COURS
ECHANTILLONNAGE ET ESTIMATIONS

I) Echantillonnage

I.1. Introduction

Travail de l'élève 1 :

Un candidat A est élu à une élection avec 52% des voix.
On effectue un sondage sur un échantillon de cent personnes choisies au hasard à la sortie des urnes.

- ⇝ La fréquence de gens qui ont voté pour le candidat A dans l'échantillon de 100 personnes peut-elle être exactement 52% ?
- ⇝ Peut-elle être différente de 52% ?

Pour chaque personne interrogée, on s'intéresse à son vote :

- ⇝ Soit pour le candidat A, ce qui est considéré comme un succès
- ⇝ Soit pour un autre candidat.

On considère que chaque personne interrogée a voté indépendamment des autres.
La variable aléatoire X qui à chaque échantillon de cent personnes associe le nombre de succès suit donc la loi binomiale $B(100\ ;\ 0,52)$

Comme n est assez grand et p n'est pas trop proche de 0 ou ni de 1, on peut considérer que les conditions sont réunies pour approximer cette loi binomiale par une loi normale de paramètres $\mu = np = 52$ et $\sigma = \sqrt{np(1-p)} \simeq 5$.

On sait alors que

$$P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \simeq 95\%$$

La fréquence de gens qui ont voté pour A est donnée par $\dfrac{X}{n}$. On a donc :

$$P\left(\frac{\mu-2\sigma}{n} \leq \frac{X}{n} \leq \frac{\mu+2\sigma}{n}\right) \simeq 95\% \iff P\left(p - 2\sqrt{\frac{p(1-p)}{n}} \leq \frac{X}{n} \leq p + 2\sqrt{\frac{p(1-p)}{n}}\right) \simeq 95\%$$

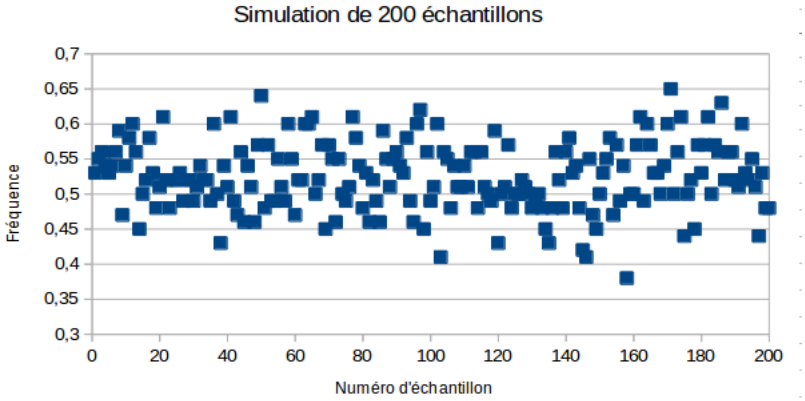
En fait, pour être plus précis, on a

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \frac{X}{n} \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \simeq 95\% \iff P\left(0.42 \leq \frac{X}{n} \leq 0.62\right) \simeq 95\%$$

Concrètement, cela signifie que si l'on réalise un grand nombre d'échantillons de 100 personnes à la sortie des urnes, environ 95% d'entre eux auront une fréquence de vote pour A comprise entre 0.42 et 0.62

On a effectué une simulation sur tableur pour 200 échantillons de 100 personnes.

Simulation du vote pour chaque personne : `=SI(ALEA()<0.52;1;0)`
Fréquence par échantillon : `=Moyenne()`



On observe sur tableur qu'environ 95% des fréquences d'échantillons de taille 100 sont bien dans l'intervalle $[0,42\ ;\ 0,62]$

I.2. Intervalle de fluctuation asymptotique à environ 95%

Définition 1. (Proposition)

Quand on prélève un échantillon de taille n dans une population qui contient une proportion p d'un caractère étudié, alors la fréquence f de ce caractère dans l'échantillon appartient à l'intervalle

$$\mathrm{I}_f = \left[p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \frac{\mathrm{X}}{n} \leq p + 1.96\sqrt{\frac{p(1-p)}{n}} \right]$$

avec une probabilité d'environ 95%.

Cet intervalle est appelé **Intervalle de fluctuation asymptotique à** 95%

Remarque : Pour avoir le droit d'affirmer cela, il suffit de pouvoir approximer la loi binomiale par la loi normale, ie que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$

Exemple :

Lors de l'élection présidentielle de 2002, 16,2% des bulletins exprimés étaient en faveur de Lionel Jospin.
A la sortie des urnes, un sondeur a interrogé un millier de personnes. Soit f la fréquence de personnes qui ont voté L. Jospin dans cet échantillon.

