STATISTIQUES INFÉRENTIELLES
ESTIMATIONS

Table des matières

I) Estimation d'une fréquence

I.1. Le problème de l'estimation

Travail de l'élève 1 :
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 ... % ?
- Aurait-on gagné en fiabilité si l'on avait augmenté la taille de l'échantillon ?
- Proposer des problèmes concrets où l'on se pose le même genre de questions.

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...

I.2. Estimation d'une fréquence

I.2.a. Estimation ponctuelle

Propriété 1.

La fréquence observée f d'un caractère sur un échantillon de population, fournit assez naturellement une **estimation ponctuelle** de la fréquence p réelle de ce caractère sur la population entière.

Exemple :

Une usine produit des vis cruciformes.

On prélève un échantillon de 150 vis et on relève 3 pièces défectueuses.

On peut alors donner une estimation de la fréquence p de vis défectueuses dans la production journalière :

On a $f = \dfrac{3}{150} = 0,02$ donc, $p = 0,02$.

Remarques :

- Notons qu'il revient exactement au même d'estimer un pourcentage : dans l'exemple précédent, on peut affirmer que 2% des vis ont une croix mal formée sur la tête.
- Cette estimation très simple a le défaut de dépendre fortement de l'échantillon (et donc du hasard qui l'a amené). Elle ne contient aucune indication de qualité de l'estimation, en particulier de la taille n de l'échantillon utilisé... Or, on comprend bien que plus n est grand, meilleure est l'information, mais de quelle façon ? Pour cela, on va utiliser nos connaissances en matière d'échantillonnage, pour donner une "fourchette".

I.2.b. Intervalle de confiance

On se place dans le cas où : $n \geq 30$, $np \geq 5$ et $np(1-p) \geq 5$.

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 $\mathscr{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

A l'aide d'un échantillon, nous allons définir, avec un coefficient de confiance choisi à l'avance, un **intervalle de confiance** de la fréquence p des éléments de la population possédant une certaine propriété.

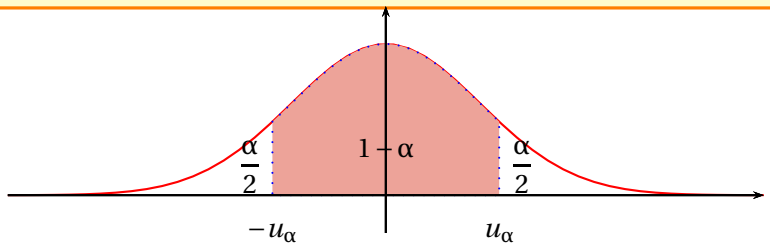
Propriété 2.

L'**intervalle de confiance** d'une fréquence p de la population avec le coefficient de confiance $1-\alpha$ est 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 le nombre t tel que $P(Z \leq t) = 1-\alpha$, avec Z suivant une loi normale centrée réduite.

Cet intervalle a pour centre la fréquence f de l'échantillon considéré.



Remarques :

- On obtient donc u_α en utilisant la calculatrice et la commande InvNorm$(1-\alpha/2)$

⇝ Les valeurs fréquentes du niveau de confiance sont $0,99$, $0,95$ et 0.90.
Pour ces trois valeurs, on obtient respectivement $t \simeq 2,575$, $t \simeq 1,96$ et $t \simeq 1.64$.

⇝ "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 ou 0.99 que cet intervalle contienne le paramètre inconnu f,

— après le tirage, le paramètre p est dans l'intervalle calculé avec une confiance 95% ou 99% ."

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, 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

On sait donc que la variable aléatoire $F = \dfrac{X}{n}$ suit une loi proche de $\mathcal{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.
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 transport. On veut déterminer, au seuil de risque 1%, un intervalle de confiance du pourcentage p de personnes mécontentes dans la commune :

On a : $f = 0,42$; $n = 500$; $\alpha = 1\%$ donc $u_\alpha = InvNorm(1-0,01/2) \simeq 2,575$.

