée entre nos moyennes est statistiquement significative.

Mise en œuvre du test « t »

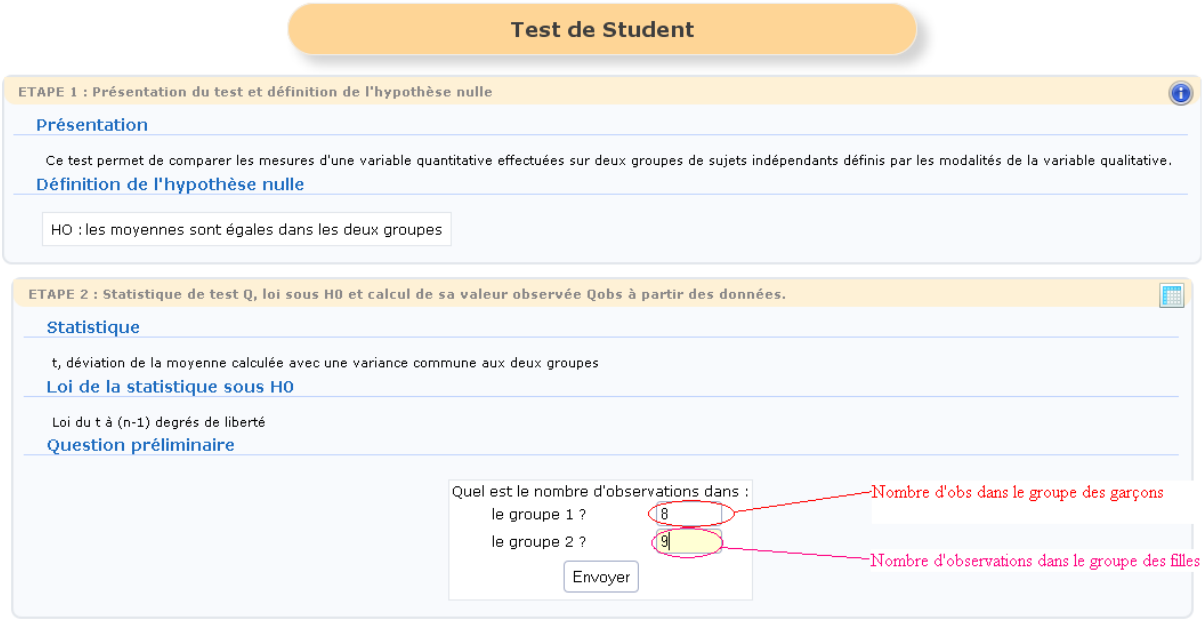
Dans sa forme originale, le t de student suppose d'effectuer un calcul relativement complexe et peu commode à mettre en œuvre sans connaissance statistique. Heureusement, depuis, l'informatique a fait son apparition, des logiciels libres ont été mis à disposition et

[5] La différence observée est-elle réelle ou est-elle le fruit du hasard de l'échantillonnage ? Les résultats auraient-ils été les mêmes si l'on avait pris une autre classe ?

certains sites proposent des outils permettant d'obtenir les résultats du test sans faire un seul calcul. C'est notamment le cas de BiostaTGV (http://marne.u707.jussieu.fr/biostatgv/). Voyons donc comment faire.

