II - Estimation d'un paramètre par intervalle de confiance

1°) - Généralités sur la construction

On veut estimer un paramètre (moyenne, proportion…) d'un caractère dans une population $\mathcal{P}$. Une estimation ponctuelle à partir d'un échantillon donné ne renseigne pas beaucoup sur le degré d'approximation du paramètre.

On détermine des réels k_1 et k_2 (dépendant éventuellement du paramètre que l'on cherche à estimer) tels que $P[k_1 \leq Z \leq k_2] = 1 - \alpha$ où Z est une variable aléatoire dont on connaît la loi d'après la théorie de l'échantillonnage et α la proportion d'échantillons donnant une valeur observée de Z jugée comme peu probable, α est le **seuil de confiance** et $1 - \alpha$, le **niveau de confiance** (en général α vaut 5 %, 10 % ou 1 %).

Un échantillon donne une valeur observée z_{obs} de Z, on fait confiance au hasard à $1 - \alpha$ et on suppose que la valeur du paramètre cherchée fait en sorte que $k_1 \leq z_{obs} \leq k_2$. Les valeurs du paramètre pour lesquelles cette inégalité est vraie constituent un **intervalle de confiance du paramètre au seuil de confiance** α.

Plus on fait confiance au hasard, plus α est petit (à la limite si on faisait totalement confiance au hasard, on devrait prendre $\alpha = 0$ et l'intervalle de confiance du paramètre serait $\mathbb{R}$).

2°) - Estimation d'une proportion par intervalle de confiance

a) - Problème

On veut estimer la proportion p d'individus ayant une certaine propriété dans une population $\mathcal{P}$, à partir de l'observation de la fréquence $f_é$ de la propriété dans un échantillon $\mathcal{E}$ de taille n.

b) - Détermination

Soit α un seuil de confiance. Soit F la variable aléatoire qui, à chaque échantillon de taille n, associe la proportion des individus ayant la propriété étudiée. Examinons ce qui se passe dans les trois cas suivants.

❶ **$n \geq 50$: la loi de $Y = \dfrac{F - p}{\sqrt{\dfrac{F\,(1 - F)}{n}}}$ est approchée par la loi normale $\mathcal{N}(0\,;1)$.**

La loi de Y ne dépend pas de p, elle permet de déterminer le réel u_α tel que $P(-u_\alpha \leq Y \leq u_\alpha) = 1 - \alpha$ ($k_1 = -u_\alpha$ et $k_2 = u_\alpha$), alors $P(|Y| \geq u_\alpha) = \alpha$ et $\Phi(u_\alpha) = 1 - \dfrac{\alpha}{2}$ où Φ est la fonction de répartition de la loi normale centrée réduite.

> En répartissant le risque de façon symétrique, la longueur de l'intervalle $[k_1\,;k_2]$ est minimale car la loi normale centrée réduite est unimodale et distribuée de façon symétrique par rapport à 0. Dans le cas où la variable aléatoire dont on connaît la loi ne remplit pas ces conditions, on répartit quand même le risque de façon symétrique mais cela n'a rien d'optimal.

On fait confiance au hasard à $1 - \alpha$ et on admet que la valeur observée $f_é$ de F à partir de l'échantillon $\mathcal{E}$ vérifie l'inégalité $-u_\alpha \leq \dfrac{f_é - p}{\sqrt{\dfrac{f_é\,(1 - f_é)}{n}}} \leq u_\alpha$, on obtient $f_é - u_\alpha\sqrt{\dfrac{f_é\,(1 - f_é)}{n}} \leq p \leq f_é + u_\alpha\sqrt{\dfrac{f_é\,(1 - f_é)}{n}}$

Un intervalle de confiance de p au niveau de confiance $1-\alpha$ est $\left[f_é - u_\alpha\sqrt{\dfrac{f_é\,(1-f_é)}{n}}\,;\, f_é + u_\alpha\sqrt{\dfrac{f_é\,(1-f_é)}{n}}\right]$.

L'intervalle aléatoire $\left[F - u_\alpha\sqrt{\dfrac{F\,(1-F)}{n}}\,;\, F + u_\alpha\sqrt{\dfrac{F\,(1-F)}{n}}\right]$ est un intervalle de confiance de p au seuil de confiance α.

