Chapitre 13

Des statistiques inférentielles pour vous porter bonheur !

Hors Sujet

Titre : « Laurence Anyways »
Auteur : Xavier Dolan
Présentation succinte : Laurence Anyways est un film franco-québécois, un mélodrame, écrit et réalisé par Xavier Dolan, sorti en 2012.
Le film a reçu le prix de meilleur film canadien au festival international du film de Toronto et le grand prix au festival du film de Cabourg, ainsi que le Prix collégial du cinéma québécois en 2013. C'est l'histoire d'un impossible amour entre un homme et une femme après que celui-ci ait décidé de changer de sexe dans les années 90.
Le jour de ses trente ans, Laurence annonce à Fred qu'il veut devenir une femme et lui demande de l'accompagner dans sa transformation. Pour Fred, c'est un coup de tonnerre, mais elle décide malgré tout de donner une chance à leur couple.
Face aux jugements et à l'incompréhension, Laurence et Fred vont tout faire pour préserver leur amour hors du commun.

Xavier Dolan, né le 20 mars 1989 à Montréal,se fait connaître en tant que scénariste et réalisateur lors de la projection de son premier long métrage : « J'ai tué ma mère » à la 41e Quinzaine des réalisateurs, au cours de la 62e édition du Festival de Cannes. Il y gagne trois prix et les trois jurys soulignent le caractère unique de sa réalisation, la vérité, la violence et la poésie de la langue, ainsi que l'acharnement du jeune cinéaste et la foi en ses projets.

Document réalisé à l'aide de LaTeX
Auteur : C. Aupérin
Site : wicky-math.fr.nf
Lycée Jules Fil (Carcassonne)

Table des matières

L'ESSENTIEL :

- ⇝ Intervalles de fluctuation sur l'ensemble du lycée et leurs caractéristiques
- ⇝ Intervalles de confiance
- ⇝ INTERPRETATION ! C'est l'essentiel ...

« *Télécharger c'est tuer l'industrie, tuons les tous* »
THURSTON MOORE (SONIC YOUTH)

DES STATISTIQUES INFÉRENTIELLES POUR VOUS PORTER BONHEUR !

Des statistiques économiques jusqu'aux extraterrestres ...

Vous maitrisez assez bien les statistiques descriptives depuis le collège (moyenne, quartiles, etc).
Mais les statistiques inférentielles sont tout autre. Vous les avez effleurés du doigt en Seconde pour la première fois, vous les avez retrouvés en première très succinctement, mais le domaine est vaste.

C'est en 1746 que les statistiques endossent leur premier grand rôle dans la vie moderne : celui d'un outil de prévision et non de description.
Le mathématicien Deparcieux établit dans un essai le « profil » de la mortalité de populations à partir de données statistiques venant d'échantillons pris dans les registres et nécrologies. En se servant des méthodes d'échantillonnage, de calcul de moyennes et d'écarts-types, Deparcieux crée les premières « tables de mortalité », permettant d'évaluer le risque moyen de mort d'un individu en fonction de son profil (âge, sexe, profession, ...). Ce risque est alors directement transformé en pécule, par exemple dans le calcul du montant de rentes viagères (rente versée à quelqu'un durant toute sa vie en échange de l'acquisition de son bien à sa mort). Avec Deparcieux, la statistique fait son entrée dans l'économie.

Mais ce n'est qu'au XIX^e siècle que la statistique finit de prendre la place qu'est la sienne aujourd'hui : celle d'une science mathématique mais aussi humaine, omniprésente dans le débat publique. C'est Adolphe Quételet, astronome belge, qui intègre dans un ouvrage toutes les lois de probabilités développées depuis Pascal et Fermat : mesure des erreurs, méthode des moindres carrés, loi binomiale, etc.
Quételet croit à la possibilité d'établir une science mathématique humaine capable de relier les phénomènes sociaux de masse à des lois.
Ainsi, il tente d'établir des lois statistiques des suicides et des crimes en fonction de paramètres comme l'origine sociale, l'âge, le sexe, le climat, le niveau d'études, le revenu, etc. Mais Quételet se voit reprocher de faire de l'homme un être dont le comportement est prédéterminé par des lois mathématiques...

De plus, avec les statistiques sociales, une question se pose : sont-ce les statistiques qui déterminent nos comportements ou l'inverse ?

Par exemple, si le taux de meurtriers dans la population est de 5% par an, cela signifie-t-il qu'il existe une sorte de loi qui nous dépasse et qui « oblige » 5% de personnes à se transformer en meutriers ?
Ce type de questionnement hantera tout le XX^e siècle et conduira à de tragiques dérapages où l'on voudra « neutraliser » dès le berceau tout homme né avec les « paramètres statistiques du crime »...

Dans tous les cas, la statistique inférentielle, armée des lois de probabilité, envahit le XX^e siècle avec une force qui déborde la planète. Ainsi, depuis les années 1960, les astronomes se servent des statistiques pour pister les extraterrestres !