Un intervalle de confiance du pourcentage p est donc :

$$\left[0,42 - 2,575\sqrt{\frac{0,42\times 0,58}{500}}; 0,42 + 2,575\sqrt{\frac{0,42\times 0,58}{500}}\right] = [0,36; 0,48] = [36\%; 47\%].$$

Exemple :

Dans l'exemple de la bouteille, 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 90% associé est :

$$\left[0.59 - 1.64\sqrt{\frac{0.59\times 0.41}{300}}\ ;\ 0.59 + 1.64\sqrt{\frac{0.59\times 0.41}{300}}\right] \simeq [0.54; 0.64]$$

soit environ entre 8 et 12 perles noires. L'intervalle de confiance à 95% associé est :

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

L'intervalle de confiance à 99% associé est :

$$\left[0.59 - 2.58\sqrt{\frac{0.59\times 0.41}{300}}\ ;\ 0.59 + 2.58\sqrt{\frac{0.59\times 0.41}{300}}\right] \simeq [0.51; 0.67]$$

soit environ entre 7 et 13 perles noires.

En réalité, il faut arrondir par défaut et par excès en fonction des bornes inf et sup ...

Exercice 1 : **Objectifs sur la base d'un énoncé d'examen :**

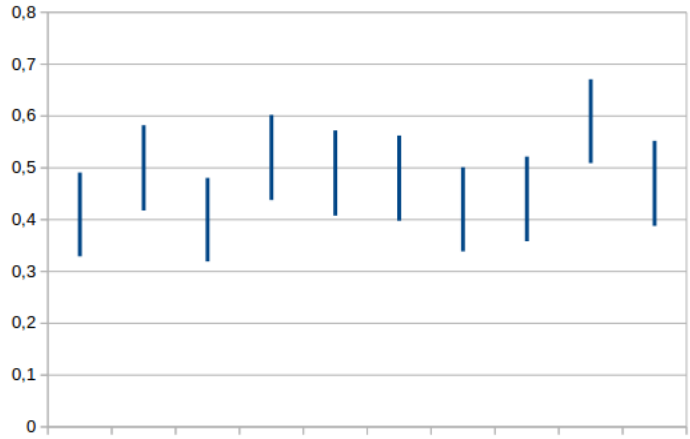
- ⇝ Déterminer un intervalle de confiance pour la moyenne de la population et étudier l'impact du coefficient de confiance et de la taille de l'échantillon.
- ⇝ Expérimenter la dépendance des intervalles de confiance à l'échantillon choisi.

Travaillons sur un exemple 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%

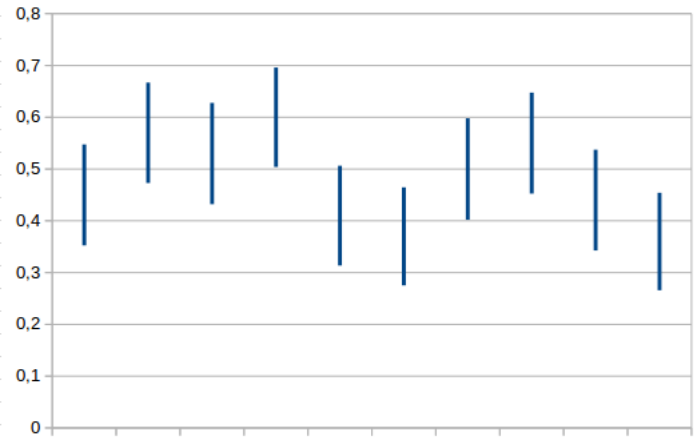
Ouvrir la feuille de calculs et expliquer sa construction.
Sur un exemple de 10 sondages,

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 ?

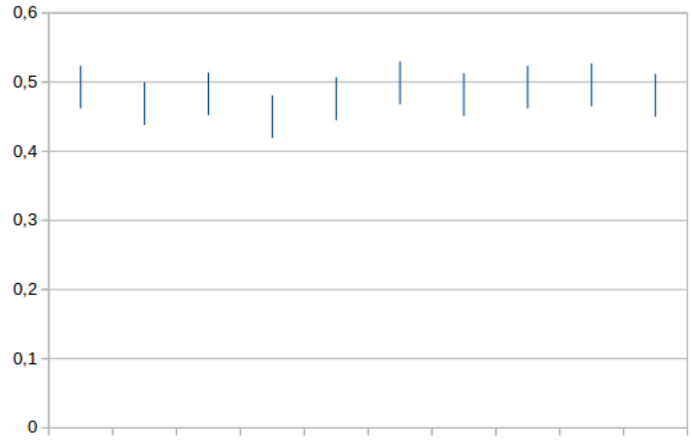


Les intervalles de confiance à 95% se "trompent" moins souvent (statistiquement, on peut observer qu'environ 95% contiennent la valeur p) parce qu'ils sont plus "longs".

