

Série d'exercices n°1-Solutions

Questions à Choix Multiples :

I) Identifier la bonne réponse

- La moyenne $\bar{X}$ est une
A. variable déterministe
B. caractéristique de la population
C. estimation de μ
D. statistique-échantillon
- Pour l'estimation de la moyenne μ , l'erreur d'échantillonnage (v.a.) est
A. $E(|\bar{X} - \mu|)$
B. $|\bar{X} - \mu|$
C. $E[(\bar{X} - \mu)^2]$
D. $\sigma_{\bar{X}}$
- La probabilité que la moyenne de l'échantillon soit exactement égale à la moyenne d'une grande population est
A. petite
B. nulle
C. grande
D. moyenne
- Pour un échantillon de grande taille, la distribution d'échantillonnage de $\bar{X}$ est une
A. normale
B. binomiale
C. normale standard
D. Student
- Pour l'estimation d'une proportion p , l'erreur-type du $\bar{p}$ d'un ESSR, est
A. $\sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$
B. nulle
C. $\sqrt{\frac{p(1-p)}{n}}$
D. $\sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$
- L'échantillonnage stratifié est adéquat lorsque les variances internes des strates sont
A. importantes
B. faibles
C. égales
D. différenciées
- Les échantillons ESAR et ESSR donnent, approximativement, les mêmes résultats lorsque la taille de l'échantillon représente au plus de la population
A. 25%
B. 5%
C. 10%
D. 15%

II) Identifier la mauvaise réponse

- La valeur prise par la v.a. $\bar{X}$ au moment de l'observation est une
A. réalisation d'une variable aléatoire
B. valeur aléatoire
C. valeur déterministe
D. une valeur dans l'espace d'observation E
- le sondage stratifié
A. consiste à diviser la population en sous populations homogènes
B. est utilisé lorsque les variances internes des strates sont faibles
C. est une méthode d'échantillonnage non aléatoire
D. est utilisé lorsque les variances externes des strates sont grandes
- le sondage par grappe
A. est utilisé lorsque la population est composée de différents groupes intérieurement hétérogènes
B. doit éviter les grosses grappes au profit de grappes de tailles le plus uniformes que possible
C. est une méthode d'échantillonnage aléatoire
D. peut fixer au préalable la taille finale de l'échantillon

Exercice 1 : Une compagnie d'assurance constate que chaque année 0.4% de ces assurés meurent à la suite d'une maladie. Posons X le nombre de sinistre à payer au cours d'une année dans laquelle la compagnie gère 100 000 dossiers.

1. Déterminer la loi exacte de la variable aléatoire X et déduisez la probabilité que la compagnie n'ait pas de sinistre à payer (calculer $P(X = 0)$).
2. En utilisant une approximation adéquate, déterminez la probabilité que la compagnie payera
 - a. exactement 440 risques assurés (calculer $P(X = 440)$).
 - b. plus de 440 risques assurés (calculer $P(X > 440)$).

Solution de l'exercice 1

Soit X le nombre de sinistre à payer au cours d'une année. On dénombre seulement les sinistrés qui meurent à la suite d'une maladie.

- 1- Si on prend au hasard un dossier parmi 100 000 dossiers, il y'a une probabilité de 0.004 que le dossier correspond à un sinistre qu'on doit payer, la loi décrite ici est la loi Bernoulli $B(0.004)$:

$$P(Y = 1) = 0.004 \text{ et } P(Y = 0) = 0.996$$

La somme des 100 000 observations de Bernoulli est la v.a. X qui suit une loi binomiale $B(100000, 0.004)$ et on a :

$$P(X = k) = C_n^k (0.004)^k (0.996)^{n-k} ; \quad n = 100000; \quad k = 0, 1, 2, \dots, n$$

$$P(\text{la compagnie nait pas de sinistre à payer}) = P(X = 0) = C_{100000}^0 (0.004)^0 (0.996)^{100000} \\ = (0.996)^{100000} = 0.$$

- 2- $X \sim B(100000 ; 0.004)$; $E(X) = 100000 * 0.004 = 400$;

$$Var(X) = 400 * 0.996 = 398.4 ; \quad \sigma_X = 19.96.$$

$n = 100000 \gg 30$, d'après le théorème central limite (TCL), on aura une bonne approximation de la loi de X par la loi normale,

$$\frac{X - E(X)}{\sigma_X} \sim N(0,1)$$

- a. $P(X = 440) = P(439.5 < X < 440.5)$

$$= P\left(\frac{439.5 - 400}{19.96} < \frac{X - 400}{19.96} < \frac{440.5 - 400}{19.96}\right) \\ \approx P(1.98 < Z < 2.03) = \phi(2.03) - \phi(1.98) = 0.9788 - 0.9761 = 0.0027 \\ = 0.27\%$$

($Z \sim N(0, 1)$, ϕ est sa fonction de répartition)

- b.

$$P(X > 440) = P\left(\frac{X - 400}{19.96} > \frac{440 - 400}{19.96}\right) \approx P(Z > 2.004) = 1 - \phi(2) = 1 - 0.9772$$

$$P(X > 440) = 2.28\%$$

Si on va utiliser la table de la loi normale standard, on trouvera seulement les quantiles $z_{97.72\%} = 2$ et $z_{97.79\%} = 2.01$, on choisira celui le plus proche. Dans cet exemple on a remplacé $\phi(2.004)$ par $\phi(2)$.

Exercice 2 : Un appareil contient un composant essentiel dont l'espérance de la durée de vie est $\mu = 10$ h, avec un écart-type de $\sigma = 2$ h. L'appareil doit fonctionner pendant 2200 heures.

