Exploiter statistiquement des données d'enquête

Intervention de Denis la Mache
Docteur en sociologie
Responsable de formation au CREPS Poitou-Charentes

I – Lire un tableau statistique : tableaux uni-variés et tableaux de contingence

Dans un précédent cours, nous avons vu comment se construisent une variable et ses modalités. Je vous propose maintenant de voir comment on résume ces variables et leurs distributions. On se situe désormais en aval de la collecte statistique : on a les chiffres, il s'agit de les « faire parler ».

Il s'agit donc :

- D'examiner variables séparément
- D'analyser les relations qu'elles peuvent entretenir entre elles

Avertissement : Nous allons amorcer ici un propos plus mathématique. Mais il ne s'agit pas pour autant de faire un cours de mathématiques. L'objectif sera d'aller au plus simple. Il n'y aura pas de démonstration et je proposerai les formules dans leurs version les plus simples et pratiques possible.

A – La question de l'intervalle de confiance

Une fois que l'on a interrogé l'échantillon et récolté les données, une première question se pose : *Quelle est leur fiabilité ? Quelle chance a-t-on que le résultat obtenu sur l'échantillon soit identiques dans la population parente ?*

Ceci amène à traiter la question de l'intervalle de confiance.

Pour comprendre le principe, partons d'un exemple : Un résultat de 53 % obtenu sur échantillon a 100 % de chance qu'il soit compris entre O et 100 % dans la population parente. Il a un peu moins de chances d'être compris entre 10% et 90%, encore moins d'être compris entre 52 et 54 etc.

L'intervalle de confiance est donc : l'intervalle dans lequel la valeur réelle dans la population à le plus de chances de se trouver

Autrement dit : Calculer intervalle de confiance, c'est calculer intervalle dans lequel se trouve le résultat en fonction d'une probabilité plus ou moins grande de se tromper.

Cet intervalle dépend donc de 2 éléments :

- **L'échantillon** (sa composition et la valeur qui y est observée) : on va en déduire « E » : l'erreur qu'il est susceptible de générer quant à la précision de l'observation
- **Le risque que l'on accepte de prendre pour donner une réponse juste** (plus on veut donner une réponse précise plus on prend de risque de se tromper) : on va en déduire z : l'amplitude de l'incertitude liée au risque pris

Formule applicable sur un échantillon de plus de 100 individus :
Celle-ci repose sur des fondements probabilistes que n'aborderons pas ici)

E = erreur d'échantillonnage

p = pourcentage observé dans l'échantillon
n = taille de l'échantillon

$$E = \pm \sqrt{\frac{p(1-p)}{n}}$$

L'erreur d'échantillonnage va être apprécié en fonction d'un risque admis auquel va être affectée la valeur z (qui va agir comme coefficient multiplicateur de E) : Moins on admet un risque important plus grande sera l'incertitude sur l'intervalle probable.

Extrait de la table de Laplace-Gauss

Risque admis	0.01 (99%)	0.02 (98%)	0.03 (97%)	0.04 (96%)	0.05 (95%)	0.06 (94%)	0.07 (93%)	0.08 (92%)	0.09 (91%)
Valeur de z (arrondie)	2.58	2.32	2.17	2.05	1.96	1.88	1.81	1.75	1.69

Intervalle de $p \pm E_z$ **confiance :**

Exercice : On a demandé à 1000 pers si elles avaient une voiture. 74 % ont répondu ***oui***. Quel est l'intervalle de confiance pour être certain à 99% de chances de ne pas se tromper que cette proportion est identique dans la population parente ?

$$E = \pm \sqrt{\frac{0,74(1-0,74)}{1000}}$$

$$E = \pm 0,01387$$

L'intervalle de confiance pour un risque 0,01 correspond à P± 2,58 E. Le résultat dans la population parente à donc 99% de chances de se trouver entre 70,5% et 77,5%.