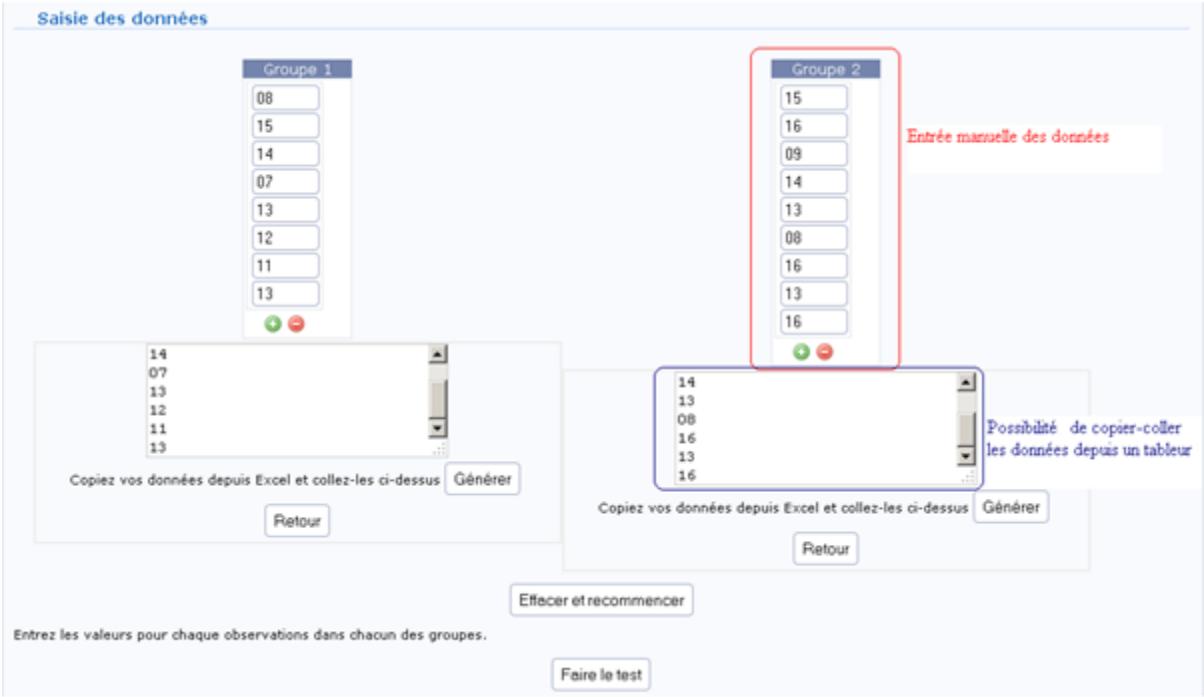
Etape 1 :

Il convient tout d'abord de se rendre à l'interface de calcul du t de student disponible ici : http://marne.u707.jussieu.fr/biostatgv/?module=tests/student



On nous demande ici de renseigner le nombre d'observations dans le groupe 1 et le groupe 2. Il s'agit dans notre exemple du groupe des garçons (8 observations) et des filles (9 observations). On clique sur « envoyer ».

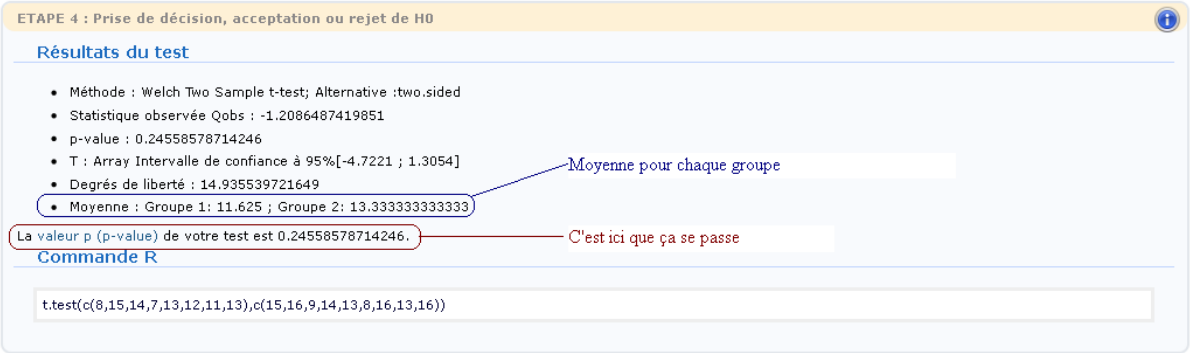
Etape 2 :



Il s'agit ici d'insérer les valeurs de chaque observation (dans notre cas la moyenne générale de chaque élève). On peut procéder manuellement ou effectuer un copier-coller

depuis un tableur (notamment lorsque l'on travaille sur de grands échantillons). Il suffit ensuite de « faire le test ».

Etape 3 :



Les résultats du test s'affichent sans que l'on ait besoin d'effectuer de calcul. La moyenne pour chaque groupe nous est donnée, ainsi qu'une série d'informations. Mais comment interpréter ces résultats ?

Pour que l'on puisse affirmer que les moyennes sont statistiquement différentes, il faut que *la p-value soit inférieure à 0,05*[6] (*faire attention à la présence de puissances* lors de la lecture du résultat).

Dans notre exemple, la p-value est d'environ 0,2456 et est supérieure à la limite fixée de 0,05. On ne peut donc pas dire que la différence observée entre les moyennes est statistiquement significative. Autrement dit, les résultats observés pour notre échantillon ne nous permettent pas d'affirmer que l'ensemble des filles ont une réussite scolaire supérieure à celle des garçons.

