

Analyse sociologique de données d'enquête par questionnaire

Olivier Martin

Cours de l'UE « Méthodes d'investigation sociologique »

L3 de sciences sociales – Université Paris Descartes – 2011-2012

Le test du χ^2 ?

En résumé...

- 1) On fait l'hypothèse que les deux variables sont indépendantes.
- 2) Sous cette hypothèse, pour chacun des écarts possibles à l'indépendance, on détermine si chacun de ces écarts est très probable, assez probable, moyennement probable, peu probable, pas du tout probable...
- 3) On en déduit la valeur de la probabilité pour la distance entre notre tableau empirique et le tableau d'indépendance...
- 5) On en conclut la valeur de plausibilité (crédibilité) de l'hypothèse initiale (indépendance).

En pratique...

1) On fait l'hypothèse que les deux variables sont indépendantes.

2) ORDINATEUR – STATISTICIEN – LOGICIEL
probable, pas du tout probable...

3) ORDINATEUR – STATISTICIEN – LOGICIEL

5) On en conclut la valeur de plausibilité (crédibilité) de l'hypothèse initiale (indépendance).

En pratique...

- 1) On réalise le tri croisé souhaité.
- 2) On calcule... ou plutôt on fait calculer la distance du khi² et surtout la probabilité associée.
- 3) Si la probabilité est faible (< 5 % ou < 10 %) on en déduit que les deux variables ne sont probablement pas indépendantes... Elles sont probablement dépendantes.



Revenons à l'exemple utilisé :

| Tableau empirique | | | | | |
|-------------------|------------------------------|----------------------------|------------------------------|--|-------|
| | tu aimes beaucoup les livres | tu aimes plutôt les livres | tu es indifférent aux livres | tu n'aimes pas les livres mais tu dois t'en servir malgré tout | Total |
| femme | 504 | 312 | 37 | 11 | 864 |
| homme | 186 | 146 | 30 | 7 | 369 |
| Total | 690 | 458 | 67 | 18 | 1233 |

$$\text{Distance} = 0,8689 + 0,2487 + 2,1083 + 0,2063 + \dots + 2,0344 + 0,5824 + 4,9364 + 0,4831 = 11,4684$$

Probabilité associée à cette distance = $0,009444708\dots \approx 0,01 = 1 \%$

Tableau peu probable... donc rejet de l'hypothèse d'indépendance... donc « dépendance » des variables.

Exemple (sous Modalisa)

The screenshot shows the SPSS interface with a contingency table and chi-square test results. The table is titled "45. Q37 sexe / 43. Q35 Degré 'd'attachement' au livre". The columns represent levels of attachment to books, and the rows represent gender. The chi-square test results are displayed at the bottom of the window.

| | tu aimes beaucoup les livres | tu aimes plutot les livres | tu es indifferent aux livres | tu n'aimes pas les livres mais tu dois t'en servir malgré to | Total |
|-------|------------------------------|----------------------------|------------------------------|--|-------|
| femme | 443 | 282 | 32 | 11 | 768 |
| homme | 173 | 131 | 29 | 7 | 340 |
| Total | 616 | 413 | 61 | 18 | 1108 |

Chi2=10,9 ddl=3 p=0,012 (Très significatif)

Distance

Probabilité

Probabilité faible ? Probabilité élevée ?

- La réponse dépend de la discipline...
- En sciences sociales / sociologie :
 - $p < 1\% = 0,01 \rightarrow$ Dépendance très probable
 - $p < 5\% = 0,05 \rightarrow$ Dépendance probable
 - $p < 10\% = 0,1 \rightarrow$ Dépendance assez probable
- On parle de significativité et on utilise des **** (étoiles).
 - $p < 1\% = 0,01 \rightarrow$ Très significatif = ***
 - $p < 5\% = 0,05 \rightarrow$ Significatif = **
 - $p < 10\% = 0,1 \rightarrow$ Assez significatif = *
- Attention, c'est une convention : elle peut varier d'une publication à l'autre, d'un auteur à l'autre, d'un échantillon à l'autre...

Comment utiliser le χ^2 ?

Le χ^2 : une solution miracle ?

La réponse est évidemment négative !

- A) Ne s'applique pas à tous les tableaux...
- B) Est un outil probabiliste...
- C) N'est pas la démonstration d'un lien sociologique entre deux aspects du social...
- D) Est un jugement global sur les relations entre deux variables
- E) Pas un indicateur d'intensité de la relation
- F) Sensible aux recodages !

Le Khi² : une solution miracle ?

A) Ne s'applique pas à tous les tableaux...

- seulement sur les tableaux de contingence (les individus doivent être représentés une et une seule fois)
- donc pas les croisements de variables multiples

Le Khi² : une solution miracle ?

A) Ne s'applique pas à tous les tableaux...