B – l'examen des variables séparément : le tri à plat

Il s'agit de décrire l'information récoltée pour chaque variable. On va commencer par transcrire le *comportement* de la variable :

- modalités de la variable
- occurrence de la modalité

Ex : 810 étudiants sont interrogés : « Etes vous d'accord avec le fait que les gens cultivés écoutent plus France inter qu'RTL ? »

Modalité	**Effectifs concernés**	**Pourcentage**
Tout à fait d'accord	527	65
Plutôt d'accord	218	27

Plutôt pas d'accord	35	4
Pas du tout d'accord	6	1
Non réponse	24	3
total	**810**	**100**

Quelques remarques préliminaires pour alimenter le commentaire d'un tri à plat :

- les pourcentages sont intéressants mais il faut toujours s'enquérir des effectifs qui ont servis de base aux calculs. Ce sont eux qui renseignent sur l'échantillon (donc sur la fiabilité des résultats) et servent de bases aux calculs que verront ensuite. Un document statistique qui ne fait apparaître que des pourcentages n'est pas un document statistique exploitable.

- Etant donné la marge d'incertitudes (voir précédemment), les pourcentages sont souvent arrondis (Il est en effet inutile d'aller jusqu'à 1à chiffres après la virgule sur un nombre qui est d'emblée une approximation). Donc les totaux ne sont pas toujours égaux à 100%

- Les non réponses peuvent être gardées ou enlevées (les pourcentages seront adaptés en conséquence). D'un point de vue statistique, les 2 sont possibles : leur présence ou absence est donc un choix de l'analyste. Les chiffres ne disent que ce qu'on leur demande de dire !

  Ex : lors d'un sondage électoral, enlever les non réponse permet de prévoir le résultat (dans les limites de ce qui a déjà été dit précédemment). Les garder donne des renseignements sur l'incertitude des électeurs

1- Résumé statistique : les valeurs centrales

Il existe 3 types de valeurs centrales : le mode, la médiane, la moyenne

- Mode : (se calcule sur tous types de variables) valeur ou classe qui a l'effectif le plus grand

Ex : La réponse la plus donnée à un questionnaire

- Médiane : (se calcule sur les variables métriques et ordinales) valeur qui subdivise l'échantillon en 2 sous-parties égales. Autrement dit, la population est répartie en deux sous-parties égales

C'est l'observation qui est au milieu de la distribution ordonnée, c'est-à-dire qui voit les effectifs se distribuent autant en dessous que au dessus. On peut parler selon les cas de valeurs ou de classes médianes.

Le cas d'une variable ordinale :

Ex : A la question « Notez votre satisfaction entre 1 et 10 », il y a eu 6 réponses : 7, 4, 4, 6, 4, 5
Pour connaître la médiane, on classe les réponses avec leurs effectifs respectifs

Note	1	2	3	4	5	6	7	8	9	10
Effectifs	0	0	0	3	1	1	1	0	0	0

IL y a 3 individus jusqu'à 4 (inclus) et 3 à partir de 5. La médiane est donc la classe 4

Truc pour les tout petit effectif : on classe toutes les occurrences 444567 en ordre croissant (ou décroissant) et on prend valeur (ou intervalle de valeur) qui sépare distribution en 2

Le cas d'une variable métrique :

Le principe est le même, mais on adapte la méthode : dans ce cas on peut calculer la **valeur médiane**

Ex : 300 personnes ont répondu à un test. Les notes obtenues se répartissent de la manière suivante :

Notes	0	1	2	3	4	5	6	7	8	9
Effectifs	1	6	5	11	75	98	80	15	5	4

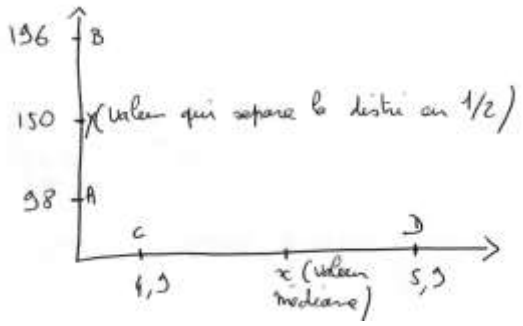
La médiane est la valeur de note qui à 150 pers au dessus et 150 en dessous

Première étape : Il faut savoir où l'on dépasse les 150 : On passe donc par le calcul des effectifs cumulés.