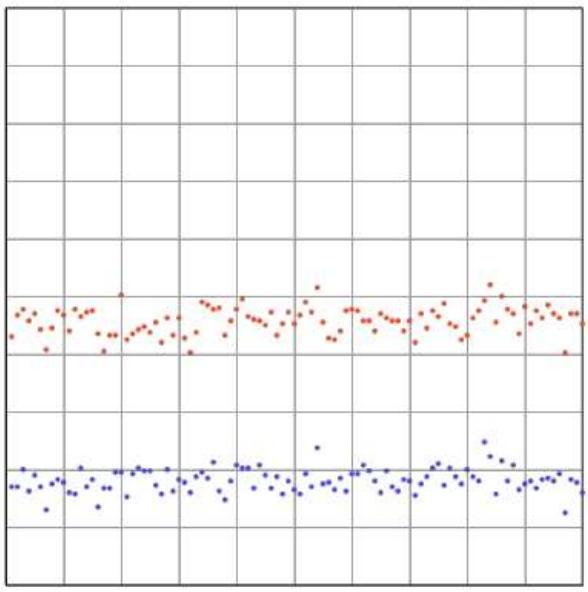
I) Echantillonnage : rappels de Seconde et de Première

Travail de l'élève 1 : En Syldavie, Norbert fait des études statistiques sur le daltonisme des éléphants syldaves pratiquant le saut en parachute.
Il pense que 46% d'entre eux sont des mâles et 18% des éléphants ont plus de 60 ans.
Il a prélevé un échantillon de 400 éléphants parmi lesquels 195 sont des mâles et 313 ont moins de 60 ans.
L'échantillon a été réalisé par un tirage au hasard et il peut être assimilé à un tirage avec remise.

1. L'échantillon de Norbert est-il représentatif ?
2. L'étude de Norbert montre que dans cet échantillon, 29% des éléphants sont daltoniens.
   Estimer la proportion d'éléphants daltoniens dans la population totale.

Démarche pour mener l'activité : On laisse les élèves chercher et comprendre eux-mêmes l'expression « échantillon représentatif », ainsi que la nécessité d'un critère de décision.
On attend notamment les mots « fréquence » et « fluctuation d'échantillonnage ».
On peut parallèlement proposer aux élèves le programme suivant sur algobox qui simule la réalisation de 100 échantillons de tailles 400, et renvoit les nuages de points associés aux fréquences d'apparition des caractères mâles et plus de 60 ans.

```
VARIABLES
    a EST_DU_TYPE NOMBRE
    Males EST_DU_TYPE NOMBRE
    Vieux EST_DU_TYPE NOMBRE
    compteur EST_DU_TYPE NOMBRE
    compteur2 EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
    POUR compteur ALLANT_DE 1 A 100
        DEBUT_POUR
        Males PREND_LA_VALEUR 0
        Vieux PREND_LA_VALEUR 0
        POUR compteur2 ALLANT_DE 1 A 400
            DEBUT_POUR
            a PREND_LA_VALEUR random()
            SI (a<=0.46) ALORS
                DEBUT_SI
                Males PREND_LA_VALEUR Males+1
                FIN_SI
            SI (a<=0.18) ALORS
                DEBUT_SI
                Vieux PREND_LA_VALEUR Vieux+1
                FIN_SI
            FIN_POUR
        TRACER_POINT (compteur,Males/400)
        TRACER_POINT (compteur,Vieux/400)
        FIN_POUR
FIN_ALGORITHME
```



La notion d'échantillon représentatif est une question délicate, en particulier lorsqu'elle concerne des personnes dans le cadre d'un sondage. Elle l'est clairement moins lorsqu'il s'agit d'un échantillon de pièces dans une chaine de fabrication. Cette notion d'échantillon représentatif est évoquée ici afin de contextualiser un peu l'activité mais ne constitue en aucun cas un objectif du programme.
Il convient également de souligner que, dans les sondages, les tirages sont pour la plupart effectués sans remise

mais peuvent s'apparenter à des tirages avec remise dès que la taille de l'échantillon est petite devant la taille de la population totale, ce qui est le cas dans les sondages classiques.
On peut d'ailleurs observer que dans le cas contraire, l'intérêt de ne questionner qu'un échantillon diminue.

Critère de décision
Pour une étude statistique sur certains caractères connus de la population (âge, sexe, taille, etc), on considère qu'un échantillon est **représentatif**, si la fréquence f observée de ces caractères est dans l'**intervalle de fluctuation au seuil de** 95%.

Remarques :

- ⇝ Il s'agit d'un intervalle dans lequel se situent 95% des échantillons établis dans ces conditions.
  Cet intervalle dépend évidemment de la taille n de l'échantillon et de la probabilité p d'apparition du caractère dans la population.
- ⇝ Dans la pratique, on connait n et on peut faire en sorte que notre échantillon soit établi au hasard.
  Par contre, pour p, à moins d'avoir regarder la population entière (et dans ce cas, ce chapitre n'a plus d'intérêt) il s'agit d'une hypothèse.
  Cet intervalle de fluctuation servira donc surtout à vérifier la crédibilité d'une hypothèse sur p.
- ⇝ Ainsi, si la fréquence observée n'appartient pas à l'intervalle de fluctuation au seuil de 95%, on considérera que l'échantillon n'est pas représentatif, ie non établi au hasard, ou encore que l'hypothèse de départ sur p est **mauvaise** et on la rejettera.
  On a un risque de se tromper dans 5% des cas, puisque 5% des fréquences établies dans ces conditions ne sont pas dans cet intervalle. On a donc refusé 5% d'échantillons « en trop ».
  Par contre, si f est dans l'intervalle de fluctuation au seuil de 95%, on considérera l'échantillon comme représentatif, ie établi au hasard, ou encore que notre hypothèse sur p est **crédible**, sans connaître le risque d'erreur, ie le nombre d'échantillons acceptés « en trop ».
  Dans tous les cas, on est sûr de rien ! Donc on évitera le vocabulaire « vrai » ou « faux ».
- ⇝ Tout est une question d'équilibre :
  Si l'on veut diminuer l'erreur de rejet, par exemple en prenant un intervalle de fluctuation au seuil de 100%, on ne rejettera donc aucun échantillon « en trop », par contre, on les acceptera tous, donc évidemment, beaucoup trop. Toute hypothèse sur p semblera crédible, ce qui n'a aucun intérêt.
  Ainsi, dans la pratique, on utilise surtout les seuils de 95% et de 99%.