1. Déterminer l'intervalle de fluctuation asymptotique de f à 95%. Interpréter.
2. Qu'arrive-t-il à l'intervalle si l'échantillon devient de taille $n = 2000$? et $n = 10000$?

I.3. Prise de décision

Introduction : On cherche à savoir si un dé est pipé ou non.
Pour cela, on peut par exemple le lancer un grand nombre de fois, et regarder la fréquence d'apparition du 1.
Si le dé n'est pas pipé on devrait trover une fréquence proche de $\frac{1}{6}$. Mais qu'entend-on par «proche» ?
C'est là où l'I_f peut fournir un critère de décision :

- ⇝ On calcule l'I_f correspondant à notre hypothèse $f = \frac{1}{6}$ pour les n lancers effectués.
- ⇝ Si la fréquence observée d'apparition du 1 dans l'échantillon est dans l'I_f, alors on peut penser qu'elle est assez proche de $\frac{1}{6}$, et donc accepter l'hypothèse que le dé n'est pas pipé.
- ⇝ Sinon, on peut considérer que la fréquence observée est trop loin de $\frac{1}{6}$, et donc penser que le dé est pipé

Remarque : Est-on pour autant sûr de notre réponse, dans un cas comme dans l'autre ? ?
Evidemment non !

- ⇝ Dans le cas où l'on accepte l'hypothèse, on n'est quand même sûr de rien, mais on n'est pas capable d'évaluer le risque d'erreur
- ⇝ Dans le cas où l'on refuse l'hypothèse, c'est parce que l'on considère que notre échantillon n'est pas cohérent avec la théorie, puisque l'I_f ne contient pas f. Mais la théorie dit justement que seuls 95% des échantillons le sont ! Donc on a un risque de 5% de se tromper.

Exemple :
Dans une serre où sont élevées des drosophiles, le pourcentage de mouches du type «yeux rouge» est censé être de 80% si aucun facteur extérieur ne vient provoquer des mutations.
On prélève un échantillon de 1000 mouches et on en compte 760 du type yeux rouges.
Peut-on détecter la présence d'un facteur extérieur ?

Méthode à suivre

1. On repère dans l'énoncé l'hypothèse faite sur une proportion p dans la population.
2. On détermine l'intervalle de fluctuation asymptotique à 95% des fréquences sur les échantillons de taille n.
3. On prend une décision :
   - ⇝ Si la fréquence f observée dans l'échantillon appartient à l'intervalle de fluctuation, on accepte l'hypothèse selon laquelle la proportion est bien p dans la population, puisque l'ensemble est cohérent.
   - ⇝ Si la fréquence f observée dans l'échantillon n'appartient pas à l'intervalle de fluctuation, on rejette l'hypothèse selon laquelle la proportion est bien p dans la population, puisque cela semble peu cohérent

II) Estimation

II.1. Introduction

Travail de l'élève 2 :
Matériel : On dispose d'une bouteille opaque que l'on ne peut ouvrir, contenant un certain nombre de billes noires et blanches (resp. 9 et 8, non dit aux élèves évidemment). Quand on retourne la bouteille, on distingue près du bouchon trois billes à chaque fois.

Montrer simplement la bouteille aux élèves et leur dire que l'on cherche à connaître la proportion p de billes noires.
Question : Proposer un protocole.

Voici diverses questions à poser aux élèves au fur et à mesure :

- ⇝ Le terme «connaître» est-il approprié ? Que pourrait-on dire de mieux ?
- ⇝ Aurait-on obtenu la même estimation en prélevant un autre échantillon de même taille ?
- ⇝ La taille de l'échantillon est-elle suffisante au vu du nombre de billes dans le sac ?
- ⇝ Quelle confiance accorder au fait que l'échantillon ait conduit à une proportion de ... % ?
- ⇝ Proposer un intervalle dans lequel on peut estimer que p soit, avec une confiance de 95%.
- ⇝ Aurait-on gagné en précision si l'on avait augmenté la taille de l'échantillon ?
- ⇝ Proposer des problèmes concrets où l'on se pose le même genre de questions.