Remarque :

Si α diminue, u_α augmente et l'amplitude de l'intervalle augmente.

Exercice 3

On considère la population d'une grande ville. On veut estimer la proportion de personnes de la ville nées en janvier et pour cela, on prélève dans cette population un échantillon de 400 personnes dans lequel on constate que 32 personnes sont nées au mois de janvier.

Donner une estimation de la proportion de personnes de la ville nées en janvier par intervalle de confiance au niveau de confiance de 95 %.

❷ $n \geq 30$, $np \geq 15$ et $np(1-p) > 5$: la loi de $X = \dfrac{F-p}{\sqrt{\dfrac{p\,(1-p)}{n}}}$ est approchée par la loi normale $\mathcal{N}(0\,;\,1)$.

On détermine le réel u_α tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1-\alpha$.

On fait confiance au hasard à $1-\alpha$ et on admet que $f_é$ vérifie l'inégalité $-u_\alpha \leq \dfrac{f_é - p}{\sqrt{\dfrac{p\,(1-p)}{n}}} \leq u_\alpha$, en résolvant cette double inéquation du second degré en p,on obtient :

$$\frac{f_é + \dfrac{u_\alpha^{\,2}}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{f_é\,(1-f_é)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}} \leq p \leq \frac{f_é + \dfrac{u_\alpha^{\,2}}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{f_é\,(1-f_é)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}$$

Un intervalle de confiance de la proportion p au niveau de confiance $1-\alpha$ est

$$\left[\frac{f_é + \dfrac{u_\alpha^{\,2}}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{f_é\,(1-f_é)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\,;\,\frac{f_é + \dfrac{u_\alpha^{\,2}}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{f_é\,(1-f_é)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\right].$$

$\left[\dfrac{F + \dfrac{u_\alpha^{\,2}}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{F\,(1-F)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\,;\,\dfrac{F + \dfrac{u_\alpha^{\,2}}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{F\,(1-F)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\right]$ est un intervalle de confiance de p au seuil de confiance α.

❸ Dans tous les cas, la loi de nF est la loi binomiale $\mathcal{B}(n\,;\,p)$.

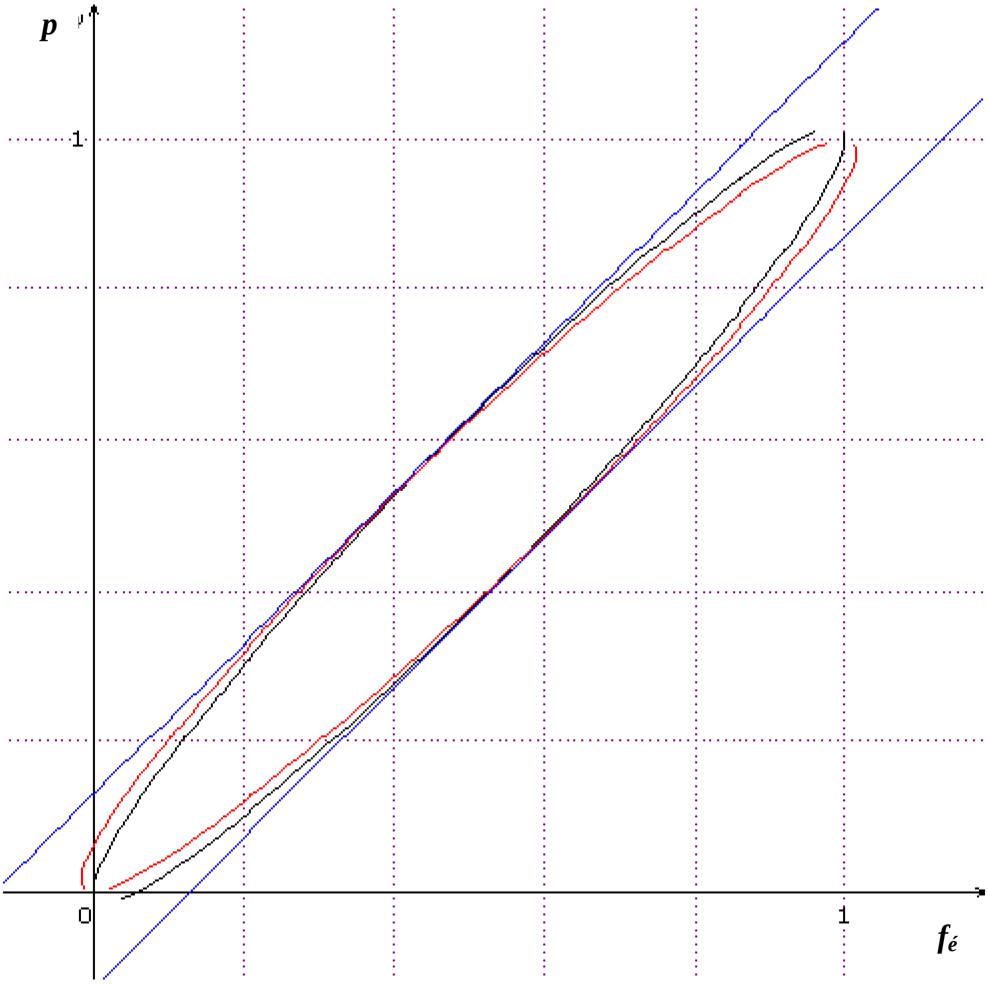
La loi de nF dépend de p et est discrète, k_1 et k_2 dépendent de p.