On en vient aux rappels de seconde et de première.

En seconde
Pour un caractère donné, on note p sa probabilité d'apparition et f sa fréquence d'apparition dans un échantillon.
La fréquence f se situe dans au moins 95% des cas dans l'intervalle $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$.

- **Avantages** : rapide à déterminer
- **Inconvénients** : valable uniquement si $n \geq 25$ ET $0.2 \leq p \leq 0.8$
  Sans parler du fait que cet intervalle sort de nulle part et n'a jamais été justifié par votre enseignant ...

On peut appliquer ce résultat pour le sexe des éléphants, mais pas pour l'âge.
On trouve I = [0.41; 0.51] et $f = \frac{195}{400} = 04875 \in$ I.
Donc l'échantillon est représentatif pour le caractère sexe, ou encore on accepte l'hypothèse $p = 0.46$.

En Première

On appelle X la variable aléatoire qui compte le nombre de personnes d'un échantillon de taille n, possèdant un certain caratère, de probabilité d'apparition p. On a donc $X \hookrightarrow B(n,p)$

On note $f = \dfrac{X}{n}$ la fréquence d'apparition de caractère.

La fréquence f se situe dans au moins 95% des cas dans l'intervalle $\left[\dfrac{a}{n}, \dfrac{b}{n}\right]$, où :

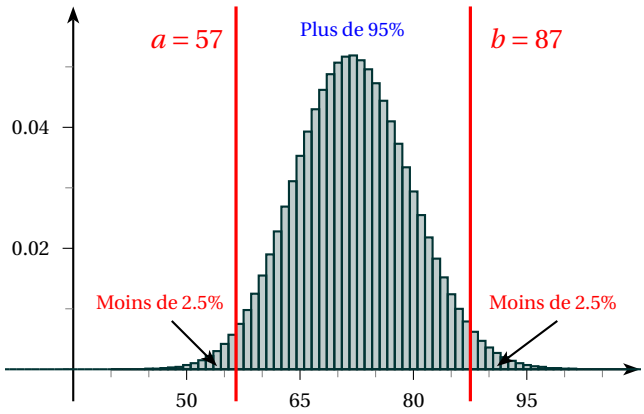
- a désigne le plus petit entier k à partir duquel $P(X \leq k)$ dépasse 0.025.
- b désigne le plus petit entier k à partir duquel $P(X \leq k)$ dépasse 0.975.

- **Avantages** : valable pour toutes les valeurs de n et p
  Sans parler du fait que cet intervalle est justifié par vos connaissances sur la loi binomiale ...
- **Inconvénients** : Très fastidieux si n est grand.

On peut appliquer ce résultat pour l'âge des éléphants.

k	$P(X \leq k)$
...	...
54	0.00953
55	0.01375
56	0.01944
57	**0.02699**
58	0.03678
59	0.04924
...	...

k	$P(X \leq k)$
...	...
83	0.94587
84	0.94587
85	0.95825
86	0.96821
87	**0.97609**
88	0.98225
...	...



Pour l'âge, on trouve $I' = [0.1425; 0.2175]$ et $f = \dfrac{400-313}{400} = 0.2175 \in I'$.
Donc l'échantillon est représentatif pour le caractère âge, ou encore on accepte l'hypothèse $p = 0.18$.
Notons que pour le sexe, on trouve $I' = [0.4125; 0.51]$ et $I \subset I'$. De plus, $f = \dfrac{195}{400} \in I'$.
Donc l'échantillon est représentatif pour le caractère sexe.

Donc l'échantillon est représentatif pour les deux caractères.

Pour la question suivante, on ignore la proportion p d'éléphants daltoniens dans la population.
On cherche cette fois un *intervalle de confiance*.

En Seconde

Pour un caractère donné, on note f sa fréquence d'apparition dans un échantillon de taille n.
Sa probabilité p d'apparition dans la population totale se situe dans au moins 95% des cas dans l'intervalle

$$\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$$

- **Avantages** : rapide à déterminer
- **Inconvénients** : valable uniquement si $n \geq 30$ ET $nf \geq 5$ ET $n(1-f) \geq 5$
  Sans parler du fait que cet intervalle sort de nulle part et n'a jamais été justifié par votre enseignant ...

Ainsi, on estime que la proportion de daltoniens est dans l'intervalle $[0.24; 0.34]$.

Conclusion

Soit p la probabilité d'apparition d'un caractère dans une population totale.

