Statistiques

Nathalie GUYADER
nathalie.guyader@gipsa-lab.fr

Année 2021-2022

Chapitres précédents

- Séance 1:
  - Introduction générale
  - Chapitre 1: Probabilités
  - Chapitre 2: Variables aléatoires
  - Chapitre 3: Statistiques descriptives

- Séance 2:
  - Introduction à Python
  - Prise en main d'un Jupyter Notebook
  - TP1: Python et Statistiques descriptives

- Séance 3:
  - TP2: Statistiques descriptives (données sur les véhicules) et compte-rendu sous Jupyter Notebook

Chapitre 4

Estimation par intervalle de confiance

Introduction à la statistique inférentielle

Dénition (Larousse)

- L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon.

On considère une population dont les éléments possèdent un caractère mesurable qui est la réalisation d'une variable aléatoire X de loi a priori inconnue

On prélève un échantillon de cette population : on suppose que la population est infinie ou si elle est finie que l'échantillonnage se fait avec remise

Toutefois cette dernière hypothèse peut être relaxée si la population est très grande : un tirage sans remise est alors équivalent à un tirage avec remise

Introduction à la statistique inférentielle

Le but de ce chapitre est l’estimation. Il est construit en 2 parties:

- une partie sur l’échantillonnage (qui permet de mieux comprendre celle sur l’estimation); cette partie n’est pas nécessaire et elle n’est pas forcément présentée dans d’autres cours sur les statistiques inférentielles et l’estimation.

- une partie sur l’estimation ponctuelle et par intervalle de confiance

Introduction à la statistique inférentielle

On sait calculer des indicateurs numériques à partir d'un échantillon de données, mais :

- comment généraliser à la population entière ?
- quelles informations sur la population obtient-on en étudiant l'échantillon ?
- quelle confiance peut-on accorder à ces informations

Introduction à la statistique inférentielle

- Idée principale :

à partir d'un échantillon représentatif on va faire des conclusions sur toute la population.

On étudie une variable X, dont on observe des réalisations. On suppose que X suit une loi connue dépendant d'un paramètre .

Plus précisément, on choisit parmi les modèles existants la loi la plus appropriée pour décrire le phénomène observe. Seule la valeur numérique de α est inconnue :

$$X \sim P(\alpha)$$

Exemple : Soit X la durée de vie des ampoules fabriquées dans une usine. On suppose que X suit une loi exponentielle de paramètre α. On va donc chercher à estimer α à partir d'un échantillon de données

Théorie de l’échantillonnage

- <u>Population et variable aléatoire X:</u>

La population E est un ensemble fini. Les éléments de E sont appelés *individus*. Chaque individu E_k a une mesure réelle x_k. On considère l’expérience aléatoire mesure qui consiste à tirer un individu au hasard dans E. Soit alors la variable aléatoire (VA) mesure X définie à partir de cette expérience par :

$$X : E \to \mathbf{R}$$

$$e_k \to x_k$$

On note:

$$E(X) = \frac{1}{N}\sum_{i=k}^{N} x_k = \mu \qquad V(X) = \frac{1}{N}\sum_{i=k}^{N} (x_k - \mu)^2 = \sigma^2$$

Théorie de l'échantillonnage

- n-échantillonnage de E:

On considère l'expérience aléatoire n-échantillonnage qui consiste à tirer, avec remise, n individus au hasard dans E. Soit alors, la VA définie à partir de cette expérience par :

$$X_k : E^n \to \mathbf{R}$$

qui associe au n-uplet tiré la mesure du k° individu tiré. On peut alors considérer le vecteur aléatoire :

$$(X_1, \ldots, X_n) : E^n \to \mathbf{R}^n$$

$X_1, \ldots, X_n$ suivent la même loi et sont indépendantes

Théorie de l'échantillonnage

- <u>Statistique de n -échantillonnage:</u>

  Par définition c'est une VA qui est fonction des VA X_1, X_2 …X_n. Les plus connues sont les statistiques d'échantillonnage:

Moyenne d'échantillonnage:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Variance d'échantillonnage:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Moyenne d'échantillonnage

- **Fluctuations de la moyenne d'échantillonnage**

  Pour être en mesure d'estimer la moyenne de la population par intervalle de confiance ou encore d'effectuer un test d'hypothèse sur la moyenne , il faut connaître la VA $\bar{X}$

  Distribution d'échantillonnage de $\bar{X}$*:* c'est la distribution des différentes valeurs que peut prendre la moyenne d'échantillonnage calculée sur tous les échantillons possibles de même taille d'une population donnée.

