B2 - Intervalle de confiance d'une moyenne avec écart-type inconnu dans le cas d'une population Gaussienne

Dans le cas précédent, on a construit l'IdC à partir de la var $\dfrac{\overline{X}_n - m}{\sigma/\sqrt{n}}$. Mais, maintenant σ étant inconnu, il convient de le remplacer par son estimateur sans biais qui sera la racine carrée de la variance d'échantillon :

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}.$$

C'est-à-dire :

$$S_n = \frac{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}}{\sqrt{n-1}}.$$

On considère alors la var $Z_n = \dfrac{\overline{X}_n - m}{S_n/\sqrt{n}}$. La var X suivant une loi $G(m,\sigma)$, Z_n suit une loi de Student [1] à $n-1$ degrés de libertés (ddl en abrégé).
Rappelons la densité de cette loi : $\forall t \in \mathbb{R}$,

$$f_{Z_n}(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}},$$

où, $\forall m \in \mathbb{N}^*, \Gamma(m) = \displaystyle\int_0^{+\infty} x^{m-1}\mathrm{e}^{-x}\,\mathrm{d}x$.

L'expression, assez rébarbative, de la densité de cette loi ne doit pas vous inquiéter car elle figure dans tous les logiciels qui permettent d'effectuer des calculs statistiques (sur les tableurs notamment).

Donnons un exemple d'utilisation de cette loi pour déterminer un intervalle de confiance. Supposons que $X \sim G(m,\sigma)$ et qu'un échantillon (iid) de cette loi donne :

$$5.134; 6.582; 4.269; 6.425; 4.465; 4.400; 5.507; 5.212; 5.134; 4.852; 5.231; 3.933; 4.925$$
$$5.901; 3.166; 3.184; 3.915; 6.315; 4.678; 4.997$$

Déterminer un intervalle de confiance pour m, au seuil 0.99.

On commence par déterminer, grâce à un logiciel adapté (tableur par exemple) fournissant les valeurs inverses de la loi de Student à $20-1=19$ ddl, la valeur $t_{0.99}$ telle :

$$P[-t_{0.99} \leqslant \frac{\overline{X}_{20} - m}{S_{20}/\sqrt{20}} \leqslant t_{0.99}] = 0.99.$$

Puis, on détermine l'intervalle de confiance aléatoire de m au seuil 0.99 :

$$P\left[m \in [\overline{X}_{20} - \frac{t_{0.99}}{\sqrt{20}}S_{20}; \overline{X}_{20} + \frac{t_{0.99}}{\sqrt{20}}S_{20}]\right] = 0.99.$$

1. Student est le pseudonyme d'un statisticien anglais, de son vrai nom William GOSSET (1876-1937).

Enfin, à partir de l'échantillon, on calcule l'écart-type d'échantillon s_{20} et $\overline{x}_{20}$, pour trouver une réalisation de cet intervalle alétoire qui sera l'intervalle de confiance particulier de m au seuil 0.99, associé à l'échantillon de taille 20 donné par l'énoncé.
On trouve $[5.08; 5.39]$.

B3 - Intervalle de confiance d'une moyenne pour une population quelconque avec écart-type connu

La différence essentielle avec les deux cas précédents est que la construction de l'intervalle va s'appuyer sur une **loi limite** et non plus sur une loi exacte. Pour que l'intervalle de confiance soit assez précis, il faudra donc que la taille de l'échantillon soit « grande ». En pratique, $n \geqslant 50$ garantit une bonne précision.

Si X suit une loi (quelconque) d'espérance mathématique m, alors le théorème central-limite affirme que :

$$Z_n = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \text{ converge en loi vers } G(0;1), \text{ lorsque } n \text{ tend vers } +\infty.$$

Rappelons que la convergence en loi de Z_n vers la loi normale centrée réduite signifie que :

$$\forall x \in \mathbb{R}, \lim_{n\to+\infty} P[Z_n \leqslant x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t.$$

À partir de ce résultat on construit un intervalle de confiance aléatoire au seuil α comme précédemment :
On détermine t_α à partir de la loi normale centrée réduite et l'intervalle est de la forme :

$$[\overline{X}_n - \frac{t_\alpha}{\sqrt{n}}\sigma; \overline{X}_n + \frac{t_\alpha}{\sqrt{n}}\sigma].$$

Pour n grand, la probabilité que m appartienne à cet intervalle (aléatoire) est approximativement égale à α.
À partir d'un échantillon de taille n, on détermine un intervalle de confiance particulier en calculant $\overline{x}_n$ à partir de l'échantillon, l'idC est $[\overline{x}_n - \frac{t_\alpha}{\sqrt{n}}\sigma; \overline{x}_n + \frac{t_\alpha}{\sqrt{n}}\sigma]$.