⇝ **Lorsque l'on connait** p, ou l'on fait une hypothèse sur la valeur de p, on utilise un **intervalle de fluctuation à** 95% pour estimer la fréquence d'apparition f du caractère dans un échantillon de taille n.

$$\text{Si } n \geq 25 \text{ , et } 0.2 \leq p \leq 0.8 \text{ :} \qquad \text{I} = \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$$

$$\text{Ou} \qquad \text{I}' = \left[\frac{a}{n}, \frac{b}{n} \right]$$

où si $\text{X} \hookrightarrow \text{B}(n, p)$ et :

- a désigne le plus petit entier k à partir duquel $\text{P}(\text{X} \leq k)$ dépasse 0.025.
- b désigne le plus petit entier k à partir duquel $\text{P}(\text{X} \leq k)$ dépasse 0.975.

⇝ **Lorsque l'on ne connait pas** p, on utilise un **intervalle de confiance à** 95% pour estimer la valeur de p à partir de la fréquence f d'apparition du caractère dans un échantillon de taille n.

$$\text{Si } n \geq 30 \text{ , } nf \geq 5 \text{ et } n(1-f) \geq 5 \text{ :} \qquad \text{J} = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right]$$

II) Intervalle de fluctuation asymptotique

Travail de l'élève 2 : On reprend le contexte de l'activité précédente.

Norbert veut effectuer une étude plus précise avec un échantillon plus grand.

Il a un échantillon de 1200 éléphants parmi lesquels 562 sont des mâles et 951 ont moins de 60 ans.

1. **a.** Pourquoi aimeront-on trouver une autre méthode que celles employées précédemment pour savoir si l'échantillon de Norbert est représentatif ?

   b. Quel nouvel outil de Terminale pourrait-on utiliser ?

   c. Proposer un nouvel intervalle de fluctuation au seuil de 95%.
   Quelle différence a-t-il avec les précédents ? Quels sont ses avantages ? ses inconvénients ?

2. L'étude de Norbert montre que dans cet échantillon, 32% des éléphants sont daltoniens.
   Estimer la proportion d'éléphants daltoniens dans la population totale.

Théorème 1. (Définition)

Soit X_n une variable aléatoire suivant une loi binomiale $\mathscr{B}(n,p)$ et α un réel tel que $0<\alpha<1$.
Soit Z une variable aléatoire suivant la loi normale $\mathscr{N}(0,1)$ et u_α l'unique réel tel que :

$$P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$$

On note I_n l'intervalle :

$$I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

Alors $\lim\limits_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$

Ainsi, l'intervalle I_n contient la fréquence $F_n = \frac{X_n}{n}$ avec une probabilité qui se rapproche de $1-\alpha$ lorsque n augmente : on dit que

- α est le risque (5%, 1% ...)
- I_n est un **intervalle de fluctuation asymptotique** de F_n au seuil $1-\alpha$.

Preuve **ROC**

On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ et on applique le théorème de Moivre-Laplace.
Si X suit la loi normale $\mathscr{N}(0,1)$:

$$\lim_{n\to+\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$$

Or

$$\begin{aligned}
-u_\alpha \leq Z_n \leq u_\alpha &\iff -u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \\
&\iff np - u_\alpha\sqrt{np(1-p)} \leq X_n \leq np + u_\alpha\sqrt{np(1-p)} \\
&\iff p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + u_\alpha\frac{\sqrt{np(1-p)}}{\sqrt{n}} \\
&\iff F_n \in I_n
\end{aligned}$$

Donc $\lim\limits_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = \lim\limits_{n\to+\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = 1-\alpha$

Remarques :

- I_n est un intervalle déterminé à partir de p et n et qui contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.
- Quand on sait qu'une suite (u_n) converge vers une limite L, on peut considérer que pour n assez grand le terme u_n constitue une approximation de L.
  Ici, on inverse les rôles. On connaît la limite, mais pas les valeurs des termes de la suite. On admet donc que, sous certaines conditions, on peut approcher le terme de rang n de la suite $P\left(\frac{X_n}{n} \in I_n\right)$ par sa limite $1-\alpha$.
- On considère que l'approximation est satisfaisante dès que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$

Corollaire 1.

L'intervalle de fluctuation asymptotique au seuil de 95% pour une variable aléatoire X_n suivant une loi binomiale $\mathscr{B}(n,p)$ est l'intervalle :

$$I_n = \left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

Preuve

On a vu au chapitre précédent que $u_{0.05} \simeq 1.96$

Remarques :

- Ainsi, environ 95% des fréquences observées se situent dans l'intervalle ci-dessus.
- Quel est l'intervalle de fluctuation asymptotique au seuil de 99% ?
- L'intervalle de fluctuation vu en seconde est une approximation de l'intervalle I_n au seuil de 95%.

  En effet, $\quad p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

  De plus, en étudiant la fonction f définie sur $[0,1]$ par $f(x) = x(1-x)$, on constate que son maximum est $\frac{1}{4}$.

  Ainsi, pour tout $p \in [0,1]$, on a $\quad p(1-p) \leq \frac{1}{4} \iff \sqrt{p(1-p)} \leq \frac{1}{2}$.

  Par conséquent, $p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p + \frac{1}{\sqrt{n}}$

  De même on a : $p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \geq p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \geq p - \frac{1}{\sqrt{n}}$

  Ainsi $\left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \subset \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$.

  Ceci prouve que l'intervalle de fluctuation vu en seconde est un intervalle de fluctuation asymptotique à un seuil au moins égal à celui de Terminale (proche de 0.95).