  Pour mieux comprendre à quoi correspond $\bar{X}$ nous allons prendre un EXEMPLE

  Attention dans la suite du cours nous n'aurons jamais accès à toutes les valeurs prises par $\bar{X}$!!!

Moyenne d'échantillonnage

Exercice:

- Expérience d'échantillonnage avec remise dans la population : détermination de la distribution de la moyenne de 5-échantillonnage.

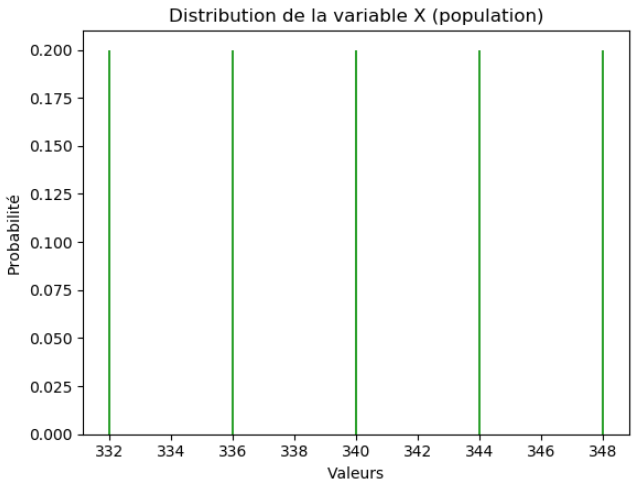
Supposons qu'une population consiste en 5 contenants (numérotés de 1 à 5) et que le poids respectif de chacun est : x_1=332g, x_2=336g, x_3=340g, x_4=344g et x_5=348g.

1/ Déterminer la moyenne, la variance et la distribution de la population (du caractère « poids » de la population). Soit X la VA qui associe à chaque contenant son poids, alors :

Moyenne : $E(X) = 340$

Variance : $\mathrm{V}(X) = 32$ Attention le calcul de la variance se fait sur la population(normalisation en $1/n$)!

Distribution :



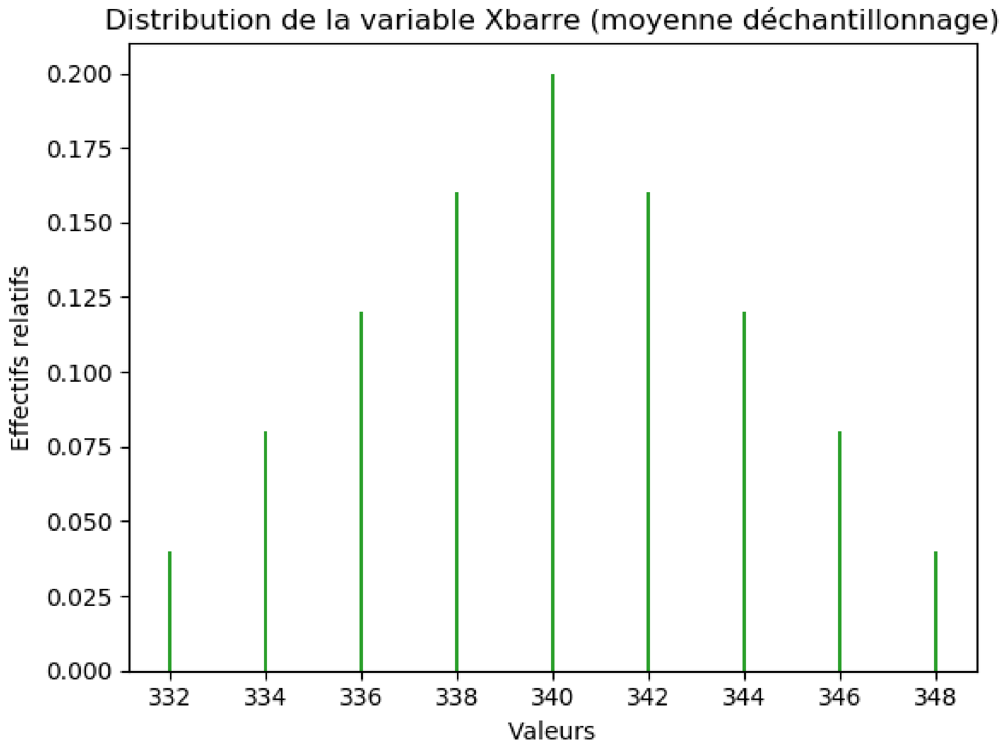
Supposons maintenant que nous voulons former tous les échantillons possibles de taille n = 2 de cette population en effectuant un échantillonnage avec remise. Il y a dans ce cas 5^2 échantillons différents possibles. Chaque échantillon a donc une probabilité égale à 1/25 d'être choisi.