Résultats dans une classe : Stéphane a regardé 100 fois trois perles à la fois. Même si ce n'est pas le cas à cause de ma bouteille mal construite, on va considérer que Stéphane a regardé un échantillon de 300 perles, dans un tirage **avec remise** (ce qui sera toujours le cas dans les exercices).
Il a obtenu 177 perles noires contre 123 perles blanches, donc on peut estimer la fréquence de perles noires à $\frac{177}{300} =$

$\frac{59}{100}$, soit environ 10 perles sur les 17 présentes.
L'intervalle de confiance à 95% associé est :

$$\left[0.59 - 1.96\sqrt{\frac{0.59\times 0.41}{300}}\ ;\ 0.59 + 1.96\sqrt{\frac{0.59\times 0.41}{300}}\right] \simeq [0.53;0.65]$$

Il faut arrondir par défaut et par excès en fonction des bornes inf et sup ...

Le problème de l'*estimation* est celui qui se pose dans la pratique des sondages ou des contrôles de qualité : à partir d'une fréquence f observée sur un échantillon, estimer la fréquence p correspondante, dans la population.
Sauf à prendre comme échantillon la population toute entière (et dans ce cas, c'est un recensement et non un sondage) cette estimation dépend du hasard qui a amené l'échantillon. **Il n'y aura aucune certitude.** C'est le type même du raisonnement inductif, qui n'a pas, on le verra, le caractère de la rigueur habituelle en mathématiques. Mais c'est ça ou rien...

On procède par sondages pour des raisons économiques. Il se peut, en particulier, qu'un contrôle de qualité nécessite la destruction des pièces testées, comme pour la mesure de résistance aux chocs...

II.2. Intervalle de confiance d'une proportion

Définition 2. (Proposition)

Dans un échantillon de taille n, on observe une fréquence f d'apparition d'un caractère.
On peut alors estimer que la proportion p d'apparition de ce caractère dans la population totale appartient à l'**intervalle de confiance**

$$\mathrm{I}_c = \left[f - 1.96\sqrt{\frac{f(1-f)}{n}}\ ;\ f + 1.96\sqrt{\frac{f(1-f)}{n}}\right]$$

avec un niveau de confiance de 95% (ou encore au risque de 5%)

Remarques :

- ⇝ On se place encore une fois dans le cas où : $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$.
- ⇝ Cet intervalle a pour centre la fréquence f de l'échantillon considéré et change en fonction de l'échantillon considéré. Une même proportion p a donc une infinité d'intervalles de confiance au seuil de 95% ...
- ⇝ «On distinguera confiance et probabilité :
  - — avant le tirage d'un échantillon, la procédure d'obtention de l'intervalle de confiance a une probabilité de 0.95 que cet intervalle contienne le paramètre inconnu p,
  - — après le tirage, le paramètre p est dans l'intervalle calculé avec une confiance 95%.»

  En effet, on ne peut pas dire que p a 95 % de chances d'appartenir à un intervalle de confiance donné tel que $[0,504;0,696]$. Cette expression ne contient rien d'aléatoire, et p est, ou non, dans cet intervalle, sans que le hasard n'intervienne.
  On peut simplement dire par exemple que, **sur un grand nombre d'intervalles de confiances (obtenus à partir d'un grand nombre d'échantillons), environ 95% contiennent effectivement la valeur de** p, ou encore que l'on a 95% de chances d'exhiber un intervalle contenant p (avant le tirage de l'échantillon).

Preuve **Hors Programme**

Introduisons la variable aléatoire d'échantillonnage X qui, à tout échantillon de taille n prélevé au hasard avec remise, associe le nombre de boules noires contenues dans l'échantillon.

On sait que X suit la loi binomiale $\mathscr{B}(n,p)$ laquelle est proche de la loi normale $\mathscr{N}(np, \sqrt{np(1-p)})$.

Par conséquent, la variable aléatoire $F = \dfrac{X}{n}$, qui représente la fréquence de billes noires de l'échantillon, suit une loi proche de $\mathscr{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Ainsi, on peut déterminer le réel positif h tel que $P(p-h \leq F \leq p+h) = 0,95$. On a :

$$\begin{aligned} P(p-h \leq F \leq p+h) = 0,95 &\iff P(-h \leq F-p \leq +h) = 0,95 \\ &\iff P\left(-\frac{h}{\sigma} \leq \frac{F-p}{\sigma} \leq \frac{h}{\sigma}\right) = 0,95 \\ &\iff P\left(-\frac{h}{\sigma} \leq Z \leq \frac{h}{\sigma}\right) = 0.95 \end{aligned}$$

où Z suit la loi normale $(0,1)$. A la calculatrice, on peut alors trouver $\dfrac{h}{\sigma} = 1.96 \iff h = 1.96\sqrt{\dfrac{p(1-p)}{n}}$

Faire la dernière étape avec les élèves à l'aide du schéma de la densité de la loi normale et de la calculatrice : InvNorm$((1+\alpha)/2)$

Ainsi, pour chaque valeur de p, on sait que la variable aléatoire F prendra ses valeurs dans l'intervalle

$$I = \left[p - 1,96\sqrt{\frac{p(1-p)}{n}}\ ;\ p + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

avec une probabilité de 95%.