  Compte tenu du caractère asymptotique, il reste inexact d'affirmer que la probabilité que la variable $\frac{X_n}{n}$ prenne ses valeurs dans cet intervalle est supérieur à 0.95 pour toute valeur de n, d'où les conditions $n \geq 25$ et $0.2 \leq p \leq 0.8$.

  En réalité, on a encore simplifié les conditions, car par exemple, pour $p = 0.35$, on faudrait prendre $n \geq 31$, pour $p = 0.5$, il faudrait $n \geq 529$ ...

Exemple :

On admet que dans la population d'enfants de 11 à 14 ans d'un département français le pourcentage d'enfants ayant déjà eu une crise d'asthme dans leur vie est de 13%.
Un médecin d'une ville de ce département est surpris du nombre important d'enfants le consultant ayant des crises d'asthme et en informe les services sanitaires. Ceux–ci décident d'entreprendre une étude et d'évaluer la proportion d'enfants de 11 à 14 ans ayant déjà eu des crises d'asthme.
Ils sélectionnent de manière aléatoire 100 jeunes de 11 à 14 ans de la ville.
La règle de décision prise est la suivante : si la proportion observée est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 95% alors une investigation plus complète sera mise en place afin de rechercher les facteurs de risque pouvant expliquer cette proportion élevée.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeunes de 11 à 14 ans ayant eu une crise d'asthme dans un échantillon de taille 100.
2. L'étude réalisée auprès des 100 personnes a dénombré 19 jeunes ayant déjà eu des crises d'asthme.
   Que pouvez-vous conclure ?
3. Le médecin n'est pas convaincu par cette conclusion et déclare que le nombre de personnes interrogées était insuffisant pour mettre en évidence qu'il y avait plus de jeunes ayant eu des crises d'asthme que dans le reste du département.
   Combien faudrait-il prendre de sujets pour qu'une proportion observée de 19% soit en dehors de l'intervalle de fluctuation asymptotique ?

***Solution* :**

1. $[0,06;0,20]$
2. La valeur 0,19 est à l'intérieur de l'intervalle de fluctuation asymptotique au seuil de 95%.
   On en conclut que la règle de décision choisie ne prévoit pas de réaliser une enquête supplémentaire.
3. Il faut et il suffit que la borne supérieure de l'intervalle asymptotique de fluctuation soit inférieur à 0.19, ce qui équivaut à $0.13+1.96\dfrac{\sqrt{0.13\times 0.87}}{\sqrt{n}}<0.19$.
   On trouve $n>120$.
   La taille doit donc être de 121 sujets au minimum si on souhaite mettre en évidence une proportion anormalement élevée dans la ville étudiée.

III) Le point sur l'estimation

Il est souvent difficile pour des raisons à la fois financières et logistiques de pouvoir recueillir des données sur la population toute entière. Le plus souvent, on se contente de travailler sur un échantillon, c'est à dire une fraction ou sous-ensemble de cette population. Ceci présente bien sûr des avantages en termes de faisabilité et de coût, mais impose des contraintes pour que l'information recueillie au niveau de l'échantillon (estimation) soit la plus proche possible de celle de la population entière. La démarche pratique est donc la suivante :

- ⇝ On sélectionne un échantillon de la population que l'on étudie, on appelle cela l'échantillonnage.
- ⇝ On vérifie, selon les cas, à partir d'intervalles de fluctuation que l'échantillon ainsi obtenu est « **représentatif** » de la population pour des critères qui sont connus dans la population et en lien avec le critère étudié.
- ⇝ On généralise les informations recueillies sur notre échantillon à la population entière.