Echantillon n°	**Contenant n°**	**Résultat de l'échantillonnage**	**Moyenne des échantillons**
1	(1,1)	(332,332)	332
2	(1,2)	(332,336)	334
3	(1,3)	(332,340)	336
4	(1,4)	(332,344)	338
5	(1,5)	(332,348)	340
6	(2,1)	(336,332)	334
7	(2,2)	(336,336)	336
8	(2,3)	(336,340)	338
9	(2,4)	(336,344)	340
10	(2,5)	(336,348)	342
11	(3,1)	(340,332)	336
12	(3,2)	(340,336)	338
13	(3,3)	(340,340)	340
14	(3,4)	(340,344)	342
15	(3,5)	(340,348)	344
16	(4,1)	(344,332)	338
17	(4,2)	(344,336)	340
18	(4,3)	(344,340)	342
19	(4,4)	(344,344)	344
20	(4,5)	(344,348)	346
21	(5,1)	(348,332)	340
22	(5,2)	(348,336)	342
23	(5,3)	(348,340)	344
24	(5,4)	(348,344)	346
25	(5,5)	(348,348)	348

Moyenne d’échantillonnage

Proposer une solution pour programmer en Python la génération de la variable qui contient la moyenne des échantillons (soit la dernière colonne du tableau précédent)

2/ Déterminer ensuite la distribution de la moyenne d’échantillonnage:



Moyenne d’échantillonnage

3/ Pour finir, calculer la moyenne et la variance de la moyenne d’échantillonnage et les exprimer en fonction de la moyenne et de la variance de la population.

Moyenne : $E(\bar{X}) = 340$

Variance : $\mathrm{V}(\bar{X}) = 16$

On constate donc que :

$E(\bar{X}) = 340 = E(X)$

$\mathrm{V}(\bar{X}) = 16 = \frac{V(X)}{n}$

Moyenne d'échantillonnage

Remarque 1:

Si nous répétons cette expérience en prélevant cette fois des échantillons de taille n = 3, il y aura 5^3 = 125 échantillons possibles (tirage avec remise) et la moyenne de la distribution d'échantillonnage de sera à nouveau 340 g.

Toutefois la variance des moyennes d'échantillonnage va diminuer : on obtiendrait 32/3. De plus, la forme de la distribution d'échantillonnage de s'approchera de plus en plus de celle d'une loi normale.

Remarque 2 :

Dans la suite du cours et de manière plus générale en statistique vous n'aurez jamais accès à toutes les valeurs prises par la moyenne d'échantillonnage !

Cette partie avait pour but de vous faire comprendre à quoi correspond la variable .

Moyenne d'échantillonnage

- **Paramètres de la moyenne d'échantillonnage**

Si on prélève un échantillon aléatoire de taille n, d'une population infinie (ou d'une population finie et échantillonnage avec remise) dont les éléments possèdent un caractère mesurable (réalisation d'une VA X) qui suit une loi de probabilité de moyenne E(X) = μ et de variance V(X) = σ^2, alors la moyenne d'échantillonnage $\bar{X}$ suit une loi de probabilité moyenne :

$$E(\bar{X}) = E(X) = \mu$$

et de variance :

$$V(\bar{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

A connaître car on s'en resservira tout le temps!!!