- seulement sur les tableaux « suffisamment » remplis

(pas trop de cases dont l'effectif théorique est inférieur à 5)

(vérifier les contributions au khi²)

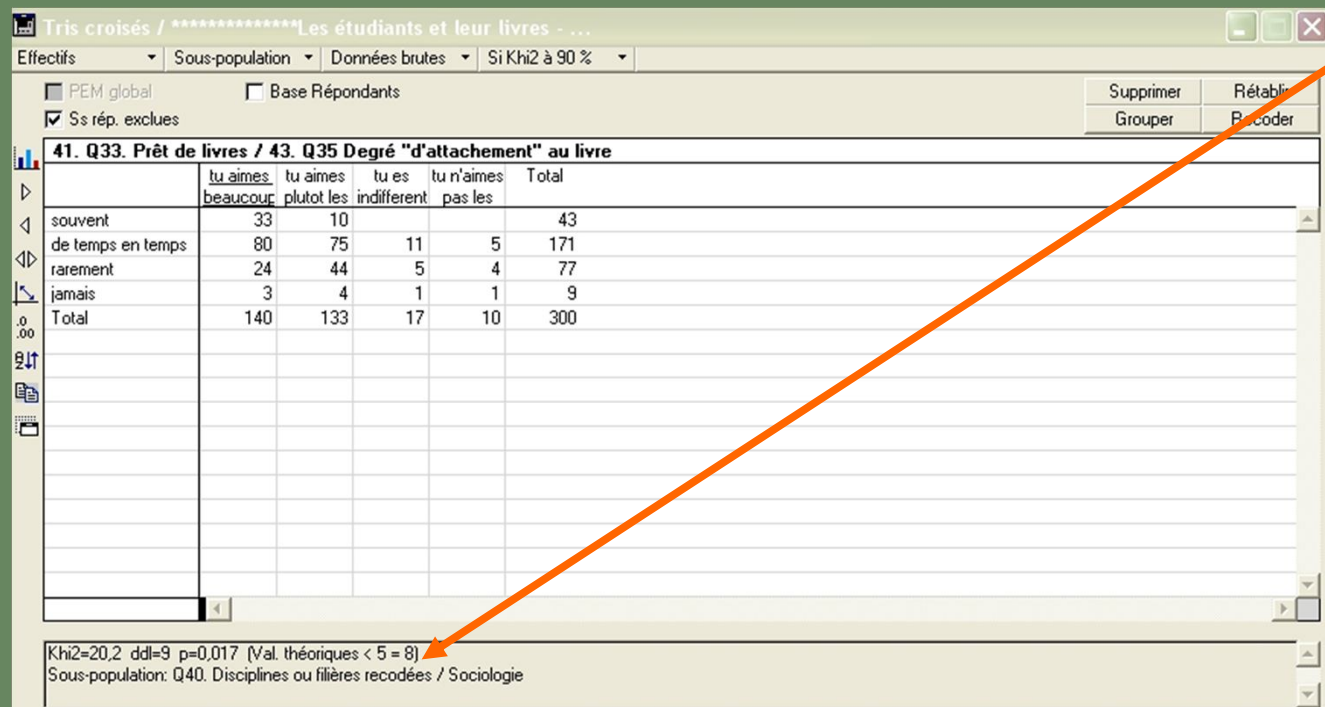
Tableau empirique

| | tu aimes beaucoup les livres | tu aimes plutôt les livres | tu es indifférent aux livres | tu n'aimes pas les livres mais tu dois t'en servir malgré tout | Total |
|-------|------------------------------|----------------------------|------------------------------|--|-------|
| femme | 10 | 5 | 1 | 3 | 19 |
| homme | 12 | 51 | 2 | 3 | 68 |
| Total | 22 | 56 | 3 | 6 | 87 |

Le Khi² : une solution miracle ?

A) Ne s'applique pas à tous les tableaux...

- être attentif aux commentaires fournis par MODALISA



The screenshot shows the SPSS Crosstabs dialog box for a 2x4 contingency table. The table displays the relationship between 'Prêt de livres' (rows) and 'Degré "d'attachement" au livre' (columns). The total sample size is 300. The warning message at the bottom indicates that the chi-square test is not applicable because the expected cell count is less than 5.

| | tu aimes beaucoup | tu aimes plutôt les | tu es indifférent | tu n'aimes pas les | Total |
|-------------------|-------------------|---------------------|-------------------|--------------------|-------|
| souvent | 33 | 10 | | | 43 |
| de temps en temps | 80 | 75 | 11 | 5 | 171 |
| rarement | 24 | 44 | 5 | 4 | 77 |
| jamais | 3 | 4 | 1 | 1 | 9 |
| Total | 140 | 133 | 17 | 10 | 300 |

Chi2=20,2 ddl=9 p=0,017 (Val. théoriques < 5 = 8)
Sous-population: Q40. Disciplines ou filières recodées / Sociologie

Le χ^2 : une solution miracle ?

B) Est un outil probabiliste...

- pas à l'abri d'un « mauvais tour » du hasard
- les biais dans l'échantillonnage
- échantillonnage jamais parfaitement aléatoire

Le Khi^2 : une solution miracle ?

C) N'est pas la démonstration d'un lien sociologique entre deux aspects du social...

- un lien statistique n'est pas synonyme d'un lien sociologique
- le sens de ce lien n'est pas indiqué par le test du khi^2
- la signification de ce lien n'est pas indiqué par le test du khi^2

Le χ^2 : une solution miracle ?

D) Est un jugement global sur les relations entre deux variables

- Deux variables peuvent être globalement indépendantes même si deux modalités sont fortement associées
- Un lien de dépendance n'indique pas quelles sont les modalités fortement associées (... sauf si on étudie les contributions du χ^2)

Le χ^2 : une solution miracle ?

E) Pas un indicateur d'intensité de la relation

- Le χ^2 indique seulement la probabilité d'indépendance des deux variables et non la force de leur lien éventuel...

Le Khi² : une solution miracle ?

F) Sensible aux recodages !

- Oui, puisqu'un recodage est en fait une redéfinition de la variable, des catégories qui s'opposent...
- Il ne faut pas hésiter à recoder (regrouper lignes ou colonnes) pour faire « parler » un tableau...