Notes	0	1	2	3	4	5	6	7	8	9
Effectifs	1	6	5	11	75	98	80	15	5	4
Effectifs cumulés	1	7	12	23	98	196				

A la fin de la classe 4 on n'est pas à 150. A la fin de la classe 5 on les a dépassés. Donc la classe médiane est « 5 ». Mais quelle est la « valeur médiane » ?

2ème étape : représentons graphiquement le problème :



Appliquons le vieux théorème de Thales :

$$\frac{CD}{AB} = \frac{Cx}{Ay}$$

Application numérique :

$$\frac{(5,5-4,5)}{(196-98)} = \frac{(x-4,5)}{(150-98)}$$

$$x = \left[\frac{(5,5-4,5)}{(196-98)} \times (150-98)\right] + 4,5$$

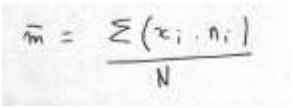
x = 5,4

La valeur médiane est de 5,4

Ah, j'oubliais un détail : Bien sûr, Excel fait parfaitement bien tout seul ces calculs laborieux !

Moyenne : Il s'agit de la somme des observations divisées par leurs effectifs (Elle se calcul uniquement sur des variables métriques)

Soient :
xi : Les valeurs correspondants aux modalités de la variable
ni : Les effectifs concernés par ces valeurs
N= somme des effectifs concernés par les valeurs

Formule : 

Ex : une personne a subi 5 épreuves qui ont été notées

Epreuve 1	20
Epreuve 2	15
Epreuve 3	18
Epreuve 4	25
Epreuve 5	17

Total : 95

Moyenne = 95*1/5 = 19

L'exercice semble facile lorsqu'une seule personne est en jeu. Le même principe s'applique lorsque plusieurs centaines d'individus sont concernés

Autre Ex : Le nombre d'enfant par famille (calculé à partir de 100 familles)

xi	ni	xi.ni
0	42	0
1	18	18
2	33	66
3	5	15
4 et plus	2	10
total	100	109

Moyenne = 109/100. Il y a donc 1,09 enfant par famille

Ce qu'il faut retenir du chapitre sur les valeurs centrales

Type de variable et indicateurs

- Variable nominale : mode
- Variable ordinale : mode, médiane
- Variable numérique : mode médiane, moyenne

Commentaire technique :

La moyenne est valeur que prendraient toutes les unités statistiques si elles étaient identiques. La moyenne est intéressante en ce qu'elle permet de reconstituer la somme des

valeurs et donc de passer d'une population à une autre.

Attention : la moyenne est très sensible aux valeurs extrêmes (éléments atypiques ou erreur de saisie).

La médiane est fondée sur l'ordre. Si on compare un tableau statistique à un cortège (d'informations) classé par ordre de grandeur, la moyenne correspondrait à un élément de synthèse qui concentre les caractères de tous les autres éléments qui défilent. La médiane indique l'élément (ou le groupe d'éléments) bien réel qui est au milieu du cortège. Ici les valeurs prises par la variable concernée n'interviennent que pour établir un classement : la médiane est donc peut sensible aux valeurs extrêmes (On parle d'indicateur robuste).

Nota : le calcul de moyenne est impossible sur les variables qualitatives (le prénom des enfants, la PCS….). En effet que signifierai une moyenne de « ouvrier virgule cinq » ?
Mais cela est possible sur une variable numérique relative à des données qualitatives

Ex : Dans *La distinction*, P. Bourdieu interroge 584 individus issus des classes moyennes dont 100 artisans, 287 employés, 78 techniciens et instituteurs et 119 « petite bourgeoisie nouvelle ». Serge Guétary est-il leur chanteur préféré ?