[6] Il s'agit en fait d'un seuil de risque de 5% (ou niveau de confiance de 95%). Ainsi, si la « p-value » est inférieure à 0,05, on peut affirmer avec moins de 5% de chances de se tromper que les moyennes sont significativement différentes.

COMPARAISON DE POURCENTAGES : Test de comparaison de proportions

L'utilisation de pourcentages et leur comparaison sont aujourd'hui parmi les formes principales de la description de phénomènes ou de la production d'informations. Aussi, lorsque l'on étudie les résultats d'une enquête ou d'un sondage et que l'on en vient à **comparer des pourcentages** en étant issus, il convient de **s'assurer qu'il existe une différence statistiquement significative entre ces pourcentages**. C'est ce que permettent les tests de comparaison de proportions.

Exemple : sexe et utilisation d'internet

Considérons que l'on souhaite étudier l'influence du sexe sur l'utilisation d'internet. Une étude sur un échantillon de 1800 individus nous donne les résultats suivants :

	N'utilise pas (ou plus) internet	Utilise internet	Total
Femme	316	718	1034
Homme	197	569	766
Total	513	1287	1800

Ce **tableau croisé** (ou **tableau de contingence**) semble nous indiquer que l'utilisation d'internet est influencée par le sexe. En effet, 74,28% des hommes (569 sur 766) de notre échantillon affirment utiliser internet contre 69,44% des femmes (718 sur 1034).

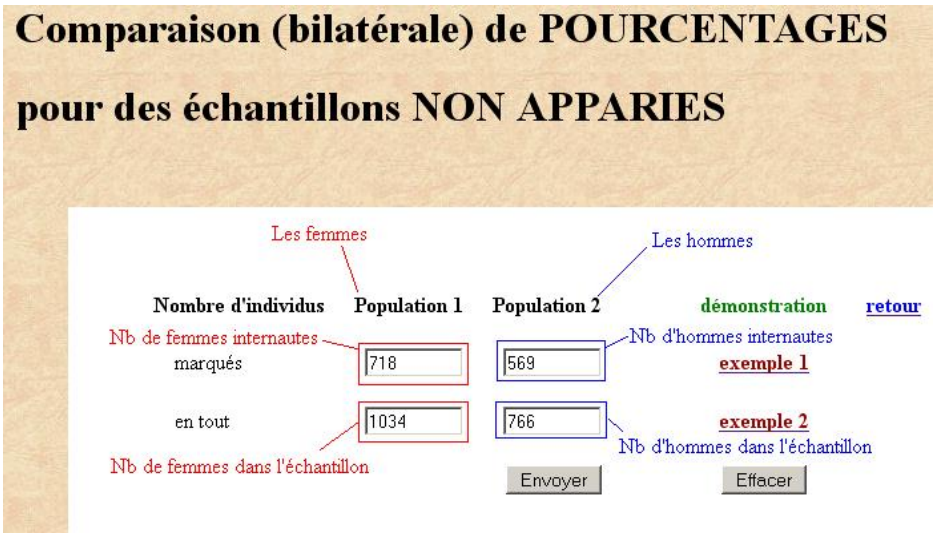
Toutefois, la différence observée ici est-elle robuste ? Ne doit-on pas ces écarts à des aléas d'échantillonnage ?

C'est pour nous assurer que la **différence observée entre ces pourcentages est bien statistiquement significative** que nous allons réaliser un test de comparaison de proportions.

Mise en œuvre

Le test de comparaison de proportions n'est actuellement pas disponible sur le site BiostaTGV. Il est toutefois possible de réaliser un tel test à l'adresse suivante : http://www.info.univ-angers.fr/~gh/wstat/comppct.php

Etape 1 :



Il nous faut dans un premier temps saisir nos données. On distingue donc nos deux populations (les femmes / les hommes) et l'on saisit pour chacune l'effectif total ainsi que le nombre d'individus « marqués » (soit ici les individus qui sont internautes). On peut ensuite « envoyer ».