B4 - Intervalle de confiance d'une moyenne pour une population quelconque avec écart-type inconnu

Si X est une var de moyenne m et d'écart-type σ, alors une généralisation du théorème central-limite, justifie que la var $Z_n = \frac{\overline{X}_n - m}{S_n/\sqrt{n}}$, tend encore en loi vers $G(0;1)$ (avec les notations introduites précédemment).
À partir de ce résultat, la construction de l'intervalle de confiance aléatoire au seuil α de m se construit exactement comme dans le cas B3. Pour un échantillon de taille n donné, on obtient un IdC particulier en remplaçant $\overline{X}_n$ par $\overline{x}_n$ et S_n par l'écart-type d'échantillon s_n.

Un cas particulier important

Il s'agit du cas où la var parente est un var de Bernouilli. C'est ce qui se produit lorsqu'on s'intéresse à la présence (ou l'absence) d'un caractère dans une population. Par exemple, avant une élection, on observe sur la population d'un certain ensemble géographique, le caractère « être favorable au candidat Untel ».
Une var de Bernouilli, X, est caractérisée par un paramètre, noté ici $p \in [0;1]$ et $E(X) = p$, $V(X) = p(1-p)$. Comme vous le comprenez, le but du travail statistique est l'estimation de p (estimation pontuelle ou par IdC). Comme p est la moyenne de la var parente, si $(X_i)_{1\leqslant i\leqslant n}$ est un échantillon iid de X, on est

ramené au cas B4 : var non nécesairement gaussienne, mais - grâce au théorème central-limite, loi de la moyenne de l'échantillon suivant une loi limite gaussienne. De plus, ni $m = E(X) = p$, ni $\sigma = p(1-p)$ ne sont connus.

Examinons un exemple.

On suppose que sur un échantillon de taille 100, le nombre de personnes favorables à Untel est égal à 62 (et le nombre de personnes non favorables alors égal à 100-62=38).

Estimation ponctuelle de p : on estime par l'estimateur habituel sans biais (et convergent) d'une moyenne $\overline{X}_{100}$. D'où l'estimation ponctuelle $\dfrac{62}{100} = 0.62$.

Estimation par IdC de p : la valeur de n étant suffisament grande, on s'appuie sur la loi limite de $\dfrac{\overline{X}_n - p}{S_n/\sqrt{n}}$ qui est la loi $G(0;1)$. D'où, pour un seuil α, comme on l'a déjà vu l'IdC aléatoire

$$[\overline{X}_{100} - \frac{t_\alpha}{10} S_{100}; \overline{X}_{100} - \frac{t_\alpha}{10} S_{100}].$$

et l'IdC particulier associé à l'échantillon observé :

$$[\overline{x}_{100} - \frac{t_\alpha}{10} s_{100}; \overline{x}_{100} - \frac{t_\alpha}{10} s_{100}].$$

Dans ce cas particulier, les expressions de $\overline{x}_n$ et de s_n s'obtiennent en codant les valeurs des var de Bernouilli par 1 (succès=« être favorable à Untel ») et 0 (échec = « ne pas être favorable à untel »). Ainsi, x_n est la fréquence observée des succès, cad ici, $\overline{x}_{100} = 0.62$. Le calcul de s_n donne : $s_n^2 = \dfrac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{n-1}$, où x_i vaut 1 un nombre de fois égal à $n\overline{x}_n$ et 0 un nombre de fois égal à $n - n\overline{x}_n = n(1-\overline{x}_n)$. On obtient :

$$\begin{aligned} s_n^2 &= \frac{(1-\overline{x}_n)^2 n\overline{x}_n + (0-\overline{x}_n)^2 n(1-\overline{x}_n)}{n-1} = \frac{(1-\overline{x}_n)^2 n\overline{x}_n + (0-\overline{x}_n)^2 n(1-\overline{x}_n)}{n} \times \frac{n}{n-1} \\ &= (1-\overline{x}_n)\,\overline{x}_n \times \frac{n}{n-1}. \end{aligned}$$

D'où, $s_n = \sqrt{(1-\overline{x}_n)\,\overline{x}_n}\sqrt{\dfrac{n}{n-1}}$.
L'expression de l'IdC particulier au seuil α est alors

$$\left[\overline{x}_n - \frac{t_\alpha}{\sqrt{n-1}}\sqrt{(1-\overline{x}_n)\,\overline{x}_n}; \overline{x}_n + \frac{t_\alpha}{\sqrt{n-1}}\sqrt{(1-\overline{x}_n)\,\overline{x}_n}\right]$$

Dans le cas particulier de l'exemple, si $\alpha = 0.95$:

$$\left[0.62 - \frac{1.96}{99}\sqrt{0.2356}; 0.62 + \frac{1.96}{99}\sqrt{0.2356}\right] = [0.573; 0.667].$$

Les chances de Untel d'être élu sont donc sérieuses.