Cet intervalle est appelé **intervalle de fluctuation**

Supposons que $n = 100$ (la taille de l'échantillon est connue) alors I = ...

Mais l'estimation est le problème inverse de la fluctuation, puisqu'on cherche p à partir de la fréquence observée f, ie de son estimation ponctuelle. Or, on a l'équivalence

$$p - 1,96\sqrt{\frac{p(1-p)}{n}} < F < p + 1,96\sqrt{\frac{p(1-p)}{n}} \iff F - 1,96\sqrt{\frac{p(1-p)}{n}} < p < F + 1,96\sqrt{\frac{p(1-p)}{n}}$$

Comme p se situe dans les deux bornes, on y substitue, faute de mieux, son estimation ponctuelle f observée.

En réalité, c'est un peu plus compliqué mais en gros, c'est l'idée

Ainsi de façon générale, on proposera comme intervalle de confiance de la fréquence p sur la population totale, au coefficient de confiance de $1-\alpha$, l'intervalle :

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

où u_α est donné par la calculatrice en utilisant la commande InvNorm pour loi normale $N(0,1)$ et de paramètre $1 - \dfrac{\alpha}{2}$.

Dans le cas précédent, pour $f = 0,6$; $n = 100$ et A = 95%, on obtient $u_{5\%} = 1,96$ et comme intervalle de confiance pour p : $[0,504;0,696]$, centré sur la valeur observée $f = 0,6$.

Exemple :

Un sondage dans une commune révèle que sur les 500 personnes interrogées, 42% sont mécontentes de l'organisation des transports. Déterminer, au risque de 5%, un intervalle de confiance du pourcentage p de personnes mécontentes dans la commune.

II.3. Mieux comprendre le sens

Objectifs :

⇝ Etudier l'impact du coefficient de confiance et de la taille de l'échantillon.

⇝ Expérimenter la dépendance des intervalles de confiance à l'échantillon choisi.

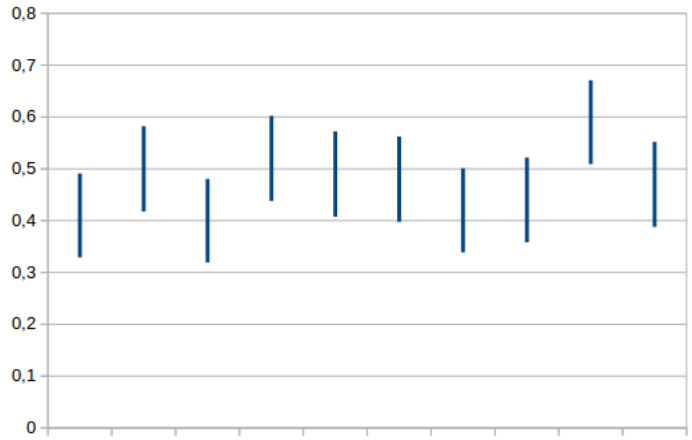
Travaillons sur un exemple d'élection où les scores étaient particulièrement serrés : Le 10 mai 1981, François Mitterrand a été élu avec 51,75% des voix, alors que Valery Giscard d'Estaing n'a recueilli que 48,25% des suffrages. On suppose que l'on effectue des sondages le jour de l'élection, pour estimer la proportion p des partisans de Giscard dans l'électorat (en réalité, $p = 0,4825$).

PARTIE A : **Taille $n = 100$ et coefficient de confiance** = 90%

Ouvrir la feuille de calculs et expliquer sa construction.