Etape 2 :

Comparaison bilatérale de Pourcentages

(échantillons NON APPARIES)

```
Population 1 :
     718  individus marqués
    1034  individus en tout
          soit une proportion de  69.4 %      % d'internautes chez les femmes

Population 2 :
     569  individus marqués
     766  individus en tout
          soit une proportion de  74.3 %      % d'internautes chez les hommes

Si on réunit les deux populations, on obtient
    1287  individus marqués
    1800  individus en tout
          soit une proportion de  71.5  %     % d'internautes dans l'ensemble de la population

 Valeur de l'écart-réduit :    2.25
    au seuil de 5 % soit la valeur 1.96
    on  peut rejeter l'hypothèse que les pourcentages sont égaux.   Le résultat est ici

 La p-value bilatérale est sans doute 0.02442032 soit en arrondi 0.024
```

Le site nous fournit alors les résultats du test. On retrouve le % d'individus marqués pour chaque population (ici le pourcentage d'internautes selon le sexe), suivi du résultat du test. Celui-ci nous indique explicitement si l'on peut ***accepter ou rejeter « l'hypothèse que les pourcentages sont égaux »***[7]. Si *l'on ne peut pas rejeter cette hypothèse*, on en conclut que la différence entre les pourcentages *n'est pas statistiquement significative*. A l'inverse, si *l'on peut rejeter cette hypothèse*, on conclura que la différence entre les pourcentages est *bien statistiquement significative*.

Dans notre exemple, « on peut rejeter l'hypothèse que les pourcentages sont égaux » (la p-value est d'environ 0,024 et est donc inférieure à limite fixée de 0,05). On peut donc dire que la différence observée entre les pourcentages d'internautes parmi les hommes et les femmes est statistiquement significative. Autrement dit, les résultats observés pour notre échantillon nous permettent d'affirmer que, dans l'ensemble, les hommes sont significativement plus connectés à internet que les femmes (74,3 contre 69,4%).

[7] On accepte ou rejette cette hypothèse pour un seuil de risque de 5% (ou niveau de confiance de 95%). Ici la p-value nous est également donnée et il nous est donc possible de vérifier qu'elle est bien inférieure 0,05.

DÉPENDANCE ENTRE DEUX VARIABLES : Le chi2

Le chi2 (également noté Khi2 ou X^2) est un test d'inférence statistique permettant d'évaluer s'il existe une **relation statistiquement significative entre deux variables** – ou si, à l'inverse, celles-ci sont **indépendantes**[8].

On le met en œuvre pour savoir si une relation découverte au sein de l'échantillon étudié est valable à l'échelle de la population totale dont est tiré cet échantillon. Autrement dit, le chi2 nous aide à déterminer si une relation observée dans un échantillon est généralisable à l'ensemble de la population.

Exemple : sexe et discussions politiques

Prenons un exemple et imaginons que l'on souhaite étudier la propension des individus à parler de politique. On cherchera plus spécifiquement à voir si cette appétence pour les discussions politiques varie entre les hommes et les femmes. Ne pouvant interroger l'ensemble des Français, nous décidons de nous adresser à un échantillon de 2000 individus.

Voici les réponses (fictives) à la question « *Vous arrive-t-il de discuter de politique ?* » selon le sexe des répondants :

Sexe / discussions politiques	**Oui**	**Non**	**Total**
Hommes	540	360	900
Femmes	330	770	1100
Total	870	1130	2000

Les résultats sont ici présentés sous la forme d'un tableau croisant *le sexe* et *le fait de discuter ou non de politique* : on parle alors de **tableau croisé** ou de **tableau de contingence**.

Une lecture rapide de ce tableau et le calcul de quelques pourcentages semblent nous indiquer que la tenue de discussions politiques est fortement influencée par le sexe des individus. En effet, 60% des hommes de notre échantillon affirment discuter de politique contre seulement 30% des femmes.

Peut-on pour autant affirmer que, à l'échelle de la population française, les hommes ont une plus forte tendance à discuter de politique que ne le font les femmes ? Les conclusions tirées de notre échantillon sont-elles généralisables à la population totale ?