Il obtient un tableau avec une variable ordinale « catégories sociales »

artisans	22%
employés	21%
techniciens	4%
« petite bourgeoisie »	9%

Transformons ce tableau en variable numérique de type « score obtenu par Guétary »

xi	ni	xi.ni
22	100	2200
21	287	6027
4	78	312
9	119	1071

La moyenne est donc de 1071/584 soit 16,45 (arrondie par P. Bourdieu à 17 car il élimine les non réponses)

On peut en déduire des « écarts à la moyenne » Ceci est très pratique pour visualiser (surtout sur des tableaux importants où plusieurs variables sont empilées) des différences significatives). On peut les faire apparaître en reprenant la présentation d'origine

Ex :

artisans	22%-17%	+5%
employés	21%-17%	+4%
Techniciens, instit	4%-17%	- 13%
« petite bourgeoisie »	9%-17%	-8%

Un indicateur simple de valeur centrale permet donc facilement de commencer à faire parler un tableau statistique.

2 - Résumé statistique : les indicateurs de dispersion

Les mesures de tendances centrales permettent d'identifier la valeur la plus représentative à l'intérieur d'un ensemble de données.

Moyenne, médiane et mode donnent des perspectives différentes du centre d'un ensemble de données. Mais la description n'est pas complète aussi longtemps qu'on ne connaît pas également la variabilité de la distribution.

La description numérique de base d'un ensemble de données exige des mesures, tant du centre que de la dispersion.

Pour les variables numériques : variance et écart type

L'écart-type : C'est la mesure de dispersion la plus couramment utilisée en statistique lorsqu'on emploie la moyenne pour calculer une tendance centrale.

Il mesure la dispersion autour de la moyenne. En raison de ses liens étroits avec la moyenne, l'écart-type peut être grandement influencé si cette dernière donne une mauvaise mesure de tendance centrale.

L'écart-type est aussi influencé par les « valeurs aberrantes » (une seule de ces valeurs pourrait avoir une grande influence sur les résultats de l'écart-type)

Il s'agit donc d'un bon indicateur de l'existence de valeurs aberrantes, ce qui en fait une mesure de dispersion très utile pour les distributions symétriques ne comptant aucune valeur aberrante.

L'écart-type est aussi utile quand on compare la dispersion de deux ensembles de données séparés qui ont approximativement la même moyenne.

De manière générale, plus l'écart-type est petit, plus les valeurs sont concentrées autour de la moyenne (moins il y a de valeurs élevées ou de valeurs faibles).

Un élément sélectionné au hasard à partir d'un ensemble de données dont l'écart-type est faible peut se rapprocher davantage de la moyenne qu'un élément d'un ensemble de données dont l'écart-type est plus élevé.

Pour calculer l'écart-type, il faut passer par un autre calcul d'un autre indicateur de dispersion : *la variance*. Etape obligée, la variance est toutefois moins utilisée parce que moins « parlante ».

La variance : C'est la moyenne des carrés des écarts à la moyenne

Pour expliquer ce point, nous distinguerons 2 cas : les distributions simples et les tableaux de fréquences.

Le cas des distributions simple, c'est-à-dire plusieurs événements et une seule unité d'observations

Ex : un individu qui subi plusieurs épreuves

On définit la variance d'une variable discrète composée de **n** observations comme suit :

$$V = \frac{\sum (x - \overline{x})^2}{n}$$

L'écart-type d'une variable discrète composée de **n** observations est la racine carrée positive des variances et se définit comme suit :

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

Utilisez cette méthode étape par étape pour trouver l'écart-type d'une variable discrète :

1. Calculez la moyenne.
2. Soustrayez de chaque observation la moyenne.
3. Calculez le carré de chacune des autres observations.
4. Additionnez ces résultats au carré.
5. Divisez ce total par le nombre d'observations (la variance, **V**).
6. Utilisez la racine carrée positive (écart-type, **σ**).

Autre exemple : 2 étudiants comparent leurs notes :

A a eu 9, 10, 11,12 et 8
B a eu 2, 15, 18,10 et 5

Ce qui nous amène à faire tableau statistique suivant :

Pour A

xi	**ni**	**xi .ni**	**$(xi-m)^2$**
9	1	9	1
10	1	10	0
11	1	11	1
12	1	12	4
8	1	8	4
Totaux :	**5**	**50**	**10**

La moyenne est 50/5 soit : 10

La variance est 10/5 soit : 2

L'écart-type est de 1.41 pt autour de la moyenne

Pour B

xi	**ni**	**xi.ni**	**$(xi.m)^2$**
2	1	2	64
15	1	15	25
18	1	18	64
10	1	10	0
5	1	5	25
Totaux	**5**	**50**	**178**