On prend pour $k_1(p)$ le plus grand entier tel que $P\left[nF \geq k_1(p)\right] \geq 1 - \frac{\alpha}{2}$ et pour $k_2(p)$ le plus petit entier tel que $P\left[nF \leq k_2(p)\right] \geq 1 - \frac{\alpha}{2}$. Ainsi $P\left[k_1(p) \leq nF \leq k_2(p)\right] \geq 1 - \alpha$.

On tire un échantillon au hasard, on admet que l'effectif observé $n f_é$ du caractère dans l'échantillon est compris entre $k_1(p)$ et $k_2(p)$. On utilise alors une table de la fonction de répartition des lois binomiales de premier paramètre n pour déterminer un intervalle contenant les valeurs de p faisant en sorte que $k_1(p) < n f_é < k_2(p)$. L'intervalle ainsi constitué est un intervalle de confiance de p au seuil de confiance α.

Pour éviter ce travail fastidieux, on utilise des abaques de loi binomiale. Un abaque est un réseau de courbes en coordonnées cartésiennes $(f\,, p)$. Chaque courbe correspond à une taille d'échantillon et donne les bornes de l'intervalle de confiance en fonction de l'observation $f_é$ de F dans l'échantillon $\mathcal{E}$.

Intervalles de fluctuation obtenus trois méthodes pour une proportion



- En rouge avec $\boldsymbol{Y = \dfrac{F - p}{\sqrt{\dfrac{p\,(1-p)}{n}}}}$

  Pour une valeur de *p*, l'intervalle de fluctuation de la fréquence d'échantillonnage *F* au niveau de probabilité de 95%, est l'intervalle $\left[p - 1{,}96\sqrt{\dfrac{p\,(1-p)}{n}}\;;\; p + 1{,}96\sqrt{\dfrac{p\,(1-p)}{n}}\right]$

- En noir avec $\boldsymbol{Y = \dfrac{F - p}{\sqrt{\dfrac{F\,(1-F)}{n}}}}$

  Pour une valeur de *p*, l'intervalle de fluctuation de la fréquence d'échantillonnage *F* au niveau de probabilité de 95%, est l'intervalle $\left[\dfrac{p + \dfrac{u_\alpha^{\,2}}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{p\,(1-p)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\;;\; \dfrac{p + \dfrac{u_\alpha^{\,2}}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^{\,2}}{4n^2} + \dfrac{p\,(1-p)}{n}}}{1 + \dfrac{u_\alpha^{\,2}}{n}}\right]$

- En bleu avec l'approximation du programme de seconde

  Pour une valeur de *p*, l'intervalle de fluctuation de la fréquence d'échantillonnage *F* au niveau de probabilité de 95%, est l'intervalle $\left[p - \dfrac{1}{\sqrt{n}}\;;\; p + \dfrac{1}{\sqrt{n}}\right]$.