Le χ^2 : une solution miracle ?

Un dernier conseil :

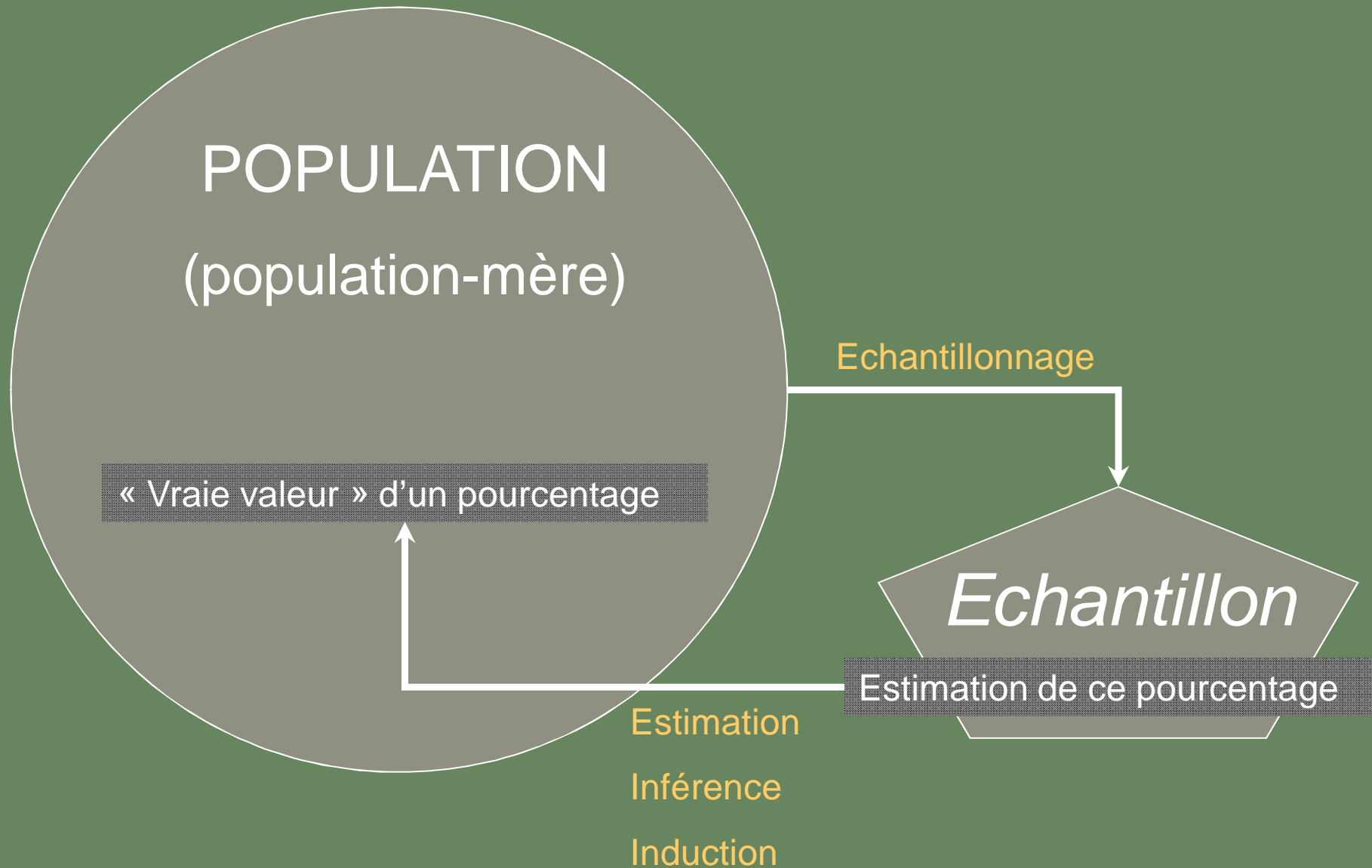
- Comme tout test ou outil statistique, le χ^2 fournit des indications, des indices...
- Il n'est ni suffisant ni nécessaire
- Il permet de disposer d'arguments supplémentaires lors d'une analyse
- L'absence de significativité du test ne signifie pas l'absence d'une interdépendance entre deux modalités...
- L'indépendance entre deux variables peut être un résultat intéressant d'un point de vue sociologique !

Séance 7 :

Comment mesurer
la « fiabilité »
des pourcentages ?

La situation courante est la suivante :

- Un population dont on souhaite connaître certaines propriétés
- Un échantillon issu de cette population et permettant de calculer des pourcentages
- Une question : puis-je faire confiance à ces pourcentages calculés sur l'échantillon pour me renseigner sur la population ?