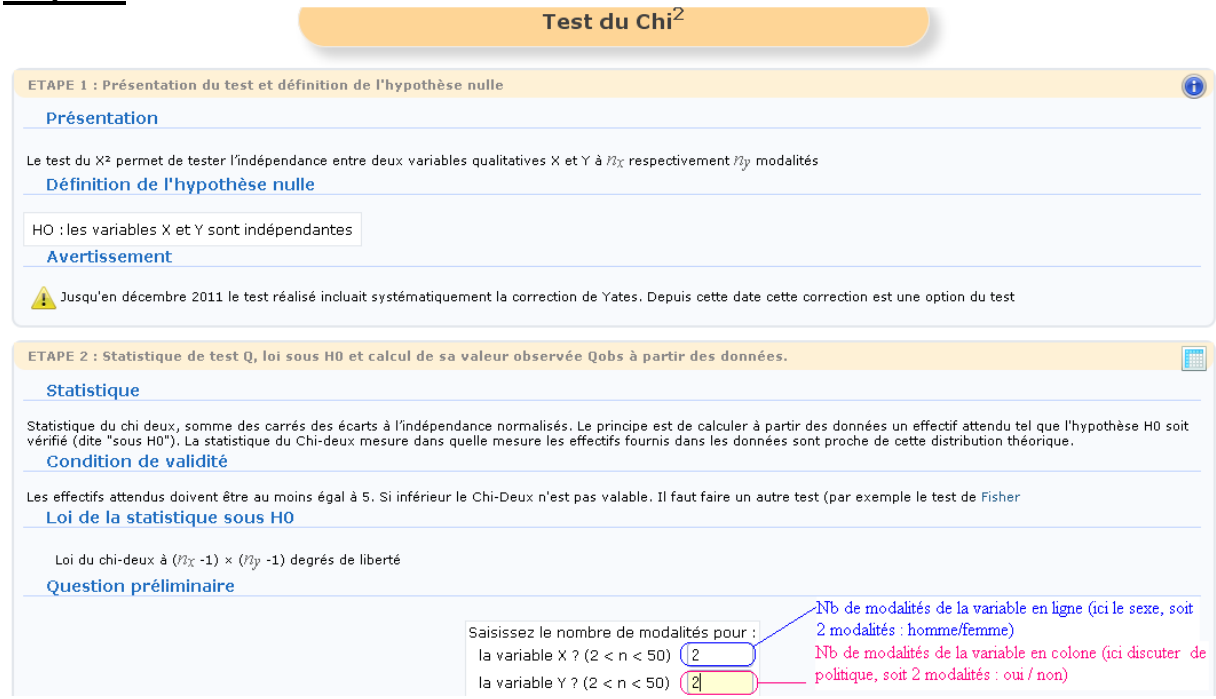
C'est afin de pallier cette difficulté que le test du chi2 va être mis en place. Celui-ci va nous indiquer **si l'on peut généraliser nos résultats à l'ensemble de la population sans avoir trop de risque de se tromper**. Autrement dit, il évalue si la relation observée entre nos variables est statistiquement significative.

Mise en œuvre du test du chi2

Réalisée à la main la procédure du test du chi2 s'avère longue et fastidieuse. Pour nous simplifier la tâche nous allons de nouveau avoir recours au site BiostaTGV. Le test du chi2 est disponible à l'adresse suivante : http://marne.u707.jussieu.fr/biostatgv/?module=tests/chideux

[8] La procédure du chi2 implique de comparer les « ***effectifs réels*** » (tels qu'existant dans notre échantillon) aux « ***effectifs théoriques*** » (ceux que l'on aurait dû observer si les variables étaient parfaitement indépendantes).

Etape 1 :