La moyenne est 50/5 soit : 10

La variance est 35,6

L'écart-type est de 5.96 pt autour de la moyenne

Exemple 3 : L'Écart-type calculé à l'aide d'un tableau de fréquences

$$\bar{m} = \frac{\sum x_i \cdot n_i}{N}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{m})^2 \cdot n_i}{N}}$$

<u>Le cas des tableaux de fréquence CAD lorsqu'il y a à la fois plusieurs événements et plusieurs unités d'observation</u>

Ex : plusieurs individus subissent individuellement plusieurs épreuves

On a demandé à 30 artisans combien d'ouvriers ils embauchent actuellement. Voici leurs réponses :

4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4

On en tire le tableau suivant :						
Travailleur (xi)	**Comptage**	**Fréquence ni** (ou f notamment dans les statistiques anglo saxonnes)	**(xi.ni)**	**(x - $\bar{x}$)**	**(x - $\bar{x}$)^2**	**(x - $\bar{x}$)^{2}ni**
0	I	1	0	-5	25	25
1	I	1	1	-4	16	16
2	II	2	4	-3	9	18
3	III	3	9	-2	4	12
4	𝍸 I	6	24	-1	1	6
5	𝍸	5	25	0	0	0
6	IIII	4	24	1	1	4
7	III	3	21	2	4	12
8	III	3	24	3	9	27
9	II	2	18	4	16	32
		30	**150**			**152**

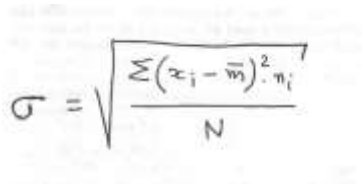
Pour calculer la moyenne :

$$\bar{m} = \frac{\sum x_i \cdot n_i}{N}$$

$$= \frac{150}{30}$$

$$= 5$$

Pour calculer l'écart-type :



$\sigma = \sqrt{(152/30)}$

$\underline{\sigma = 2.25}$

Les quartiles

Ils sont utilisables pour les variables numériques mais aussi ordinales car elles n'ont pas de moyenne donc pas question d'écart-type or les quartile reposent donc sur calcul médiane

La médiane divise les données en deux ensembles égaux.

Le quartile inférieur est la valeur du milieu du premier ensemble, dans lequel 25 % des valeurs sont inférieures à Q_1 et 75 % lui sont supérieures. Le premier quartile prend la notation Q_1.

Le quartile supérieur est la valeur du milieu du deuxième ensemble, dans lequel 75 % des valeurs sont inférieures à Q_3 et 25 % lui sont inférieures. Le troisième quartile prend donc la notation Q_3.

Il convient de noter que la médiane prend la notation Q_2, c'est-à-dire le deuxième quartile.

Exemple 1 : Quartiles supérieur et inférieur

11 individus se sont positionnés sur échelle d'attitude de 50 degrés

Données	6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36
Données ordonnées	6, 7, 15, 36, 39, 41, 41, 43, 43, 47, 49
Médiane	41
Quartile supérieur	43
Quartile inférieur	15

Écart interquartile

L'écart interquartile est une autre étendue utilisée comme mesure de la dispersion.

La différence entre les quartiles supérieur et inférieur (Q_3 - Q_1), qu'on appelle *l'écart interquartile*, indique aussi la dispersion d'un ensemble de données.

L'écart interquartile couvre 50 % d'un ensemble de données et élimine l'influence des valeurs aberrantes, parce qu'on soustrait, en effet, le quartile le plus élevé et le quartile le plus faible.

L'écart interquartile est la différence entre le quartile supérieur (Q_3) et le quartile inférieur (Q_1)

Dans l'exemple qui précède : 43-15 = 28

L'écart interquartile est de 28

3 - Le résumé en cinq nombres d'une distribution

Un résumé en cinq nombres est surtout utile pour effectuer des analyses descriptives ou pour faire une analyse préliminaire d'un vaste ensemble de données.