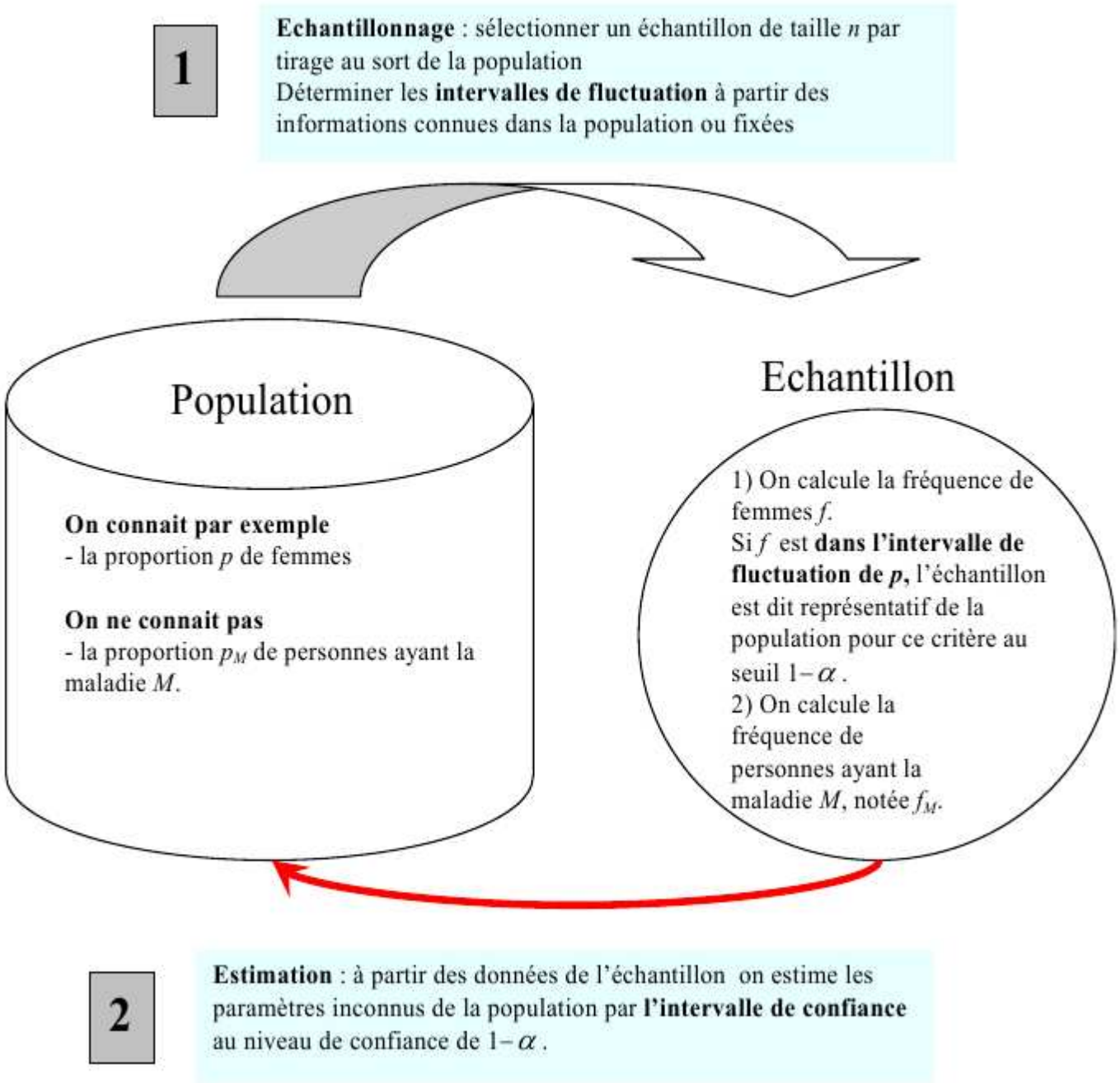
Attention !

La problématique s'inverse : à partir d'une **fréquence observée** f d'un caractère dans un échantillon, on **estime** une **proportion** p dans une population.

Cependant, on sait que notre estimation va varier d'un échantillon à l'autre, de par la fluctuation d'échantillonnage, autour de p.

Comment faire pour limiter la marge d'erreur ?

Il est donc nécessaire d'apprécier l'incertitude en fournissant une estimation par intervalle, appelé **intervalle de confiance** de p. Cet intervalle est obtenu en fonction d'un coefficient lié au niveau de confiance que l'on accorde à cette estimation.



On a déjà rappelé ce que l'on faisait en Seconde. Et bien c'est simple, on n'a pas mieux en Terminale !
Donc on reste avec le même **intervalle de confiance** au seuil de 95%, à savoir :

$$\mathrm{J} = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right]$$

Preuve

On sait que pour n suffisamment grand

$$\begin{aligned}
& \mathrm{P}\left(p - \frac{1}{\sqrt{n}} \leq \mathrm{F}_n \leq p + \frac{1}{\sqrt{n}} \right) \geq 0.95 \\
\iff \quad & \mathrm{P}\left(-\frac{1}{\sqrt{n}} \leq \mathrm{F}_n - p \leq \frac{1}{\sqrt{n}} \right) \geq 0.95 \\
\iff \quad & \mathrm{P}\left(-\frac{1}{\sqrt{n}} - \mathrm{F}_n \leq -p \leq \frac{1}{\sqrt{n}} - \mathrm{F}_n \right) \geq 0.95 \\
\iff \quad & \mathrm{P}\left(\frac{1}{\sqrt{n}} + \mathrm{F}_n \geq p \leq -\frac{1}{\sqrt{n}} + \mathrm{F}_n \right) \geq 0.95 \\
\iff \quad & \mathrm{P}\left(\mathrm{F}_n - \frac{1}{\sqrt{n}} \leq -p \leq \mathrm{F}_n \frac{1}{\sqrt{n}} \right) \geq 0.95
\end{aligned}$$

Ce qui peut se traduire ainsi : l'intervalle $\left[\mathrm{F}_n - \frac{1}{\sqrt{n}}; \mathrm{F}_n + \frac{1}{\sqrt{n}} \right]$ a une probabilité au moins égale à 0.95 de contenir p.
A partir de cet intervalle aléatoire, on obtient, en effectuant un échantillon, une réalisation de cet intervalle numérique de la forme $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$.
Si l'on fait un très grand nombre de tirages par échantillon, on sait que théoriquement on devrait [a] avoir pour au plus 5% des échantillons, des intervalles ne contenant pas la proportion inconnue p.

a. Il s'agit toujours d'un nombre fini de réalisations et il peut y avoir plus de 5% d'entre elles qui ne contiennent pas p.

Remarques :

- ⇝ On se place dans le cas $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$ pour considérer notre estimation convenable.
- ⇝ On pourrait avoir de meilleures approximations, mais elles sont hors-programme
- ⇝ Un intervalle de confiance étant un intervalle numérique, il est incorrect de conclure la détermination d'un intervalle de confiance par une phrase du type « p a une probabilité de 0,95 d'être entre $f - \frac{1}{\sqrt{n}}$ et $f + \frac{1}{\sqrt{n}}$ » car il n'y a plus d'aléatoire à ce stade. Il est en revanche convenable d'écrire : « $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$ est un intervalle de confiance de la proportion inconnue p au niveau de confiance 0,95 ».