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) Estimation d'une moyenne et d'un écart-type

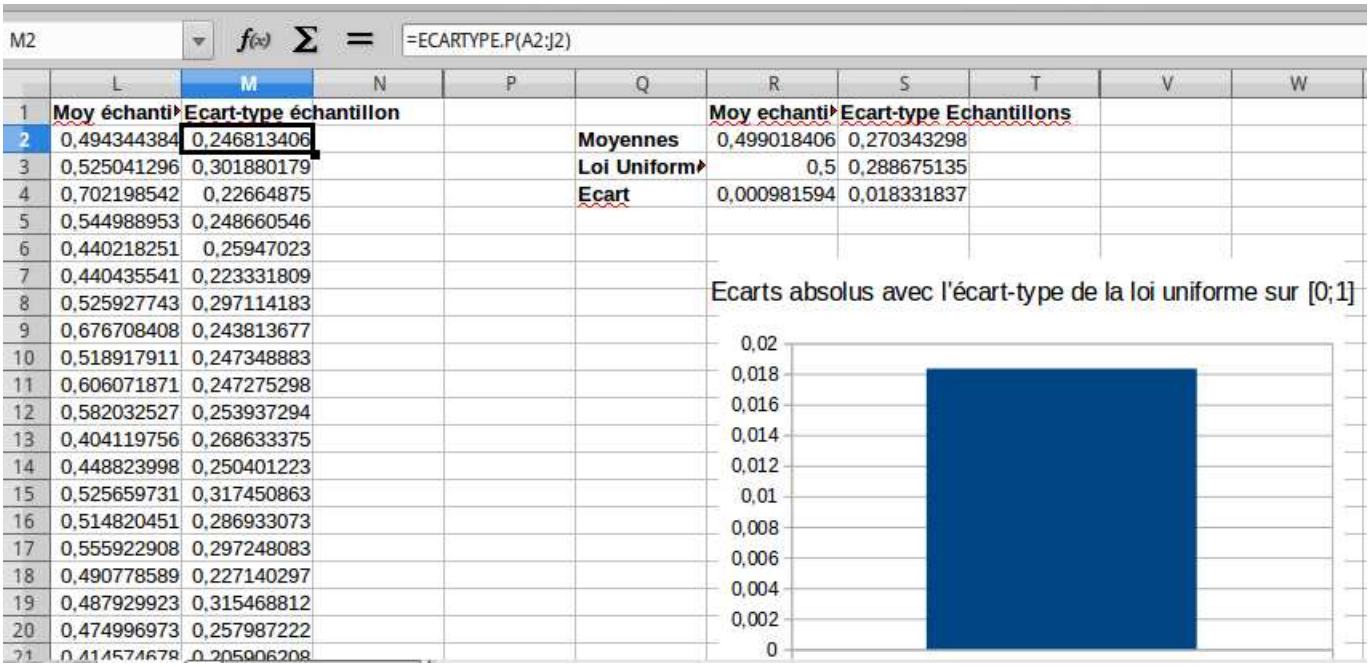
II.1. Simulation

On prélève, au hasard et avec remise, un échantillon de taille n dans une population. On calcule la moyenne $\overline{x}$ et l'écart type s_n de cet échantillon.
Il s'agit d'estimer la moyenne μ et l'écart type σ , inconnus, de la population.