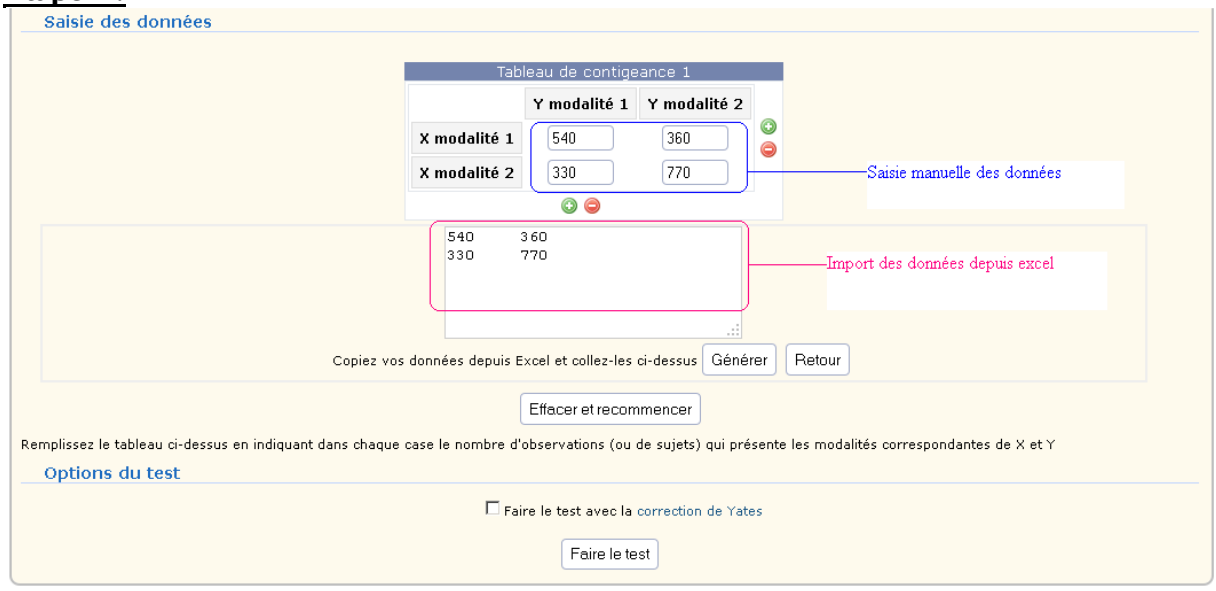


En premier lieu il nous faut préciser le nombre de modalités pour chaque variable. On a ici deux variables :

- le sexe, qui présente 2 modalités : Homme / Femme
- le fait de discuter de politique, qui présente 2 modalités : Oui / Non

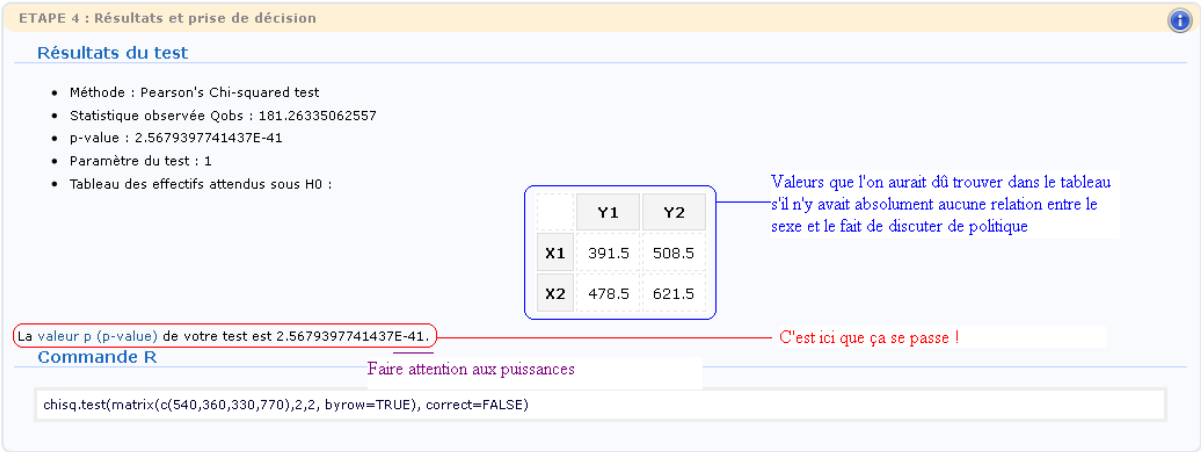
La variable **X correspond à la variable en ligne** (ici le sexe), la variable **Y à la variable en colonne** (ici le fait de discuter de politique).

Etape 2 :



Lors de cette étape il nous est demandé de rentrer les valeurs présentes dans notre tableau croisé. **Attention, on doit rentrer les valeurs dites « hors marge », c'est-à-dire sans les totaux (et sans le nom des modalités)**. On peut ensuite lancer le test.

Etape 3 :



Ici encore les résultats du test nous sont donnés sans qu'il soit nécessaire d'effectuer de calcul. La significativité de la relation nous est donnée par la valeur p (p-value). *Il convient de prendre garde à la présence de puissances* lorsqu'on lit la p-value (puissance de 10, notée E).

Pour que l'on puisse affirmer qu'il existe une relation entre les variables étudiées (ici le sexe et le fait de discuter de politique)**, il faut que *la p-value soit inférieure à 0,05*[9].**

Dans notre exemple la p-value est de 2.5679397741437E-41 et est donc inférieure au seuil fixé de 0,05. On peut donc affirmer qu'il existe une relation statistiquement significative entre le sexe d'un individu et le fait qu'il discute ou non de politique. Plus précisément, le fait d'être un homme favorise le fait de discuter de politique (il y a attraction) alors qu'être une femme freine la tenue de ce type de discussions (il y a répulsion).