Moyenne d’échantillonnage

- Nous avons observé empiriquement l’évolution de l’espérance et de la variance de la moyenne d’échantillonnage
- On peut également redémontrer ce que vaut l’espérance de la moyenne d’échantillonnage et la variance de la moyenne d’échantillonnage

Démontrer que: $E(\bar{X}) = \mathrm{E}(\mathrm{X})$

Et $V(\bar{X}) = \frac{V(X)}{n}$

La 1ère démo est faite en séance

La 2nd est à faire à la maison

Moyenne d’échantillonnage

- **Forme de la distribution de la moyenne d’échantillonnage**

Pour caractériser complètement les fluctuations de la moyenne d’échantillonnage, il faut également être en mesure de préciser la forme probabiliste des fluctuations. Pour connaître exactement la distribution de $\bar{X}$, il faut connaître la distribution de la population qui a été échantillonnée ou alors utiliser le **théorème central limite.**

Loi suivie par $\bar{X}$ lorsque l’on considère que X suit une loi normale

$\bar{X}$ en tant que somme de n VA indépendantes toutes normales, suit une loi normale de moyenne E(X) et de variance V(X) ; alors la VA centrée réduite suit une loi normale centrée sur 0 et d’écart type 1.

Moyenne d’échantillonnage

Exercice :

Les billes de roulement fabriquées par une société ont une masse moyenne de 5.02 g avec un écart type de 0.3 g. On considère que la variable qui associe à une bille sa masse moyenne suit une loi normale.

Calculer la probabilité qu’un échantillon de 25 billes ait une masse moyenne supérieure à 5.2 g

Moyenne d'échantillonnage

<u>Convergence en loi de $\bar{X}$ lorsque la taille de l'échantillon n tend vers .... même si la loi suivie par X n'est pas connue.</u>

$\forall n \in N^*, X_1, X_2, \ldots, X_n$ sont des VA à valeurs réelles qui suivent toutes la même loi (celle de X) et elles sont indépendantes

cette loi admet pour espérance μ et pour variance σ^2

si on note: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ autrement dit si on prend la VA moyenne d'échantillonnage, alors

$$\bar{X}^* = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}$$

converge en loi vers N(0,1) (la loi normale centrée réduite)

On utilisera l'approximation dès que n≥30

Moyenne d’échantillonnage

Cas important de la fréquence d’échantillonnage :

Tous les individus ont pour mesure 0 ou 1. On note p = P (X = 1).

p la **proportion d’individus de mesure 1**. X est une VA de Bernouilli de paramètre p

D’où :

$$E(X) = p$$

$$V(X) = pq = p(1-p)$$

Ici la moyenne d’échantillonnage $\bar{X}$ est appelée <u>fréquence d’échantillonnage</u> et correspond à la fréquence de sortie du 1 dans le *n*-échantillon. On pourra également noter cette variable F.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = F$$

Moyenne d’échantillonnage

<u>Exercice :</u>

Quelle est la probabilité pour qu’en 120 lancers d’une pièce équilibrée la proportion de faces soit comprise entre 40% et 60% ?

Proposer deux solutions : approche probabiliste et approche statistique.

Approche probabiliste:

On étudie la variable de Bernouilli : X=1 si face de probabilité 0.5 et la variable binomiale associée: nombre de faces en 120 lancers. Et on étudie la loi de probabilité de la variable de Bernouilli

Approche statistique:

Sur la population on étudie la loi de Bernouille: X=1 si face de probabilité 0.5. Sur l’échantillon on étudie la moyenne d’échantillonnage (ou ici fréquence d’échantillonnage) ; et on approxime la loi fréquence d’échantillonnage centrée réduite à la loi normale centrée réduite.

Estimation

L'estimation des paramètres est l'objectif fondamental de l'échantillonnage d'une population.

Introduction

Un aspect important de l'inférence statistique est celui d'obtenir à partir de l'échantillonnage d'une population, des estimations fiables de certains paramètres de cette population. Dans ce cours, les paramètres que nous allons estimer sont **la moyenne μ** (ou **la proportion p**). Ces estimations peuvent s'exprimer soit par une seule valeur (estimation ponctuelle), soit par un intervalle (estimation par intervalle).