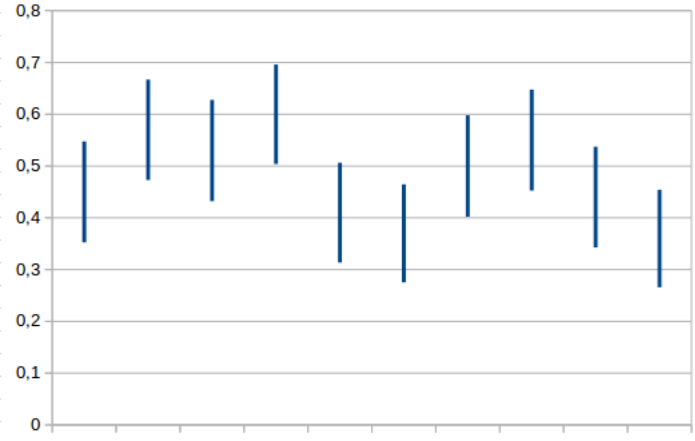
Sur un exemple de 10 sondages,

1. Sur cette feuille, combien de sondages donne Giscard vainqueur (à tort évidemment)
2. Combien d'intervalles de confiance prévoient complètement la victoire de Mitterrand ?
3. Combien d'intervalles de confiance contiennent effectivement p ?
4. Deux intervalles de confiances ont-ils obligatoirement le même centre ?
5. Deux intervalles de confiance peuvent-ils n'avoir aucun élément commun ?
6. Est-ce que $p = 0,4825$ appartient nécessairement à l'intervalle de confiance donnée par un sondage ?
7. Quel est, sur 100 sondages observés, le pourcentage d'intervalles à 90% de confiance ne contenant pas la valeur p à estimer ?



PARTIE B : **Taille $n = 100$ et coefficient de confiance** 95%

Il suffit de modifier la cellule correspondante. Quel est l'impact sur l'IC ?

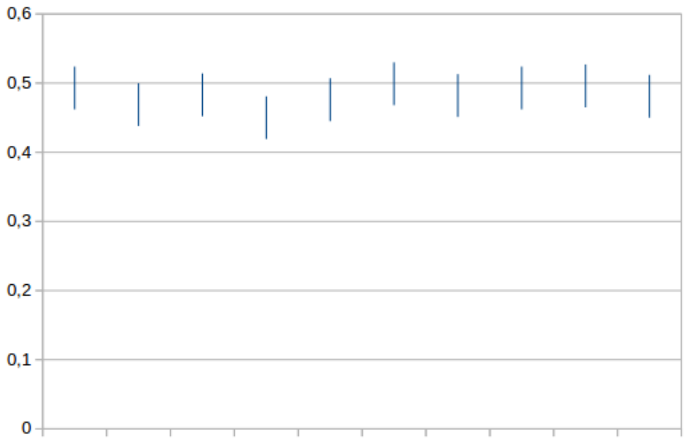


Les intervalles de confiance à 95% se "trompent" moins souvent (statistiquement, on peut observer qu'environ 95% contiennent la valeur p) parce qu'ils sont plus "longs".

On peut observer de légers écarts d'amplitude entre les différents intervalles, selon que f est petit (échantillon 6) ou grand (échantillon 4).

PARTIE C : **Taille $n = 1000$ et coefficient de confiance** 95%

On fait une nouvelle feuille de calcul avec 10 nouveaux échantillons de taille $n = 1000$. Quel est l'impact sur l'IC ?



La plus grande qualité de l'information se traduit par une amplitude notablement réduite des intervalles.
Il y a toujours, statistiquement, 95% qui contiennent p (sur l'image, l'échantillon 4 ne contient pas p).

II.4. Application : Comparaison de deux proportions

Exemple :

On s'intéresse à l'efficacité supposée d'un médicament pour soigner le dos.
Pour cela, on administre un placebo à 1300 patients et le médicament à 1300 autres.
La fréquence de personnes qui sont soulagées avec le placebo est de 40%.
La fréquence de personnes qui sont soulagées avec le médicament est de 44%.

Peut-on juger que ce médicament est efficace ?

Méthode

1. On fait l'hypothèse que deux échantillons de taille n sont issus de la même population relativement à un caractère, et donc que les deux proportions sont égales.
2. On détermine les deux intervalles de confiance à 95% de p pour chacun des échantillons
3. $\rightsquigarrow$ Si les deux intervalles de confiance sont disjoints, alors la différence entre les deux fréquences est considérée comme significative : on rejette l'hypothèse selon laquelle les deux proportions sont égales (avec un risque autour de 5%).
   $\rightsquigarrow$ Si les deux intervalles de confiance ne sont pas disjoints, on ne peut pas rejeter l'hypothèse selon laquelle les deux proportions sont égales.

Solution :

On fait l'hypothèse que les proportions sont égales, c'est-à-dire que les deux échantillons font partie de la même population (i.e. : le médicament n'est pas efficace et donc n'a pas plus d'influence que le placebo). On calcule les deux intervalles de confiance correspondant et on regarde s'ils sont disjoints ou non pour conclure.