Résumé

	Intervalle de **fluctuation** Au seuil 0.95 (p connue)	Intervalle de **confiance** Au seuil 0.95 (p inconnue)
SECONDE	$n \geq 25$, $0.2 \leq p \leq 0.8$ $I = \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$	Sensibilisation
PREMIERE	Avec la loi binomiale $I = \left[\frac{a}{n}; \frac{b}{n}\right]$	
TERMINALE	$n \geq 30$, $np \leq 5$ et $n(1-p) \leq 5$ Asymptotique $I_n = \left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$	$J = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$

Exemple :

Une entreprise désire contrôler la qualité de sa production.
Un échantillon de 370 produits a été testé. Il ressort de cette étude que 26 produits sur 370 ne sont pas conformes au cahier des charges.
On considère que la production de l'entreprise est suffisamment importante pour qu'on puisse assimiler le tirage de l'échantillon à un tirage avec remise.

1. A partir de cette étude, estimer la probabilité qu'un produit ne soit pas conforme au cahier des charges à l'aide d'un intervalle de confiance au niveau asymptotique 95%.
2. Combien faudrait-il prélever de produit pour avoir un intervalle de confiance au niveau 95% de longueur inférieure à 0.05 ?

IV) Compléments sur l'échantillonnage

L'observation d'un échantillon ne permet pas de décrire avec certitude une population mais seulement d'estimer par intervalles de confiance les valeurs de certaines caractéristiques que l'on souhaite connaître dans cette population.

« Voir invoquer la « représentativité » dans un rapport d'enquête pour justifier de la qualité d'un sondage peut presque à coup sûr laisser soupçonner que l'étude a été réalisée dans une méconnaissance totale de la théorie de l'échantillonnage. Le concept de représentativité est aujourd'hui à ce point galvaudé qu'il est désormais porteur de nombreuses ambivalences. Cette notion, d'ordre essentiellement intuitif, est non seulement sommaire mais encore fausse et, à bien des égards, invalidée par la théorie ... » (Y. Tillé, 2011, *Théorie des sondages : Echantillonnage et estimation en populations finies*)

La première chose à préciser c'est qu'avec un échantillon on ne peut pas être représentatif de l'ensemble de la population sur toutes les caractéristiques, il est donc important de définir les caractéristiques qui intéressent les responsables de l'enquête.

Pour un statisticien, l'échantillon sera dit représentatif si on peut correctement estimer les paramètres d'intérêt de la population à partir de l'échantillon. Dans le cas contraire on parlera de biais d'échantillonnage. Pour pouvoir correctement estimer les paramètres, le statisticien n'a pas nécessairement besoin que l'échantillon soit une reproduction miniature de la population, par contre il a besoin que tous les profils de la population importants pour

l'objectif de l'enquête soient représentés dans l'échantillon. Cela signifie donc que le plan d'échantillonnage utilisé dépendra de l'objectif de l'étude même si la population est la même.

La représentativité d'un échantillon nécessite que la procédure d'échantillonnage permette la constitution d'un sous-groupe recouvrant les caractéristiques qui peuvent influencer la valeur des paramètres que l'on veut estimer. La non-représentativité d'un échantillon peut par exemple être due à la sélection dans une base de sondage ne couvrant pas correctement la population.

Par exemple, supposons qu'on souhaite réaliser une enquête de prévalence d'une maladie A dans la population générale et qu'on sélectionne un échantillon à partir de la liste téléphonique (l'enquête devant se dérouler par appel téléphonique). Dans ce cas l'échantillon ne couvre pas correctement la population il y a un biais d'échantillonnage car les personnes qui répondront à l'enquête auront un téléphone et seront présentes à leur domicile, pour cette raison toutes les personnes qui seront hospitalisées à la date de l'enquête ne seront pas interrogées. Si les personnes atteintes de la maladie étudiée sont plus susceptibles de se rendre à l'hôpital, on risque de sous-estimer la prévalence de la maladie ou proportion de malades, en réalisant un échantillon comme proposé ci-dessus.

Dans tous les cas de figures on souhaite enquêter sur un nombre suffisant de sujets afin de pouvoir estimer correctement le paramètre de la population. En principe la taille de l'échantillon est indépendante de la taille de la population que l'on veut étudier. Il faut interroger autant de personnes pour estimer avec la même précision le résultat de l'élection présidentielle en France, que l'élection du maire de Bordeaux.

La taille est en revanche fonction de la marge d'erreur (amplitude de l'intervalle de confiance) que l'on accepte de prendre et qui résulte inéluctablement du fait que l'estimation est issue d'un échantillon.

Un sondage peut être effectué de multiples façons que l'on regroupe en deux grandes familles : les sondages aléatoires, dits aussi probabilistes, et les sondages non aléatoires, dits aussi empiriques ou informels.

Pour les sondage non aléatoires, la sélection des individus n'obéit plus au hasard mais est définie selon des critères de faisabilité, de ressemblance à la population et de critères subjectifs dépendant du choix des enquêteurs.

Les types de sondage satisfaisant aux critères de faisabilité ou de simplicité sont par exemple les échantillons de sujets volontaires.