Dans une feuille de tableur, on a simulé 1000 échantillons de taille 10 de choix d'un nombre aléatoire entre 0 et 1, grâce à la formule = ALEA() Autrement dit, on a simulé la loi uniforme sur $[0,1]$.
Pour chaque échantillon, on calcule sa moyenne $\overline{x}$, grâce à la formule = MOYENNE(A2 : A1001), et son écart-type s_n, grâce à la formule = ECART.TYPE.P(A2 : A1001)

Remarque : Le P précise que l'on considère l'échantillon comme la population entière, le tableur ne tient donc pas compte de la correction de l'estimation, contrairement à la formule = ECART.TYPE() qui utilise l'estimation donnée ci-après.

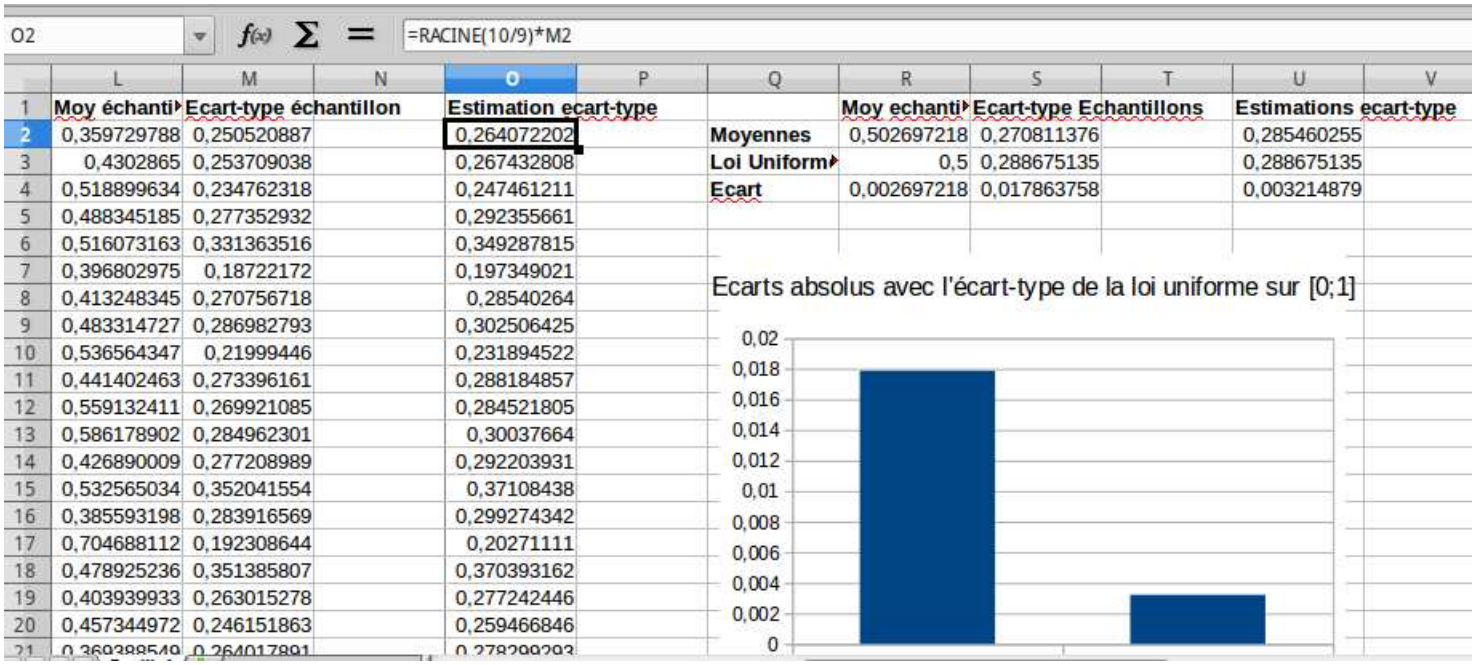
Ensuite, on compare les moyennes de chacun de ses résultats à la théorie. On constate de suite que l'estimation de la moyenne est excellente, alors que celle de l'écart-type laisse à désirer ...