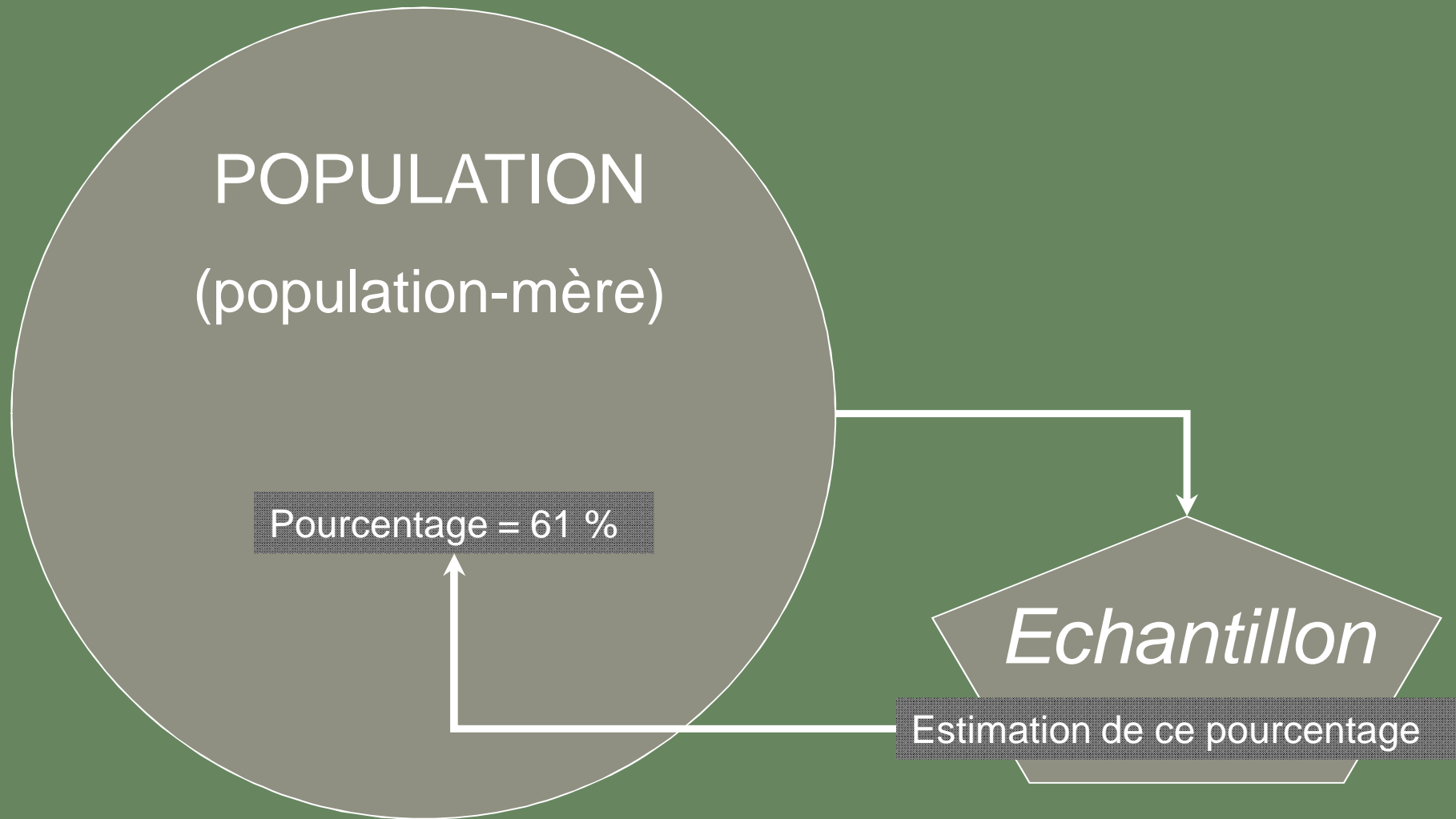
Comme pour le χ^2 , la science statistique et probabiliste nous permet de trouver une réponse à notre question...

Mais l'intuition et l'expérience empirique nous permet de « pressentir » la réponse possible à notre question

Démarches « empiriques »

Imaginons la situation suivante :

- Population de très grande taille
- On suppose que 61 % de cette population possède la propriété Y (avoir fait telle chose, avoir telle caractéristique...) : 61 % de « Oui » ; 39 % de « Non »
- Quelle estimation de ce pourcentage de « Oui » va nous fournir notre échantillon ?



Question

L'estimation sera-t-elle proche de la « vraie valeur » ?

Intuitivement

Oui...

Cf. Tirage pile ou face

Cf. Sondages d'opinion

« Loi des grands nombres »

Premier cas : un échantillon de taille 10

Un échantillon de taille 10

| | | | | | | | | | | | | | |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° 1 | : | Oui | Non | Oui | Oui | Non | Oui | Non | Non | Non | Oui | Nombre de "Oui" = | 5 |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

Un autre échantillon de taille 10

| | | | | | | | | | | | | | |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° 2 | : | Oui | Non | Oui | Oui | Oui | Non | Non | Non | Oui | Non | Nombre de "Oui" = | 5 |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

Encore un échantillon de taille 10

| | | | | | | | | | | | | | |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° 3 | : | Oui | Non | Oui | Oui | Oui | Non | Non | Oui | Non | Non | Nombre de "Oui" = | 5 |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

Encore un échantillon de taille 10

| | | | | | | | | | | | | | |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° 4 | : | Non | Non | Oui | Non | Non | Oui | Oui | Non | Non | Oui | Nombre de "Oui" = | 4 |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

Encore un échantillon de taille 10

| | | | | | | | | | | | | | |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° 6 | : | Non | Non | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Nombre de "Oui" = | 8 |
|-------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

Premier cas : un échantillon de taille 10

... encore un échantillon de taille 10

| | | | | | | | | | | | | | | |
|-----------|----|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|
| Tirage n° | 10 | | Non | Oui | Oui | Non | Oui | Oui | Oui | Oui | Non | Oui | Nombre de "Oui" = | 7 |
|-----------|----|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|---|

... encore un échantillon de taille 10

| | | | | | | | | | | | | | | |
|-----------|----|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|----|
| Tirage n° | 13 | | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Oui | Nombre de "Oui" = | 10 |
|-----------|----|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|----|

Nous pourrions continuer très longtemps...