Estimation ponctuelle

Lorsqu'une caractéristique d'une population est estimée par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé une **estimation ponctuelle**.

Soient α un paramètre de la population et Y une statistique de n-échantillonnage.

$$\boxed{\alpha \Leftrightarrow \mathrm{E(Y)}=\alpha}$$

Y est un estimateur non biaisé de α

Si y est la valeur de Y sur un n-échantillon donné alors on dit que y est une estimation non baisée de α.

L'estimation ponctuelle se fait à l'aide d'un estimateur. Cet estimateur est fonction des observations de l'échantillon. L'estimation est la valeur numérique que prend l'estimateur selon les observations de l'échantillon.

Estimation ponctuelle

Comment choisir un estimateur? Il n'existe pas de « meilleur estimateur » mais il existe des critères de comparaison:

- Biais : on souhaite que l'estimation ne soit pas systématiquement décalée par rapport à la valeur vraie
- Précision : si l'on répète l'estimation sur un autre échantillon, on souhaite obtenir une estimation cohérente, donc peu de variation d'un échantillon à l'autre.
- Convergence : si l'on peut estimer la valeur du paramètre sur toute la population, la valeur de l'estimation obtenue doit être la valeur vraie du paramètre.
- Complexité : toute estimation nécessite un calcul donc un temps. On évaluera donc la complexité des calculs
- Robustesse : il existe souvent des sources de perturbations. On souhaite que l'estimation ne soit pas sensible à la présence de valeurs aberrantes.

Extrait du cours de M. TAYACHI-PIGEONNAT

Estimation ponctuelle

- Un estimateur est une variable aléatoire
- Une estimation est une valeur déterministe! C'est la réalisation d'un estimateur
- La variance d'un estimateur mesure sa variabilité. Si l'estimateur est sans biais, cette variabilité est autour de α . Si on veut estimer α correctement , il ne faut pas que cette variabilité soit trop forte
- En pratique, si on observe plusieurs jeux de données similaires, on obtient une estimation de α pour chacun d'entre eux. Alors si l'estimateur est de faible variance, ces estimations seront toutes proches les unes des autres, et s'il est sans biais leur moyenne sera très proche de α

Estimation ponctuelle

Que dire de la moyenne d’échantillonnage par rapport à la moyenne de la population?

Estimation par intervalle de confiance

Les estimations ponctuelles ne fournissent aucune information concernant la précision des estimations. Elles ne tiennent pas compte de l'erreur possible dans l'estimation, erreur attribuable aux fluctuations d'échantillonnage.

Quelle confiance avons-nous en une valeur unique ? On ne peut répondre à cette question en considérant uniquement l'estimation ponctuelle. Il faut lui associer un intervalle qui permet d'englober avec une certaine fiabilité, la vraie valeur du paramètre correspondant.

On va ici partir de plusieurs exemples et de ce que l'on connaît (c'est-à-dire de ce qui a été vu aux chapitres précédents) pour donner les intervalles de confiance de l'estimation d'une moyenne ou d'une proportion.

Estimation par intervalle de confiance

Exercice :

Une université compte plus de 5000 étudiants. On tire un échantillon de 100 étudiants. Sa masse moyenne est de 67.45 kg et sa variance de 8.5275.

1/ Quelle est la masse moyenne de l'ensemble des étudiants ? On a évidemment envie de dire que la masse moyenne de tous les étudiants est d'environ 67.45 kg. Quel est le raisonnement qui permet en toute rigueur de le dire ?

2/ On veut ici bien entendu préciser le « environ ». On va donc chercher le *y* tel que :

P(67.45 - *y* < masse moyenne de tous les étudiants < 67.45 + *y*) = 0.95

Calculer y

Remarque : On dit que [67.45 - *y* ; 67.45 + *y*] est l'intervalle de confiance pour la masse moyenne de tous les étudiants au niveau de confiance N_C = 95%.

On repart de ce que l’on connait c’est-à-dire de la loi centrée réduite:

Estimation par intervalle de confiance

Retour à l'exercice :

2/ On veut ici bien entendu préciser le « environ ». On va donc chercher le *y* tel que :

P(67.45 - *y* < masse moyenne de tous les étudiants < 67.45 + *y*) = 0.95. Calculer y

Remarque : On dit que [67.45 - *y* ; 67.45 + *y*] est l'intervalle de confiance pour la masse moyenne de tous les étudiants au niveau de confiance N_C = 95%.