Un résumé se compose de cinq valeurs :

- les valeurs les plus extrêmes incluses dans l'ensemble de données (les valeurs maximale et minimale),
- les quartiles le moins et le plus élevés
- la médiane.

Ces valeurs sont présentées ensemble et ordonnées de la plus faible à la plus élevée : la valeur minimale, le quartile inférieur (Q_1), la valeur médiane (Q_2), le quartile supérieur (Q_3) et la valeur maximale.

On a sélectionné ces valeurs pour fournir un résumé d'un ensemble de données, parce que chaque valeur décrit une partie précise d'un tel ensemble : la médiane identifie le centre d'un ensemble de données; les quartiles supérieur et inférieur couvrent la moitié intermédiaire d'un ensemble de données; l'observation la plus faible et l'observation la plus élevée fournissent de l'information supplémentaire sur la dispersion réelle des données, ce qui fait que le résumé en cinq nombres est une mesure de dispersion importante.

Ex : Un vendeur travaille dans une boutique d'informatique. Il a enregistré le nombre d'ordinateurs qu'il a vendus chaque mois. Ces 12 derniers mois, il a vendu mensuellement le nombre d'ordinateurs suivants : 51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13. Donnez un résumé en cinq nombres du nombre d'ordinateurs vendus.

Réponses

a) Premièrement, classez les données dans l'ordre ascendant. Trouvez ensuite la médiane.

6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.

Médiane = (12 + 1) ÷ 2 = 6,5^{e} valeur
= (6^{e} + 7^{e} observations) ÷ 2
= (25 + 39) ÷ 2
= **32**

Il y a six nombres au-dessous de la médiane : 6, 7, 13, 17, 20, 25.

Q_1 = la médiane de ces six éléments
= (6 + 1) ÷ 2= 3,5^{e} valeur
= (3^{e} + 4^{e} observations) ÷ 2
= (13 + 17) ÷ 2
= **15**

Il y a six nombres au-dessus de la médiane : 39, 41, 43, 49, 51, 62.

Q_3 = la médiane de ces six éléments
= (6 + 1) ÷ 2= 3,5^{e} valeur

= (3^e + 4^e observations) ÷ 2
= **46**

Le résumé en cinq nombres pour le nombre d'ordinateurs vendus est 6, 15, 32, 46, 62.

II - La corrélation des variables

Il est très fréquent en sociologie de travailler sur des *sondages* et de rechercher des *corrélations* entre variables

Deux questions ne manquent alors pas de se poser : *Est-ce que la liaison entre les 2 variables est avérée (ou est-ce juste le fait du hasard)* ? *Est-ce qu'il est possible de généraliser la liaison sur la population parente* ?

Pour ça, on recourt à des *tests de corrélation* c'est-à-dire des tests appliqués à une distribution croisant deux variables pour conclure ou non à l'existence d'une corrélation.

Les tests sont différents selon les types de variables (nominales, ordinales, numériques).

Je vous propose d'en étudier un parmi les plus courants qui permet d'explorer la corrélation entre deux variables numériques : le X^2 (Khi deux)

1 – Corrélation de variables numériques : le X2

Concrètement, il s'agit de répondre à la question : *« Est-ce que la position d'un individu par rapport à une des deux variables permet de faire un pronostic quant à position par rapport à l'autre ? »*

Le X^2 est un test d'hypothèse c'est-à-dire qu'il repose sur la méthode *hypothético-déductive*

Prenons un tableau qui croise 2 variables :

	Modalité 1 de la Variable 1	Modalité 2 de la Variable 1	Modalité 3 de la Variable 1
Modalité 1 de la Variable 2	*Effectif correspondant*	*Effectif correspondant*	*Effectif correspondant*
Modalité 2 de la Variable 2	*Effectif correspondant*	*Effectif correspondant*	*Effectif correspondant*
Modalité 3 de la Variable 2	*Effectif correspondant*	*Effectif correspondant*	*Effectif correspondant*

Concrètement 3 étapes sont essentielles :

1 : On formule l'hypothèse que les variables sont indépendantes (hypothèse nulle H0)

2 : On fait le tableau statistique théorique qui confirmerait cette hypothèse (calcul des effectifs théoriques)