... par exemple 100 fois (*ie* 100 échantillons de taille 10)

Quelle serait la répartition de nos échantillons,
si on les classe selon le nombre de « Oui » et de « Non » ?

Premier cas : un échantillon de taille 10

Répartition de nos échantillons, selon le nombre de « Oui »

| Nombre de "Oui" dans l'échantillon | Nombre d'échantillon obtenus parmi 100 |
|------------------------------------|--|
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |
| 3 | 0 |
| 4 | 10 |
| 5 | 16 |
| 6 | 33 |
| 7 | 18 |
| 8 | 16 |
| 9 | 4 |
| 10 | 2 |

Premier cas : un échantillon de taille 10

Un autre ensemble de 100 échantillons
donnera des résultats différents :

| Nombre de "Oui" dans l'échantillon | Nombre d'échantillon obtenus parmi 100 |
|------------------------------------|--|
| 0 | 1 |
| 1 | 0 |
| 2 | 0 |
| 3 | 6 |
| 4 | 8 |
| 5 | 11 |
| 6 | 17 |
| 7 | 30 |
| 8 | 21 |
| 9 | 4 |
| 10 | 2 |

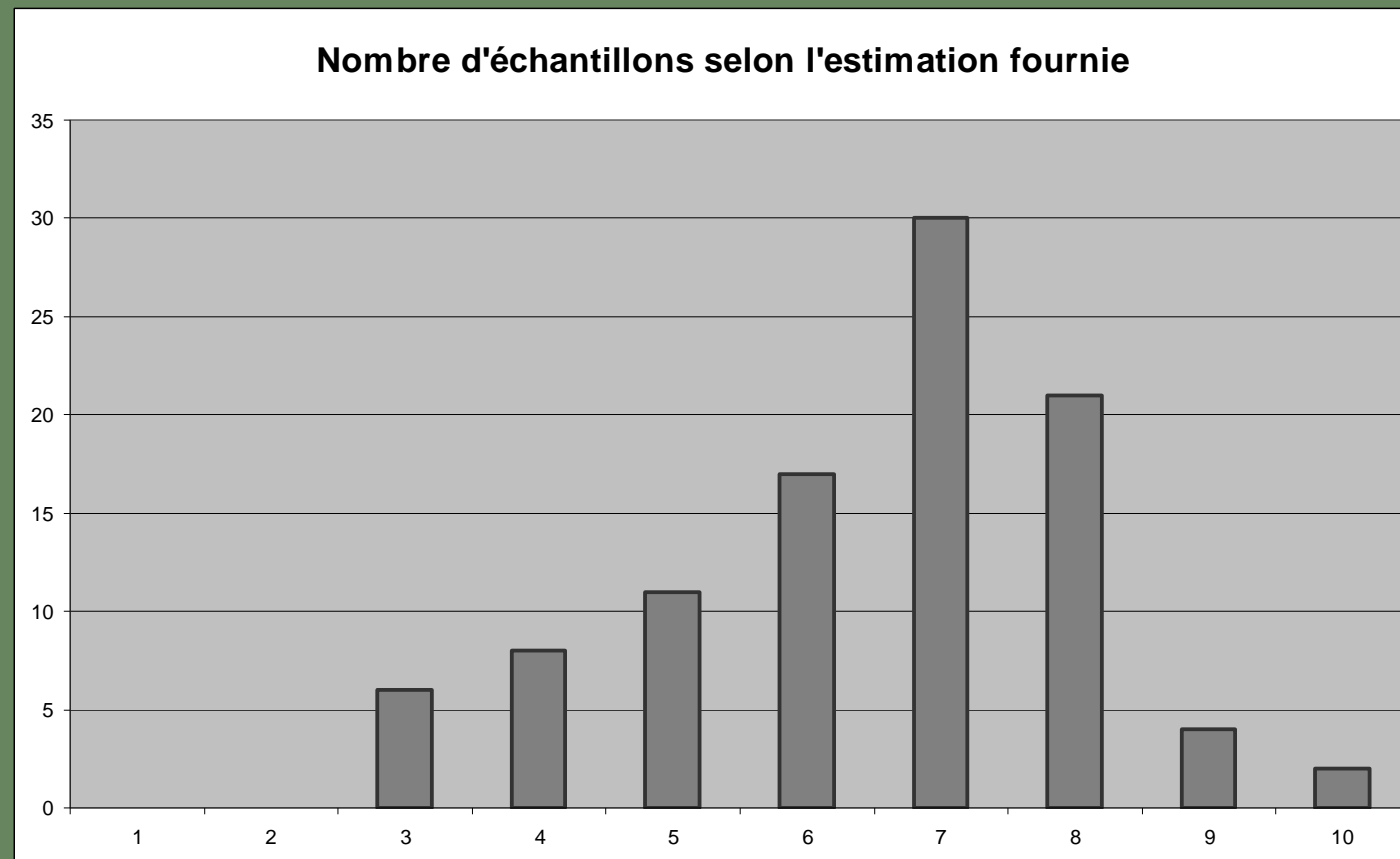
Premier cas : un échantillon de taille 10