La situation se complique si la fréquence observée des opinions favorables est voisine de 0.5.
Imaginons que $\overline{x}_n = 0.51$. L'amplitude de l'IdC associé à un échantillon (toujours au seuil 0.95) est alors égale à $2 \times 1.96 \times \dfrac{\sqrt{0.51 \times 0.49}}{\sqrt{n-1}}$, soit encore $\dfrac{1.96}{\sqrt{n-1}}$.

Si l'on veut que cette amplitude reste inférieure à 0.01 - ce qui fournira un IdC particulier « à droite » de 0.5 - il faut interroger un nombre n d'électeurs potentiels tel que $196^2 + 1 \leqslant n$, c'est-à-dire au minimum 38417 personnes - ce qui est un énorme échantillon ! Avec les réserves d'usage : c'est seulement pour 95% des échantillons, en moyenne, que l'IdC contiendra la vraie valeur de p.

Pour terminer ce paragraphe, il est intéressant d'examiner l'introduction à l'estimation d'une moyenne qui est proposée dans les programmes de Seconde.
Le détail des PO est le suivant.

CONTENUS	CAPACITÉS ATTENDUES	COMMENTAIRES
Échantillonage Notion d'échantillon. Intervalle de fluctuation d'une fréquence au seuil 0,95. Réalisation d'une simulation.	Concevoir, mettre en œuvre et exploiter des simulationsde situations concrètes à l'aide d'un tableur ou d'ue calculatrice. Exploiter et faire une analyse critique d'un résultat d'échantillonage.	Un échantillon de taille n est constitué de n répétitions indépendantes de la même expérience. À l'occasion de la mise en place d'une simulation on peut : – utiliser les fonctions logiques d'un tableur ou d'une calculatrice – mettre en place des instructions conditionnelles dans un algorithme. L'objectif est d'amener les élèves à un questionnement lors des activités suivantes. L'estimation d'une proportion inconnue à partir d'un échantillon, la prise de décision à patir d'un échantillon.

Le commentaire suivant est ajouté.

> « L'intervalle de fluctuation au seuil 0,95, relatif aux échantillons de taille n, est l'intervalle centré autour de p, où se situe - avec une probabilité égale à 0,95%, la fréquence observée dans l'échantillon de taille n. Cet intervalle peut être obtenu de façon approchée par simulation. Le professeur peut indiquer aux élèves le résultat suivant, utilisable dans la pratique pour des échantillons de taille $n \geqslant 25$ et des proportions p comprises entre 0,2 et 0,8 : si f désigne la fréquence du caractère dans l'échantillon, f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ avec une probabilité d'au moins 0,95. Le professeur peut faire percevoir expérimentalement la validité de cette propriété, mais **elle n'est pas exigible**. ».

Vous remarquerez que c'est l'expression « intervalle de fluctuation » qui est retenue. Elle correspond au cas où p est connue et où on cherche à confirmer (ou à infirmer) la valeur de p par celle d'une fréquence observée et d'un intervalle construit autour de cette fréquence.
Un intervalle de confiance est un intervalle qui a une probabilité donnée de contenir la valeur exacte de p **qui n'est pas supposée connue**.
La nuance entre les deux notions est faite dans certains manuels, mais elle n'est pas mentionnée dans le

PO.

Interrogez-vous sur la légitimité de l'approximation (de l'intervalle) qui est proposée par le PO dans l'encadré ci-dessus.

Exercices d'entrainement

Ce sont des exercices types des classes de STS - section TPIL (**T**echniques **P**hysiques pour l'**I**ndustrie et le **L**aboratoire) du groupement A (à l'exception du premier exercice).

Exercice 1 : On considère un stock très important de boulons. On note Y la var qui, à chaque boulon tiré au hasard dans le stock, associe le diamètre, en mm, de son pied.
La var Y suit la loi normale de moyenne inconnue m et d'écart-type $\sigma = 0,1$. On désigne $\overline{Y}$ la va qui, à chaque échantillon aléatoire de 100 boulons prélevé dans le stock, associe la moyenne des diamètres des pieds de ces 100 boulons (le stock est assez important pour qu'on puisse assimiler ces prélèvements à des tirages sans remise).

1) Justifier que $\overline{Y}$ suit une loi $G(m; 0,01)$.
2) Déterminer $h_1 > 0$ tel que $P[m - h_1 \leqslant \overline{Y} \leqslant m + h_1] = 0,9$.
3) Déterminer $h_2 > 0$ tel que $P[m - h_2 \leqslant \overline{Y} \leqslant m + h_2] = 0,95$.
4) Déterminer $h_3 > 0$ tel que $P[m - h_3 \leqslant \overline{Y} \leqslant m + h_3] = 0,99$.