3°) - Estimation d'une moyenne par intervalle de confiance

a) - Problème

On veut estimer la moyenne μ d'un caractère quantitatif dans une population $\mathcal{P}$ à partir de l'observation d'un échantillon $\mathcal{E}$ de taille n. Soit σ l'écart-type du caractère dans la population.

b) - Détermination

Soit α un seuil de confiance. Soit $\overline{X}$ et $\hat{S}$ les variables aléatoires qui, à chaque échantillon de taille n, associent respectivement la moyenne du caractère étudié et son écart-type corrigé.

❶ **σ est connu, la loi de $U = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ est**

$$\begin{cases} \bullet \text{ la loi normale } \mathcal{N}(0\,;\,1) \text{ si le caractère est distribué normalement dans la population,} \\ \bullet \text{ approchée par la loi normale } \mathcal{N}(0\,;\,1) \text{ si } n \geq 30 \end{cases}$$

La loi de U donnée par la théorie de l'échantillonnage ne dépend pas de μ, elle permet de déterminer le réel u_α tel que $P(-u_\alpha \leq U \leq u_\alpha) = 1 - \alpha$ $(k_1 = -u_\alpha$ et $k_2 = u_\alpha)$, alors $P(|U| \geq u_\alpha) = \alpha$ et $\Phi(u_\alpha) = 1 - \dfrac{\alpha}{2}$ où Φ est la fonction de répartition de la loi normale centrée réduite.

Si on fait confiance au hasard à $1 - \alpha$, on peut supposer que la valeur observée $m_é$ de $\overline{X}$ dans l'échantillon vérifie l'inégalité $-u_\alpha \leq \dfrac{m_é - \mu}{\dfrac{\sigma}{\sqrt{n}}} \leq u_\alpha$, on obtient $m_é - u_\alpha \dfrac{\sigma}{\sqrt{n}} \leq \mu \leq m_é + u_\alpha \dfrac{\sigma}{\sqrt{n}}$

Un intervalle de confiance de la moyenne μ au seuil de confiance α est $\left[m_é - u_\alpha \dfrac{\sigma}{\sqrt{n}}\,;\, m_é + u_\alpha \dfrac{\sigma}{\sqrt{n}} \right]$.

L'intervalle aléatoire $\left[\overline{X} - u_\alpha \dfrac{\sigma}{\sqrt{n}}\,;\, \overline{X} + u_\alpha \dfrac{\sigma}{\sqrt{n}} \right]$ est un **intervalle de confiance** de μ au seuil de confiance α.

❷ σ est inconnu et on connaît la loi de $T = \dfrac{\overline{X} - \mu}{\dfrac{\widehat{S}}{\sqrt{n}}}$

La loi de T donnée par la théorie de l'échantillonnage ne dépend pas de μ, elle permet de déterminer le réel t_α tel que $P(-t_\alpha \leq T \leq t_\alpha) = 1 - \alpha$ ($k_1 = -t_\alpha$ et $k_2 = t_\alpha$), alors $P(|T| \geq t_\alpha) = \alpha$ et $\Phi(t_\alpha) = 1 - \dfrac{\alpha}{2}$.

On fait confiance au hasard à $1 - \alpha$ et on admet que les valeurs observées $m_é$ et $\widehat{s_é}$ dans l'échantillon de $\overline{X}$ et $\widehat{S}$ respectivement, vérifient l'inégalité $-t_\alpha \leq \dfrac{m_é - \mu}{\dfrac{\widehat{s_é}}{\sqrt{n}}} \leq t_\alpha$, on obtient $m_é - t_\alpha \dfrac{\widehat{s_é}}{\sqrt{n}} \leq \mu \leq m_é + t_\alpha \dfrac{\widehat{s_é}}{\sqrt{n}}$.

Un intervalle de confiance de la moyenne μ au seuil de confiance α est $\left[m_é - t_\alpha \dfrac{\widehat{s_é}}{\sqrt{n}} \,;\, m_é + t_\alpha \dfrac{\widehat{s_é}}{\sqrt{n}} \right]$.