Chacun de ces types d'échantillon donnera une estimation de la part de « Oui » dans la population

| Nombre de "Oui" dans l'échantillon | Estimation | Nombre d'échantillon obtenus parmi 100 |
|------------------------------------|------------|--|
| 0 | 0% | 1 |
| 1 | 10% | 0 |
| 2 | 20% | 0 |
| 3 | 30% | 6 |
| 4 | 40% | 8 |
| 5 | 50% | 11 |
| 6 | 60% | 17 |
| 7 | 70% | 30 |
| 8 | 80% | 21 |
| 9 | 90% | 4 |
| 10 | 100% | 2 |

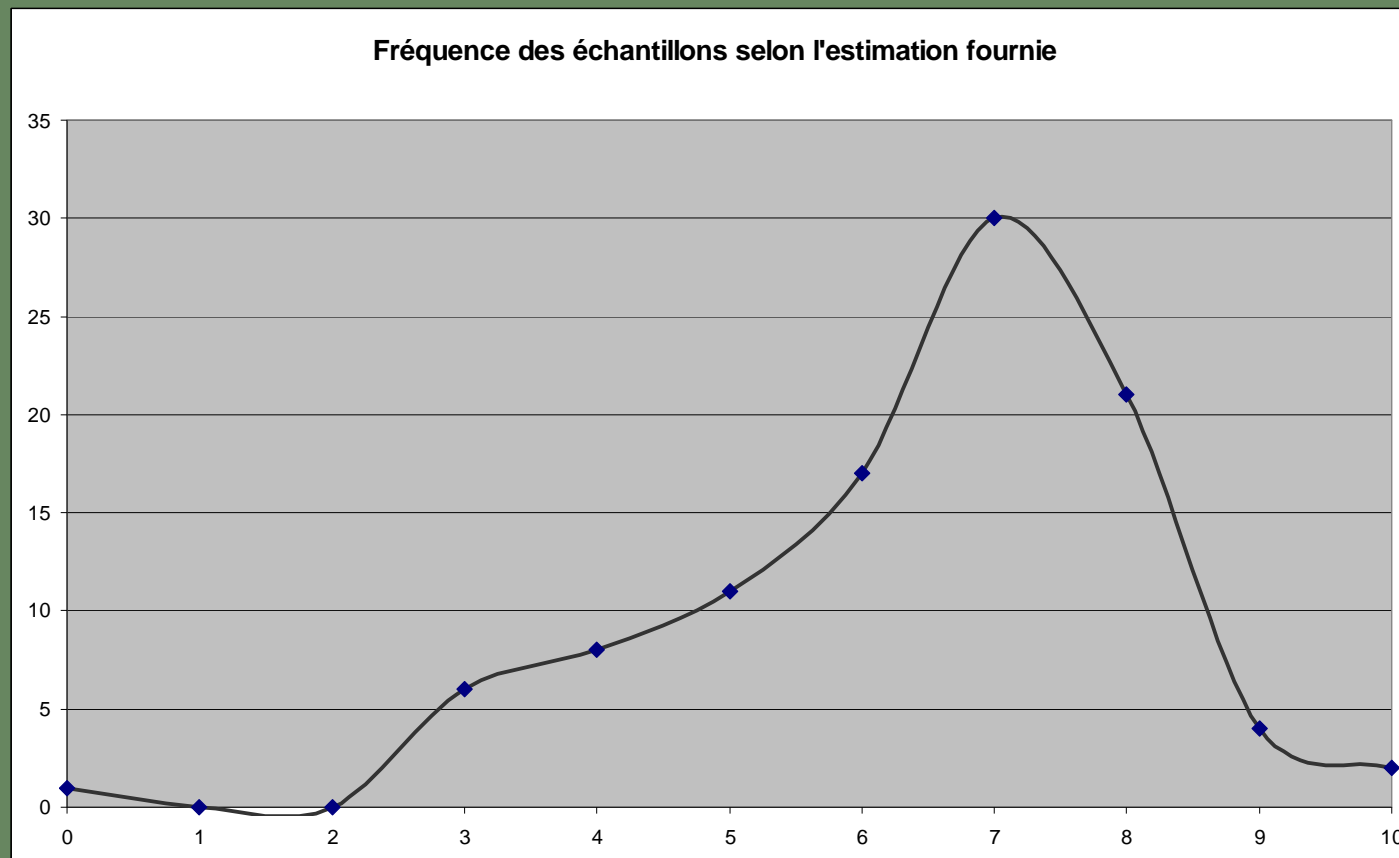
Premier cas : un échantillon de taille 10

Chacun de ces types d'échantillon donnera une estimation de la part de « Oui » dans la population



Premier cas : un échantillon de taille 10

Chacun de ces types d'échantillon donnera une estimation de la part de « Oui » dans la population



Premier cas : un échantillon de taille 10

On peut recommencer encore et encore...

→ Simulation sous EXCEL

Premier cas : un échantillon de taille 10

Que constate-t-on ?