> *Nb : **Corrélation ne signifie pas causalité**. Il faut prendre garde à ne pas surinterpréter les résultats et garder à l'esprit que l'existence d'une relation statistiquement significative entre des variables ne signifie pas que l'une détermine l'autre.*
> Exemple : *Il est prouvé qu'il existe une corrélation entre le nombre de glaces consommées et le nombre de cambriolages. Peut-on pour autant en conclure que le fait de manger des glaces entraine un sursaut des cambriolages ? Evidemment non. Les résultats sont en réalité ici le fait d'une variable cachée : la consommation de glaces augmente en été, tout comme le nombre de cambriolage (les voleurs profitant de l'absence des propriétaires lors de la période estivale pour réaliser leurs méfaits).*

[9] Il s'agit en fait d'un seuil de risque de 5% (ou niveau de confiance de 95%). Ainsi, si la « p-value » est inférieure à 0,05, on peut affirmer avec moins de 5% de chances de se tromper qu'il existe une relation statistiquement significative entre les variables étudiées.

NIVEAU DE CONFIANCE ET MARGE D'ERREUR

Lors de l'interprétation de sondages d'opinion ou de données d'enquête, on oublie souvent de lire la marge d'erreur qui accompagne les résultats (quand celle-ci est clairement indiquée). Cette lecture est pourtant cruciale et permet d'estimer la robustesse des résultats.

Définitions

La **marge d'erreur** estime l'étendue dans laquelle est en réalité contenue la valeur que l'on cherche à mesurer. La marge d'erreur nous donne en fait une « fourchette » de valeurs dans laquelle est contenue la véritable valeur. Ainsi, un sondage attribuant 32% d'intentions de vote à un candidat avec une marge d'erreur de +/- 2%[10] signifie que l'intention réelle de vote pour ce candidat est comprise entre 30 et 34% (32-2=30 ; 32+2=34).

Exemple : Si une étude indique avec une marge d'erreur de 2,5% que 35% des individus souhaitent changer de voiture, il faut en réalité lire qu'entre 32,5 et 37,5 % (35% +/- 2,5) des individus ont l'intention de changer de voiture.

On comprend donc que plus la marge d'erreur est grande, plus on doit être prudent dans l'interprétation des résultats : plus la marge d'erreur est forte, moins les résultats sont précis.

Le **niveau de confiance** permet d'estimer la fiabilité des résultats. Il correspond au pourcentage de chances que la véritable valeur soit effectivement comprise dans la marge d'erreur. Autrement dit, le niveau de confiance donne la probabilité que les résultats trouvés (+ ou - la marge d'erreur) soient exacts. Le plus souvent, le niveau de confiance est fixé à 95%.

Exemple : Une enquête présentant un niveau de confiance de 95% et une marge d'erreur de +/- 5% nous apprend que 50% des adolescents fument.
Cela signifie que, avec 95% de chances de ne pas se tromper, on peut affirmer que 45 à 55% (50% +/- 5) des adolescents fument.

Précisions

La marge d'erreur est calculée en prenant en compte deux facteurs principaux : la taille de l'échantillon et le niveau de confiance. La marge d'erreur est généralement donnée pour un pourcentage observé de 50% car elle est alors maximale. Le tableau suivant permet de se faire une idée de la marge d'erreur à laquelle on est confrontée pour un pourcentage donné et selon la taille de l'échantillon (ici pour un niveau de confiance de 95%)[11].

	Pourcentage				
Taille de l'échantillon	50%	60% ; 40%	70% ; 30%	80% ; 20%	90% ; 10%
500	4,38	4,29	4,02	3,51	2,63
1000	3,10	3,04	2,84	2,48	1,86
1500	2,53	2,48	2,32	2,02	1,52
2000	2,19	2,15	2,01	1,75	1,31

Pour des résultats plus précis (tester un pourcentage donné avec une taille d'échantillon donnée), le calculateur suivant peut s'avérer utile pour estimer la marge d'erreur liée à une enquête : http://www.rmpd.ca/calculators.php

NB : La notion de « marge d'erreur » ne recouvre pas toutes les sources d'erreurs possibles. En effet, seule l'erreur liée à l'échantillon est ici mesurée. Or d'autres biais peuvent exister : non-réponses, erreur d'observation / biais de collecte, erreur de calcul, défaut de couverture, etc.