Moy échanti	Ecart-type échantillon
0,494344384	0,246813406
0,525041296	0,301880179
0,702198542	0,22664875
0,544988953	0,248660546
0,440218251	0,25947023
0,440435541	0,223331809
0,525927743	0,297114183
0,676708408	0,243813677
0,518917911	0,247348883
0,606071871	0,247275298
0,582032527	0,253937294
0,404119756	0,268633375
0,448823998	0,250401223
0,525659731	0,317450863
0,514820451	0,286933073
0,555922908	0,297248083
0,490778589	0,227140297
0,487929923	0,315468812
0,474996973	0,257987222
0,414574678	0,205906208

	Moy echanti	Ecart-type Echantillons
Moyennes	0,499018406	0,270343298
Loi Uniform	0,5	0,288675135
Ecart	0,000981594	0,018331837

On affiche alors les colonnes cachées qui proposent une nouvelle estimation pour l'écart-type. On constate de suite qu'elle semble bien meilleure.



Moy échanti	Ecart-type échantillon	Estimation ecart-type
0,359729788	0,250520887	0,264072202
0,4302865	0,253709038	0,267432808
0,518899634	0,234762318	0,247461211
0,488345185	0,277352932	0,292355661
0,516073163	0,331363516	0,349287815
0,396802975	0,18722172	0,197349021
0,413248345	0,270756718	0,28540264
0,483314727	0,286982793	0,302506425
0,536564347	0,21999446	0,231894522
0,441402463	0,273396161	0,288184857
0,559132411	0,269921085	0,284521805
0,586178902	0,284962301	0,30037664
0,426890009	0,277208989	0,292203931
0,532565034	0,352041554	0,37108438
0,385593198	0,283916569	0,299274342
0,704688112	0,192308644	0,20271111
0,478925236	0,351385807	0,370393162
0,403939933	0,263015278	0,277242446
0,457344972	0,246151863	0,259466846
0,369388549	0,264017891	0,278299293

	Moy echanti	Ecart-type Echantillons	Estimations ecart-type
Moyennes	0,502697218	0,270811376	0,285460255
Loi Uniform	0,5	0,288675135	0,288675135
Ecart	0,002697218	0,017863758	0,003214879

II.2. Estimation ponctuelle

Propriété 3.

La valeur moyenne observée $\overline{x}$ d'un caractère sur un échantillon de population, fournit assez naturellement une **estimation ponctuelle** de la moyenne μ réelle de ce caractère sur la population entière.
L'écart-type observé s_n d'un caractère observé sur un échantillon de population fournit une estimation *faussée* de l'écart-type réel σ de ce caractère dans la population entière.
Une meilleure estimation de σ est le nombre $\sqrt{\dfrac{n}{n-1}}s_n$ où n est la taille de l'échantillon servant au calcul de s_n.

Exemple :

Une usine produit des vis cruciformes. On souhaite estimer la moyenne des longueurs des vis dans la production de la journée qui s'élève à 10000 pièces.
On choisit un échantillon de 150 vis et on obtient une moyenne de $\overline{x} = 4,57$ cm.
On en déduit donc que la longueur moyenne des vis de la production journalière est $\mu = 4,57$ cm.

La mesure de la longueur des vis produites dans l'échantillon précédent de 150 pièces conduit à relever un écart-type de 3 mm.
La meilleure estimation possible de l'écart-type de la production journalière n'est pas de 3 mm comme dans le cas précédent pour la moyenne, mais de $\sigma = 3\sqrt{\dfrac{150}{149}} \simeq 3,01$ mm.

Remarques :

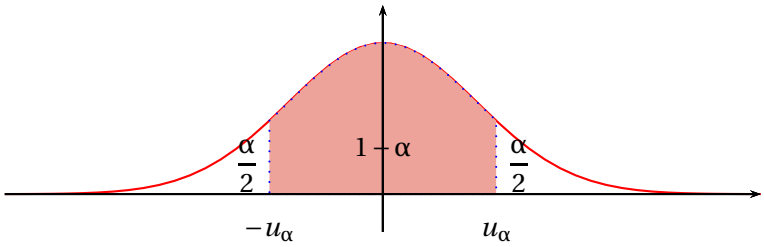
- L'estimation de σ peut étonner. Elle est ainsi faite pour que, sur un grand nombre d'échantillons, la moyenne des estimations soit égale à σ. En effet, s_n a tendance à être généralement inférieur à σ (biais). *Mais ceci se démontre !*
- La correction de s_n devient cependant assez rapidement minime lorsque la taille de l'échantillon augmente car
$$\lim_{n\to\infty} \sqrt{\frac{n}{n-1}} = 1$$
La correction est ainsi de l'ordre de 0,5% pour des échantillons de taille 100, et de l'ordre de 0,05% pour des échantillons de taille 1000.