- Rares sont les échantillons donnant une estimation « très fausse » (0 %, 10%, 20 %, ou 100 %) à la place de 61 %
- Assez fréquents sont les échantillons donnant une estimation « proches » (50%, 60 %, 70 %, 80 %) des 61 %

En d'autres termes :

→ *On a de « grandes chances » d'avoir un échantillon qui donne une estimation assez proche de la vraie valeur...*

Deuxième cas : un échantillon de taille 50

On suit le même raisonnement

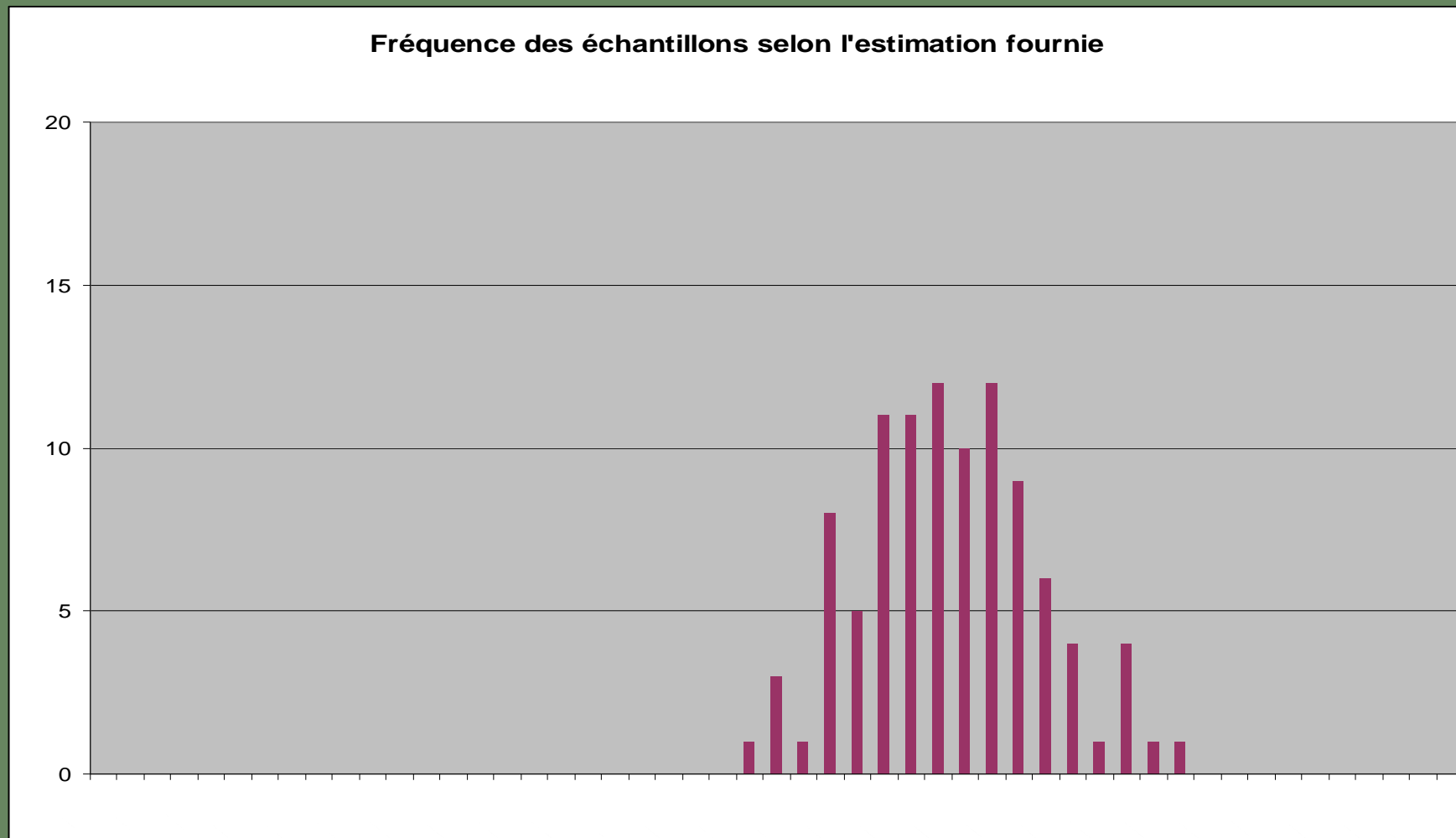
- Tirage d'un premier échantillon... dénombrement des « oui » parmi les 50
- Tirage d'un deuxième échantillon...
- Tirage d'un troisième échantillon...
- ...
- ...
- Tirage d'un 100^{ième} échantillon... dénombrement des « oui » parmi les 50