Les types de sondage satisfaisant aux critères de ressemblance à la population sont appelés échantillonnage par choix raisonné. La méthode des quotas, qui est la méthode la plus utilisée parmi les sondages non aléatoires et dans les sondages d'opinion, fait partie de cette catégorie de sondage. Les enquêteurs doivent inclure un nombre donné d'individus présentant telle ou telle caractéristique dans des proportions voisines de celles de la population. Du moment que le quota est respecté, le mode de sélection des individus est laissé au libre choix de l'enquêteur. La méthode des quotas consiste à construire un échantillon qui soit une maquette, un modèle réduit de la population étudiée, en conservant les mêmes proportions. La plupart des sondages politiques effectués en France utilisent cette méthode.

La date cruciale pour l'histoire de l'échantillonnage est le mardi 3 novembre 1936, jour de la publication des résultats de l'élection présidentielle aux États-Unis. Le journal « Literary Digest » avait réalisé des sondages préélectoraux, comme à leur habitude, par consultation individuelle d'électeurs (appelés « votes de paille » à cette époque). Cette méthode ne fait appel à aucune notion de représentativité, mais est réalisée sur un nombre important d'électeurs et jusqu'en 1936 elle donne des résultats tout à fait satisfaisants. Ce journal comme bien d'autres prédit alors l'élection de Lanton, mais finalement F.D. Roosevelt est élu. Seuls trois sondages l'avaient donné gagnant, tous réalisés par une méthode empirique appelée la méthode des quotas. Ce fut le début des grandes structures de sondages telles que la société de sondage Gallup aux Etats-Unis.

Dans un plan d'échantillonnage aléatoire, tous les individus de la population ont une probabilité connue et non nulle d'être sélectionnés pour faire partie de l'échantillon. La sélection des individus constituant l'échantillon s'effectue par un plan d'échantillonnage à un ou plusieurs degrés et à chaque degré une procédure de tirage au sort est spécifiée ; il peut s'agir d'une procédure de sondage aléatoire simple, ou systématique, ou d'une procédure stratifiée, avec sélection équiprobable ou à probabilité proportionnelle à la taille. Logiquement seuls les sondages aléatoires permettent de fournir des estimations avec une précision donnée, c'est- à -dire avec un intervalle de confiance.

Les sélections aléatoires à partir d'une liste d'individus peuvent s'effectuer de différentes façons. Prenons l'exemple d'une enquête que l'inspection académique souhaiterait réaliser auprès des élèves des lycées d'un département afin d'étudier les difficultés scolaires rencontrées par ceux-ci. Il est impossible d'interroger la totalité des élèves et le souhait est de pouvoir obtenir des informations sur un échantillon représentatif de 500 élèves. Pour cette dernière raison il est décidé de sélectionner aléatoirement les élèves, mais plusieurs méthodes peuvent être proposées.

1. si la liste de tous les élèves est accessible de manière électronique on peut sélectionner 500 élèves dans la liste en utilisant par exemple un tableur, il y a plusieurs méthodes pour cela :
   a. créer pour chaque élève un nombre aléatoire suivant une loi uniforme, puis choisir de trier la liste en fonction de ce nombre aléatoire créé, cela revient à mélanger de façon aléatoire la liste. On sélectionne finalement les 500 premiers noms qui sont dans la liste triée. Cette méthode permet de réaliser une sélection simple sans remise.
   b. Numéroter tous les élèves de la liste, puis utiliser la fonction aléatoire du tableur pour sélectionner uniquement 500 nombres, les élèves correspondant à ces nombres seront sélectionnés. En appliquant cette méthode un nombre peut être sélectionné plusieurs fois. Cela revient donc à réaliser un échantillon avec remise.
   c. On peut aussi utiliser la méthode de sélection systématique, c'est-à-dire que si le nombre d'élèves est égale à N on tire au sort un nombre, noté d, entre 1 et N puis on sélectionne de manière régulière sur la liste le $d + \dfrac{\text{N}}{500}$ -énième élèves, si ce nombre 500 dépasse le rang du dernier élève on reprend la liste au début.

   Les trois méthodes présentées ci-dessus sont des sélections que l'on peut considérer équiprobables car chaque sujet a la même probabilité d'être sélectionné.
2. On peut souhaiter effectuer une enquête en face à face, c'est-à-dire qu'un enquêteur doit se déplacer pour interroger l'élève, il est donc important d'essayer de gérer le nombre de déplacements. Dans les méthodes proposées précédemment rien n'est contrôlé et l'enquêteur peut devoir traverser le département pour interroger un et un seul élève. Afin d'améliorer cela on peut décider de sélectionner un certain nombre d'établissements et de sélectionner un certain nombre d'élèves dans chaque établissement. On parlera alors de sondage à plusieurs degrés. Dans ce cas la sélection n'est pas toujours équiprobable.

   Exemple :
   Supposons que 10 des 70 lycées soient sélectionnés et dans chaque lycée sélectionné on sélectionne 30 élèves. Dans ce cas la probabilité que l'élève A soit sélectionné est environ égale à

   $$\frac{10}{70} \times \frac{30}{\text{nb d'élèves du lycée d'appartenance de l'élève A}}$$

   on remarque que cette probabilité dépend de la taille du lycée et donc non équiprobable.

   On peut vouloir construire un échantillon représentant les lycées généraux et professionnels. Dans ce cas et afin de forcer cette représentativité, on commence par partager en deux paquets la liste : liste des lycées professionnels et liste des lycées généraux et on effectue un échantillon dans chacune des deux listes. On parle alors de sondage stratifié.