Intervalle de confiance d'une moyenne

Population normale de variance connue ou grand échantillon (n≥30)

A partir d'un échantillon aléatoire de taille *n* d'une population normale de variance connue

σ^2 on définit, en prenant comme estimation ponctuelle de μ la moyenne de l'échantillon $\bar{x}$, un intervalle de confiance ayant un niveau de confiance N_c% de contenir la vraie valeur de comme suit :

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

où $z_{\frac{\alpha}{2}}$ est la valeur de la variable normale centrée réduite telle que la probabilité que Z soit compris entre $-z_{\frac{\alpha}{2}}$ et $z_{\frac{\alpha}{2}}$ est 1-α = N_c

Intervalle de confiance d'une moyenne

Population normale de variance inconnue ou grand échantillon (n≥30)
Dans le cas d'un grand échantillon ($n \geq 30$) provenant d'une population de variance inconnue mais estimée par la variance d'échantillon s^2 , alors l'intervalle de confiance

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Population normale de variance inconnue et n<30
A partir d'un échantillon aléatoire de petite taille ($n<30$), prélevé d'une population normale de moyenne m (inconnue) et de variance s^2 inconnue, alors on définit, en prenant comme estimation ponctuelle de m la moyenne $\bar{x}$ de l'échantillon, un intervalle de confiance ayant un niveau de confiance N_c% de contenir la valeur vraie de m comme suit :

$$\bar{x} - t_{\frac{\alpha}{2};\nu} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2};\nu} \frac{s}{\sqrt{n}}$$

Intervalle de confiance d’une moyenne

Exercice: Détermination de la taille de l’échantillon requise pour un essai de fiabilité d’un dispositif électronique

Une firme vient de développer un nouveau dispositif électronique qui entre dans la fabrication d’appareils de traitement de texte. Avant de mettre en production ce nouveau dispositif, on veut effectuer des essais préliminaires pour être en mesure d’estimer la fiabilité en terme de durée de vie. D’après le bureau d’étude de l’entreprise, l’écart type de la durée de vie de ce nouveau dispositif électronique serait de l’ordre de 100 heures.

Déterminer :

1/ Le nombre d’essais requis pour estimer, avec un niveau de confiance de 95%, la durée de vie moyenne d’une grande production de sorte que la marge d’erreur dans l’estimation n’excède pas 50 heures.

2/ Le nombre d’essais requis (pour le même niveau de confiance) pour estimer la durée de vie avec une marge d’erreur de 20 heures.

Intervalle de confiance d'une proportion

Intervalle de confiance d'un pourcentage (proportion)

A partir d'un échantillon aléatoire de taille n et en prenant comme estimation ponctuelle de *p* la fréquence observée *f* d'avoir un certain caractère qualitatif, on définit l'intervalle de confiance de *p* (la proportion d'éléments (individus) possédant un caractère qualitatif dans la population) avec un niveau de confiance N_c % :

$$f - z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

Cet intervalle de confiance est valable à condition que : $nf > 5$ et $n(1-f) > 5$

<u>Remarque:</u> on peut aussi trouver la notation $\widehat{p}$ au lieu de f

Intervalle de confiance d'une proportion

Une élection oppose deux candidats A et B. Un institut de sondage interroge 800 personnes sur leurs intentions de vote :

- 420 déclarent voter pour A

- 380 déclarent voter pour B

Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra A le jour de l'élection, en inférant sur l'ensemble de la population. L'estimation de p est la proportion:

$$f = \frac{420}{800} = 52{,}5\%$$

L'institut de sondage estime donc que le candidat A va gagner l'élection. Mais pour évaluer l'incertitude, on a besoin d'un intervalle de confiance de seuil 5% pour p. On obtient alors la réalisation suivante de l'intervalle de confiance: $[0{,}4904;\ 0{,}5596]$.

Conclusion : on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Intervalle de confiance d’une moyenne

- Ecrivez un programme Python qui retourne l’intervalle de confiance de la moyenne d’un échantillon ….