3 On mesure l'écart entre le tableau réel (avec les effectifs réels) et le tableau théorique (avec les effectifs théoriques). L'indice qui rend compte de cet écart est le X^2

Partons d'un exemple : Une enquête d'opinion auprès d'un échantillon de 373 mères divorcées remises en couple

Question : *Comment considérez vous le rôle de votre conjoint actuel par rapport à vos enfants ?*

Résultats :

	Vit en couple remarié	Vit en couple non remarié	
« un vrai père »	60	55	**115**
« un second père »	60	56	**116**
« un ami, un conseiller »	55	87	**142**
	175	**198**	**373**

Calcul du X^2

Ho : Il n'y a pas de différence d'opinion selon le statut matrimonial. Il y a donc la même proportion de femmes quelques soit leur statut matrimonial

Effectifs théoriques correspondants :

- On commence par déterminer le *degré de liberté*, c'est-à-dire le nombre minimum d'effectifs théoriques à calculer (les autres pouvant être obtenus par déduction)
  Soient : dl = degré liberté
  l =nombre de lignes
  c =nombre de colonnes
  dl = (l-1). (c-1)

  Ici dl=2

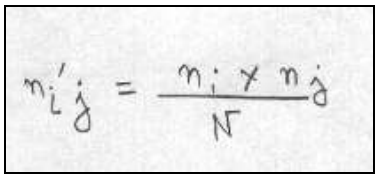
1	*Peut être déduit*
2	*Peut être déduit*
Peut être déduit	*Peut être déduit*

- on calcule les effectifs théoriques proprement dit :

Soit le tableau réel :

nij 1		ni
nij 2		ni
		ni
ji	ji	N

Donc le tableau théorique :



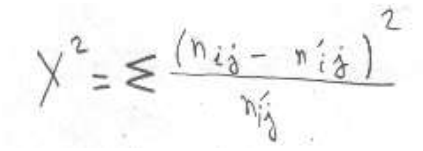
n'ij 1		ni
n'ij 2		ni
		ni

ji	ji	N

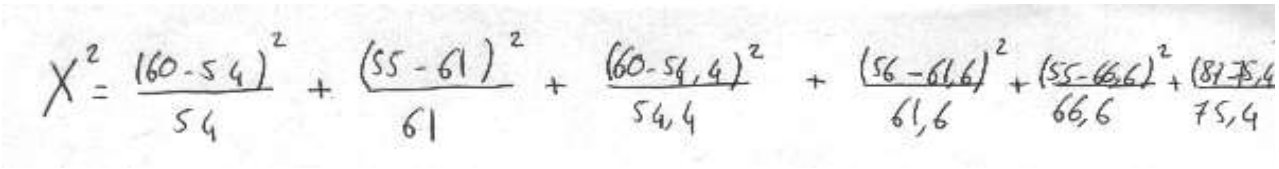
Application numérique :

54 (au lieu de 60)	61 (au lieu de 55)	115
54 (au lieu de 60)	61.6 (au lieu de 56)	116
66.6 (au lieu de 55)	75.4 (au lieu de 87)	142
175	198	N

On applique la formule générale du X^2



Il reste donc à appliquer la formule pour chaque case :



$X^2 = 0.7+0.6+0.6+0.5+2+1.8$

$X^2 = 6.2$

On regarde sur la *table des X^2* le « X^2 max » admis pour confirmer Ho et conclure à l'indépendance des variables. Comment lire cette table ?

La lecture dépend de 2 éléments : Le dl (degré de liberté) et le risque admis (C'est le même principe que pour l'intervalle de confiance)

Ici dl = 2. Prenons un risque par exemple de 0.05 (C'est à dire avec 95% de chance de ne pas se tromper)

Donc ici on lit sur la table X^2 max = 5.99. Notre X^2 observé est de 6.2. On doit donc rejeter l'hypothèse d'indépendance Ho. **Il y a donc corrélation entre les variables**

Attention :
- Un X2 fort ne signifie pas une liaison forte mais une forte probabilité de liaison
- Il existe des limites de fiabilité et des précautions d'usage du X2 (Pour approfondir la question, je vous suggère le recours à un manuel de statistiques descriptives)