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

— Soit pour

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

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

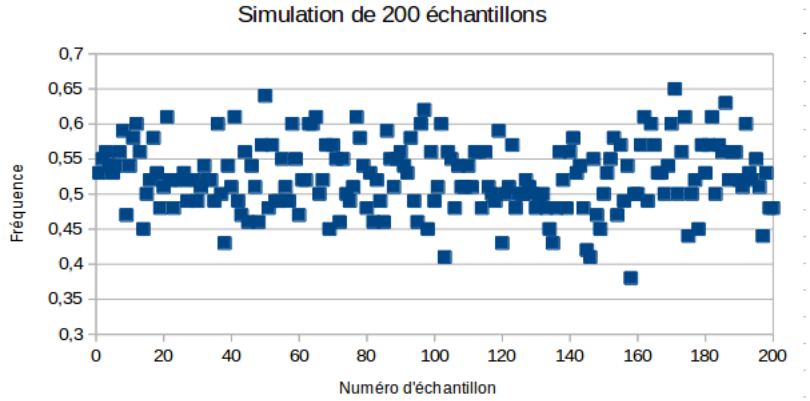
On sait alors que $P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \simeq \ldots$

La fréquence de gens qui ont voté pour A est donnée par $\frac{X}{n}$.

On a donc $P\left(\leq \frac{X}{n} \leq \right) \simeq \ldots$

Concrètement, cela signifie que ...

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



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%. Il est à connaître par coeur.

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 $n = 10000$?

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 générale 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.

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**

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

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).

Exemple :

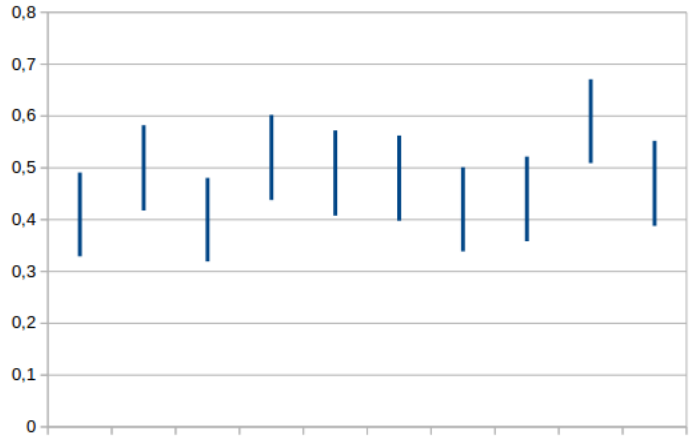
Un sondage dans une commune révèle que sur les 500 personnes interrogées, 42% sont mécontentes de l'organisation des transports. Déterminer, au risque de 5%, un intervalle de confiance du pourcentage p de personnes mécontentes dans la commune.

II.3. Mieux comprendre le sens de la confiance ou du risque

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, $p = 0,4825$ mais supposons que nous ne le savons pas encore).

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

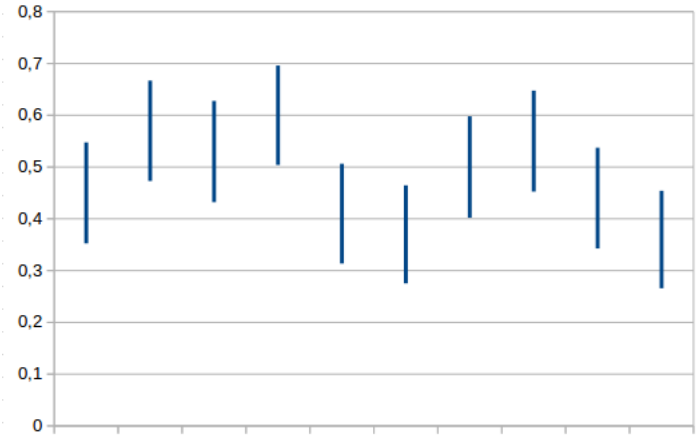
A l'aide d'un tableur, on a simulé des 10 sondages de 100 personnes à la sortie des urnes.
On a alors représenté les intervalles de confiance au niveau de confiance de 90% de la proportion p de votants pour Giscard pour chaque sondage.



1. 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 a priori, 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 niveau de confiance de** 95%

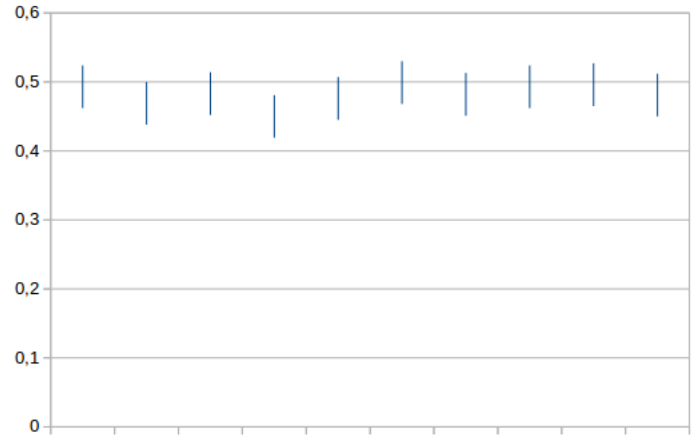
On reprend le principe précédent, mais avec un niveau de confiance de 95%.



Quel impact constatez-vous sur les intervalles de confiance ?

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

On reprend le principe précédent, mais avec 10 nouveaux sondages de 1000 personnes.



Quel impact constatez-vous sur les intervalles de confiance ?

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. — 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%).
   — Si les deux intervalles de confiance ne sont pas disjoints, on ne peut pas rejeter l'hypothèse selon laquelle les deux proportions sont égales.