Exercice 2 : On se propose d'étudier, dans une population de grand effectif, la taille d'adolescents de 13 à 14 ans. On suppose que la va X donnant la taille d'un adolescent est une va gaussienne de moyenne m et d'écart type σ.
Un échantillon de 36 adolescents, choisis au hasard dans la population étudiée, donne les résultats suivants.

Taille	[130 ;135[	[135 ;140[	[140 :145[	[145 ;150[	[150 ;155[	[155 ;160[	[160 ;165[
Effectif	1	4	7	10	8	4	2

1) a) Calculer la moyenne, notée $\overline{x}$ et l'écart-type σ_e de cet échantillon.
   b) En déduire une estimation ponctuelle de m et de σ.
   c) Donner un intervalle de confiance de m au seuil de 95%.
2) Afin d'améliorer la connaisssance de m, on décide d'augmenter la taille de l'échantillon. À partir de quel entier n_0 obtiendra-t-on un intervalle de confiance d'amplitude inférieure à 1 cm avec un seuil de 98% ?

Solutions :

Exercice 1 :

1) $\overline{Y}$ étant une moyenne d'un échantillon de var indépendantes d'une loi normale, suit aussi une loi (exacte) normale, de même moyenne que la loi parente et d'écart-type égal à $\sigma/\sqrt{n}$ si σ est l'écart-type de la loi parente et n la taille de l'échantillon.

2) $Z = \dfrac{\overline{Y} - m}{0.01}$ suit une loi $G(0;1)$, d'où

$$P\left[-\frac{h_1}{0,01} \leqslant Z \leqslant \frac{h_1}{0,01}\right] = 0,9 \Leftrightarrow 2F(\frac{h_1}{0,01}) - 1 = 0,9 \Leftrightarrow F(\frac{h_1}{0,01}) = 0,95$$

L'utilisation de la loi inverse d'une loi gausienne standard donne $h_1 = 0,0164$.

3) $h_2 = 0,0196$.

4) $h_3 = 0,0258$.

Exercice 2 :

1) a) Le type de calcul demandé ici pose un problème. On ne connait pas pour chaque individu de la population, la valeur de la taille. Il n'est donc pas possible d'effectuer les calculs demandés sans faire une hypothèse supplémentaire concernant la répartition des individus dans chaque classe. L'hypothèse la plus élémentaire consiste à assimiler chaque classe à son centre. Tout se passe alors comme si : 1 individu mesurait 132,5 cm ; 4 individus mesuraient 137,5 cm, etc.
D'autres hypothèse sont possibles - qui conduisent à d'autres valeurs de la moyenne et de l'écart-type - comme, par exemple, supposer que les individus sont équirépartis dans les classes. Alors, les 4 individus de la deuxième classe sont interprétés comme : 1 individu de taille 136 cm ; 1 individu de taille 137 cm, 1 individu de taille 138 cm et 1 individu de taille 139 cm.
Avec l'hypothèse simplificatrice, on trouve (en cm)

$$\overline{x} = \frac{1 \times 132.5 + 4 \times 137.5 + 7 \times 1452.5 + 10 \times 147.5 + 8 \times 152.5 + 4 \times 157.5 + 2 \times 162.5}{36} = 148.1$$

Et, pour l'écart-type d'échantillon $\sigma_e = 7.25$.

b) Les valeurs précédemment calculées fournissent des estimations ponctuelles des deux paramètres. Les estimateurs utilisés sont sans biais et convergents.

c) Pour déterminer un intervalle de confiance de m au seuil 95%, on utilise le fait que $\dfrac{X - m}{\sigma}$ suivant une loi gaussienne centrée réduite, $\dfrac{X - \overline{X}_{36}}{S_{36}/6}$ suit une loi de Student à 35 ddl. On a vu dans le cours que l'IdC était de la forme

$$[\overline{x}_{36} - t_{0.95}\frac{s_{36}}{6}; \overline{x}_{36} + t_{0.95}\frac{s_{36}}{6}]$$

où, $t_{0.95}$ est déterminé à partir de la loi de Student.
soit, $[148.1 - 2.03 \times \dfrac{7.25.}{6}; 148.1 + 2.03 \times \dfrac{7.25}{6}] = [145.6; 150.6]$.

2) L'amplitude de l'intervalle est $2 \times t_{0.98} \times \dfrac{7.25}{\sqrt{n}}$, où $t_{0.98} = 2.438$. Ce nombre est inférieur ou égal à 1 (en cm) dès que

$$2 \times 2.438 \times 7.25 \leqslant \sqrt{n} \Leftrightarrow n \geqslant 1250$$

Taille d'échantillon très supérieure à la taille de l'échantillon initial, mais fournissant un IdC particulier très étroit.