[10] Si par abus de langage on dira souvent qu'une marge d'erreur est (par exemple) de +/-5%, il faudrait en réalité dire qu'elle est de + /- 5 points de pourcentage.

[11] Exemple de lecture : Si l'échantillon est de 1000 individus et que le pourcentage que l'on observe pour une caractéristique donnée est de 70%, alors la marge d'erreur est de 2,84. Autrement dit, la véritable valeur du pourcentage est comprise entre 67,16 et 72,84%.

« ERREURS DE BASE » A EVITER

Evolutions (données chronologiques)

Les coefficients multiplicateurs ne s'additionnent pas : ils se multiplient.

Ainsi, si de 2009 à 2010 le nombre de chômeurs a été multiplié par 1,3 puis qu'il est multiplié par 1,7 de 2010 à 2011 alors, sur la période 2009-2011, le nombre de chômeurs a été multiplié par 2,21 (1,3x1,7) et non pas par 3 (1,3+1,7).

De même, **on n'additionne (ou ne soustrait) pas des taux de variation** :

Exemple : Le prix d'un objet valant 100 euros augmente de 70%. Quelques mois plus tard le prix diminue de 70%. Ne prenant pas garde, d'aucuns auront tôt fait de considérer que le vélo est revenu à son prix initial de 100 euros. Hors ce n'est pas le cas :

Prix initial		**Prix après l'augmentation de 70%**		**Prix après la baisse de 70%**
100€	$\longrightarrow$	$100 + 70 \times \frac{100}{100} = 170€$	$\longrightarrow$	$170 - 70 \times \frac{170}{100} = 51€$

Le prix final est donc de 51€ (et non pas de 100€). Une hausse de 70% puis une diminution successive de 70% ne revient donc pas à une évolution globale de 0% (70% - 70%). Ici il y a au total une baisse de 49% du prix.

En fait, pour trouver l'évolution globale du prix, il faut ici encore multiplier les évolutions :

$(1 + \frac{70}{100}) \times (1 - \frac{70}{100}) = 1,7 \times 0,3 = 0,51$. Et 100€ x 0,51 = 51€, soit une baisse de 49%.

Pourcentages

Il convient tout d'abord de bien faire la *distinction* entre *pourcentages en ligne* et *pourcentages en colonne*. Prenons l'exemple suivant croisant le sexe et l'utilisation d'internet :

	Non, je ne l'ai jamais utilisé	Non, je ne l'utilise plus	Oui, je l'utilise	Total
Femme	248	68	718	1034
Homme	161	36	569	766
Total	409	104	1287	1800

Ici, environ 24% des femmes affirment n'avoir jamais utilisé internet (248/1034x100). Toutefois, 60,64% des personnes n'ayant jamais utilisé internet sont des femmes (248/409x100). La lecture des pourcentages en ligne ou en colonne renvoie donc à 2 réalités distinctes ; on décrit 2 phénomènes différents.

Lorsque l'on utilise des pourcentages, il convient également de garder à l'esprit la *taille et la structure de la population de référence* de laquelle ils sont tirés :

	Non, je ne l'ai jamais utilisé	Non, je ne l'utilise plus	Oui, je l'utilise	Total
Femme	23,98 %	6,58 %	69,44 %	100 %
Homme	21,02 %	4,70 %	74,28 %	100 %
Total	22,72 %	5,78 %	71,50 %	100 %

Dans notre échantillon, 74,28% des hommes utilisent internet contre 69,44% des femmes. Il y a donc plus d'hommes que de femmes qui utilisent internet ? Erreur : on oublie ici qu'il y a plus de femmes que d'hommes dans notre échantillon. Si les hommes semblent proportionnellement plus connectés à internet, il n'en reste pas moins que dans l'absolu, en effectif réel, il y a plus de femmes (718) internautes que d'hommes (569) dans notre échantillon. On parle d'*effet de structure*.

Il faut également faire attention aux abus de langage lorsque l'on compare des pourcentages dans le temps. Une évolution s'exprime en *points de pourcentage*.

Exemple : Si la proportion d'ouvriers au sein d'une entreprise est passée de 25% en 2010 à 20% en 2011, il y a eu une diminution de 5 points de pourcentage (et non pas une baisse de 5%).