L'intervalle aléatoire $\left[\overline{X} - t_\alpha \dfrac{\widehat{S}}{\sqrt{n}} \,;\, \overline{X} + t_\alpha \dfrac{\widehat{S}}{\sqrt{n}} \right]$ est un intervalle de confiance de μ au seuil de confiance α.

Exercice 4

Un grossiste achète un lot de plusieurs milliers de poulets à une coopérative agricole. Il voudrait un encadrement de la masse moyenne μ des poulets dont il serait "sûr" à 90 % (le niveau de confiance est 0,9). Pour cela, il prélève au hasard 60 poulets du lot. La moyenne de l'échantillon est $m_é = 1{,}5$ kg et son écart-type est $s_é = 0{,}2$ kg.

1°) - Donner $\widehat{s_é}$ l'écart-type corrigé de l'échantillon.

2°) - Donner une estimation de μ par un intervalle de confiance au seuil de 10 %.

Exercice 5

Un centre de transfusion sanguine désire connaître, à 0,05 près, la proportion p de personnes du groupe sanguin O (donneurs universels) dans sa zone d'action et cela au niveau de confiance de 99 %.

Déterminer la taille de l'échantillon à prélever dans cette population pour satisfaire cette demande.

Exercice 6

Une semaine avant des élections, un institut de sondage a interrogé, au hasard, n personnes (n est de l'ordre de plusieurs centaines) sur leurs intentions de vote. L'institut donne au niveau de confiance de 95 %, l'intervalle de confiance [34,72 % ; 43,37 %] pour le pourcentage d'électeurs favorables au candidat Martin.

Déterminer la taille n de l'échantillon interrogé par l'institut de sondage.

Exercice 7

On veut estimer le nombre N d'oiseaux d'une certaine espèce dans une région. Pour cela, on en capture 90 que l'on bague, puis que l'on relâche. On cherche ensuite à estimer le pourcentage p d'oiseaux bagués dans la population :

- quelque temps après, on capture 110 oiseaux ;
- après chaque capture, on observe si l'animal est bagué ou non, puis on le relâche (tirage avec remise) ;
- le nombre d'oiseaux bagués ainsi observés est 17.

❶ Déterminer une estimation ponctuelle de p.

❷ Déterminer un intervalle de confiance de p au niveau de 95 %.

❸ Utiliser les résultats précédents pour déterminer un encadrement de N.

Exercice 8

Un échantillon de 12 mesures de la résistance de rupture de certains fils de coton a pour moyenne 7,38 kg et pour écart-type 1,24 kg. On suppose que les mesures de la résistance sont réparties selon une loi normale. Déterminer un intervalle de confiance de la résistance moyenne de rupture au seuil de confiance de **a)** 5 % **b)** 1 %.

Exercice 9

Un laboratoire vérifie la résistance à l'éclatement (en kg/cm^2) des réservoirs d'essence d'un fabricant. Des essais similaires, réalisés il y a un an, permettent de considérer que la résistance à l'éclatement est distribuée normalement avec une variance de 9.

Des essais sur un échantillon de 10 réservoirs conduisent à une résistance moyenne à l'éclatement de 219 kg/cm^2.

Estimer par intervalle de confiance la résistance moyenne à l'éclatement de ce type de réservoir au niveau de confiance de 95 %.

Exercice 10

Le comptable d'une entreprise veut obtenir une estimation du coût moyen de la main d'œuvre directe pour la fabrication d'une pièce particulière. Sur un échantillon aléatoire de 12 lots, on a obtenu les coûts en euros suivants :

982 ; 990 ; 985 ; 855 ; 910 ; 947 ; 842 ; 964 ; 941 ; 760 ; 810 ; 920.

On suppose que les coûts sont répartis normalement dans l'ensemble des lots de la production.

1°) - Estimer ponctuellement la moyenne et l'écart-type du coût de main d'œuvre par lot produit.

2°) - Estimer par intervalle de confiance, au niveau de confiance de 95 %, la moyenne du coût de main d'œuvre par lot produit.