II.3. Intervalle de confiance

On considère la variable aléatoire $\overline{\mathrm{X}}$ qui, à tout échantillon de taille n, associe sa moyenne. On sait que, au moins pour n assez grand, $\overline{\mathrm{X}}$ suit (approximativement) la loi normale $\mathrm{N}\left(\mu; \dfrac{\sigma}{\sqrt{n}}\right)$.
Ceci nous permet d'en déduire la proposition suivante :

Propriété 4.

L'intervalle de confiance de la moyenne m de la population avec le coefficient de confiance $1-\alpha$ est
$$\left[\overline{\mathrm{X}} - u_\alpha \frac{\sigma}{\sqrt{n}}; \overline{\mathrm{X}} + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$



Preuve

En effet, on sait que $P(\mu - h \leq \overline{X} \leq \mu + h) = 0,95$ pour $h = 1.96\dfrac{\sigma}{\sqrt{n}}$

Ce qui équivaut à $P(\overline{X} - h \leq \mu \leq \overline{X} + h) = 0,95$ (intervalle dont les bornes sont aléatoires et qui contient μ avec une probabilité de 0,95).

De façon générale, on proposera comme intervalle de confiance de la moyenne μ de la population totale, au coefficient de confiance de $1-\alpha\%$, l'intervalle :

$$\left[\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}} \,;\, \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

calculé à partir de la moyenne $\overline{x}$ observée sur un échantillon de taille n, où u_α est donné par la calculatrice en utilisant la commande InvNorm pour loi normale $N(0,1)$ et de paramètre $1 - \dfrac{\alpha}{2}$.

Remarques :

- Lorsque σ est connu et que n est grand, la situation est plus simple que pour les fréquences.
- Ce qui précède est valable pour les grands échantillons ($n \geq 30$) ou les petits échantillons issus d'une population normale, lorsque σ est connu.
- Lorsque σ est inconnu et $n \geq 30$, on utilise la formule précédente, en remplaçant σ par son estimation ponctuelle $s = \sqrt{\dfrac{n}{n-1}}\, s_n$

Exemple :

On suppose que la durée de vie, exprimée en heures, d'une ampoule électrique d'un certain type, suit la loi normale de moyenne M inconnue et d'écart-type $\sigma = 20$.

Une étude sur un échantillon de 16 ampoules donne une moyenne de vie égale à 3000 heures.

On va déterminer un intervalle de confiance de M au seuil de risque de 10%.

On a : $\alpha = 10\%$ d'où $1 - \alpha = 0,90 \iff t = 1,645$.

Un intervalle de confiance de M est donc : $\left[3000 - 1,645\dfrac{20}{\sqrt{16}}; 3000 + 1,645\dfrac{20}{\sqrt{16}}\right] = [2992, 3008]$.

III) Tableau récapitulatif

Le tableau ci-dessous regroupe toutes les situations dans lesquelles on doit savoir fournir une estimation ponctuelle ou par intervalle de confiance :

Paramètre de la population totale à estimer	Valeur du paramètre dans l'échantillon de taille n	Estimation ponctuelle pour la population totale	Estimation par intervalle de confiance au niveau de confiance $1-\alpha$ pour la population totale
Fréquence	f	$p=f$	$\left[f-u_\alpha\sqrt{\dfrac{f(1-f)}{n}};f+u_\alpha\sqrt{\dfrac{f(1-f)}{n}}\right]$
Moyenne	$\overline{x}$	$\mu=\overline{x}$	$\left[\overline{x}-u_\alpha\dfrac{\sigma}{\sqrt{n}};\overline{x}+u_\alpha\dfrac{\sigma}{\sqrt{n}}\right]$
Écart-type	s_n	$\sigma=s_n\sqrt{\dfrac{n}{n-1}}$	