Deuxième cas : un échantillon de taille 50

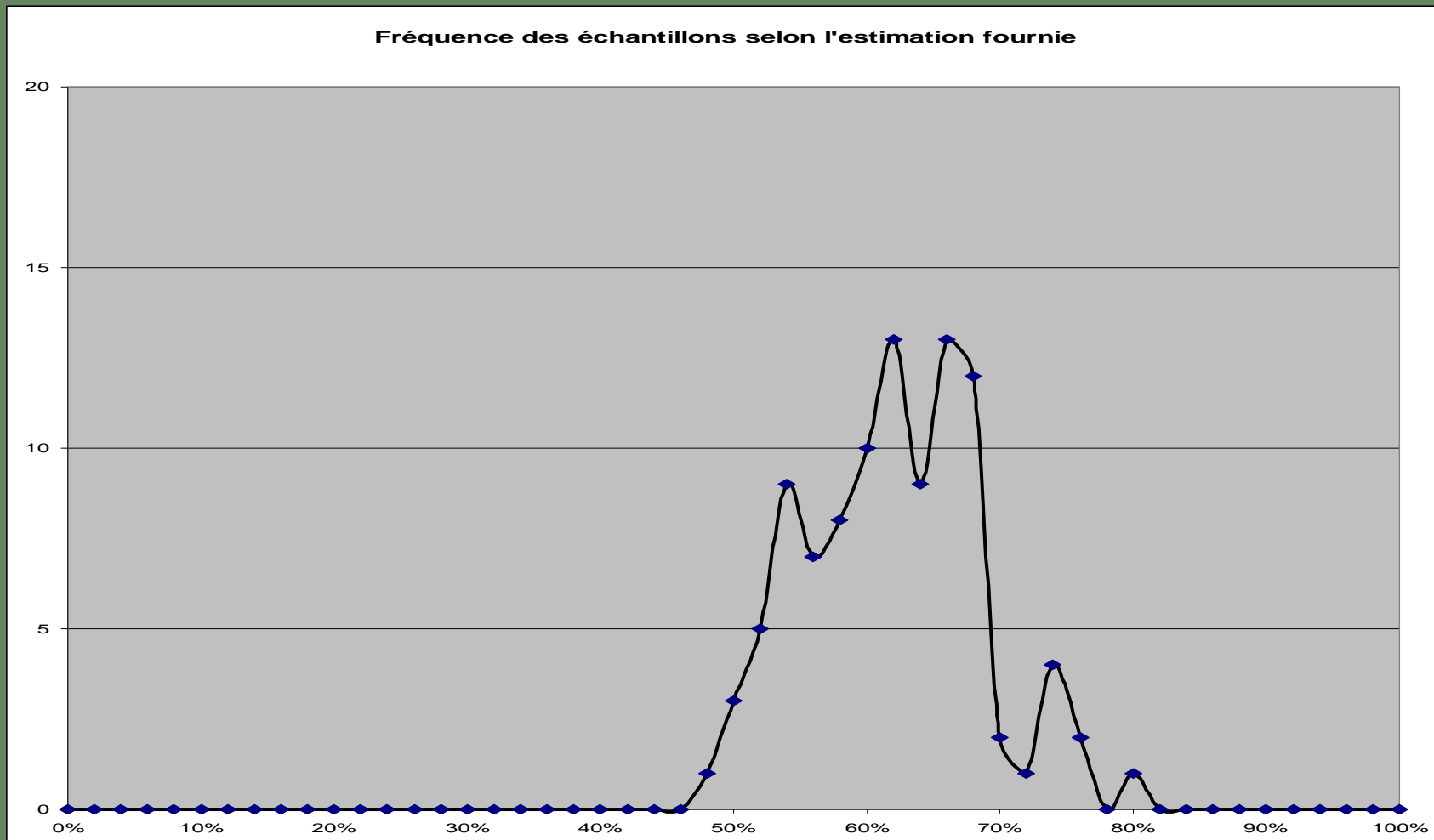
| Nombre de "Oui" dans l'échantillon | Estimation | Nombre d'échantillon obtenus parmi 100 |
|------------------------------------|------------|--|
| 0 | 0% | 0 |
| 1 | 2% | 0 |
| 2 | 4% | 0 |
| 3 | 6% | 0 |
| 4 | 8% | 0 |
| 5 | 10% | 0 |
| 6 | 12% | 0 |
| 7 | 14% | 0 |
| 8 | 16% | 0 |
| 9 | 18% | 0 |
| 10 | 20% | 0 |
| 11 | 22% | 0 |
| 12 | 24% | 0 |
| 13 | 26% | 0 |
| 14 | 28% | 0 |
| 15 | 30% | 0 |
| 16 | 32% | 0 |
| 17 | 34% | 0 |
| 18 | 36% | 0 |
| 19 | 38% | 0 |
| 20 | 40% | 0 |
| 21 | 42% | 0 |
| 22 | 44% | 0 |
| 23 | 46% | 0 |
| 24 | 48% | 1 |

| | | |
|----|------|----|
| 25 | 50% | 3 |
| 26 | 52% | 5 |
| 27 | 54% | 9 |
| 28 | 56% | 7 |
| 29 | 58% | 8 |
| 30 | 60% | 10 |
| 31 | 62% | 13 |
| 32 | 64% | 9 |
| 33 | 66% | 13 |
| 34 | 68% | 12 |
| 35 | 70% | 2 |
| 36 | 72% | 1 |
| 37 | 74% | 4 |
| 38 | 76% | 2 |
| 39 | 78% | 0 |
| 40 | 80% | 1 |
| 41 | 82% | 0 |
| 42 | 84% | 0 |
| 43 | 86% | 0 |
| 44 | 88% | 0 |
| 45 | 90% | 0 |
| 46 | 92% | 0 |
| 47 | 94% | 0 |
| 48 | 96% | 0 |
| 49 | 98% | 0 |
| 50 | 100% | 0 |

Deuxième cas : un échantillon de taille 50



Deuxième cas : un échantillon de taille 50



Deuxième cas : un échantillon de taille 50

On peut recommencer encore et encore....

→ Simulation sous EXCEL

Deuxième cas : un échantillon de taille 50 Que constate-t-on ?

- La même chose que pour les échantillons de taille 10 mais en « pire » (en plus prononcée)
- Encore plus rares sont les échantillons donnant une estimation « très fausse » (0 %, 10%, 20 %, ou 100 %) à la place de 61 %
- Encore plus fréquents sont les échantillons donnant une estimation « proches » (50%, 60 %, 70 %, 80 %) des 61 %

En d'autres termes :

→ *On a de « grandes chances »
d'avoir un échantillon qui donne
une estimation assez proche de la vraie valeur...*

... à suivre

INFORMATION

Mise en place d'un serveur et forum sur ce cours et des TD à l'adresse :

<http://moodle.univ-paris5.fr/>

Plus précisément : <http://moodle.univ-paris5.fr/course/category.php?id=175>

Identifiants nécessaires (intranet Paris Descartes)