Combien de composants faut-il avoir en réserve pour garantir une fiabilité de 99% ?

Solution de l'exercice 2

Soit n le nombre de composants nécessaire pour un fonctionnement continu de 2200 heures.

Soit X la v.a. durée de vie d'un composant, soit $X_1, X_2, \dots, X_n$ les n durées de vie des n composants.

$$E(X) = 10 = \mu, \quad \sigma = \sigma(X) = 2$$

Soit $S_n = \sum_{i=1}^n X_i$ la durée totale du fonctionnement de l'appareil, $E(S_n) = n\mu = 10n$

$$Var(S_n) = \sum_{i=1}^n Var(X_i) = n\sigma^2 = 4n \quad \text{et} \quad \sigma(S_n) = 2\sqrt{n}$$

(Les X_i sont indépendants, de même loi puisque les composants sont supposés de même type)

Calculons le nombre n , il doit vérifier la condition suivante,

$$P(S_n \geq 2200) = 99\%$$

D'après le TCL, S_n suit une loi normale quand n est grand,

$$\frac{S_n - 10n}{2\sqrt{n}} \sim N(0, 1)$$

$$\begin{aligned} P(S_n \geq 2200) &= P\left(\frac{S_n - 10n}{2\sqrt{n}} \geq \frac{2200 - 10n}{2\sqrt{n}}\right) \approx P\left(Z \geq \frac{2200 - 10n}{2\sqrt{n}}\right) = 1 - \Phi\left(\frac{2200 - 10n}{2\sqrt{n}}\right) \\ &= \Phi\left(\frac{-2200 + 10n}{2\sqrt{n}}\right) = 99\% \end{aligned}$$

$$\frac{-2200 + 10n}{2\sqrt{n}} \text{ est le quantile d'ordre } 0.99$$

On observe dans la table de la loi normale standard, on trouve $z_{0.99} = 2.33$. On doit résoudre l'équation suivante,

$$\frac{-2200 + 10n}{2\sqrt{n}} = 2.33$$

$$\Leftrightarrow 10n - 4.66\sqrt{n} - 2200 = 0$$

On calcule $\Delta = 4.66^2 + 4 * 10 * 2200 = 88021.7156$ et $\sqrt{\Delta} = \sqrt{88021.7156} = 296.68$

Il y a deux solutions à cette équation, l'une prend une valeur négative, l'autre prend une valeur positive, c'est celle qui nous intéresse. On trouve ainsi :

$$\sqrt{n} = \frac{4.66 + 296.68}{20} = 15.067 \Rightarrow n = 227.02$$

On prendra $n = 228$.

Il faudra avoir 228 composants en stock pour garantir (au moins) un fonctionnement de 220 heures avec une fiabilité de 99%.

Exercice 3 : Afin d'estimer leurs moyennes respectives, on échantillonne deux populations. On utilise un ESAR de taille n_1 pour la population P_1 , qui présente un écart type égal à σ_1 . Pour la population P_2 dont l'écart type vaut $\sigma_2 = 2\sigma_1$, on prend un ESAR de taille $n_2 = 2n_1$. Pour quel échantillon l'estimation de la moyenne de la population est-elle précise ?

Solution de l'exercice 3

Soit $\bar{X}^1$ la moyenne empirique calculée sur l'échantillon de la population P_1 et $\bar{X}^2$ celle calculée sur l'échantillon de la population P_2 .

$\bar{X}^1$ est un e.s.b. de μ_1 moyenne de la population P_1 , de même $\bar{X}^2$ est un e.s.b. de μ_2 moyenne de la population P_2 .

$$MSE(\bar{X}^1) = Var(\bar{X}^1) = \frac{\sigma_1^2}{n_1}$$

$$MSE(\bar{X}^2) = Var(\bar{X}^2) = \frac{\sigma_2^2}{n_2} = \frac{4\sigma_1^2}{2n_1} = 2 MSE(\bar{X}^1) > MSE(\bar{X}^1)$$

$$MSE(\bar{X}^2) > MSE(\bar{X}^1)$$

L'échantillon issu de la population P_1 est plus précis que celui de la population P_2 .

Le fait qu'on a doublé la taille du 2^{ème} échantillon n'améliore pas la précision de l'estimation puisque dans la population P_2 il y'a plus de variabilité (de fluctuations) que P_1 .

Exercice 4 : Trois firmes ont des inventaires différents par leur taille. La firme A a une population de 2000 pièces. La firme B a une population de 5000 pièces et la firme C a une population de 10000 pièces. L'écart type de la population pour le coût d'une pièce est $\sigma = 144$.

1. En utilisant le facteur d'exhaustivité, calculez l'erreur type pour chacune des trois firmes, étant donné que les échantillons sélectionnés sont de type ESSR et de taille 50.
2. Quelle est la probabilité que pour chaque firme, la moyenne d'échantillon $\bar{X}$ s'écarte au plus de ± 25 de la moyenne de la population μ .

Solution de l'exercice 4

Les données de l'exercice sont,

$$n = 50; N_A = 2000; N_B = 5000; N_C = 10000; \frac{n}{N_A} = \frac{50}{2000} = 2.5\%; \frac{n}{N_B} = \frac{50}{5000} = 1\%;$$

$$\frac{n}{N_C} = \frac{50}{10000} = 0.5\%.$$

1. Pour un ESSR, l'erreur-type ou l'écart-type de la moyenne empirique est égale à

$$\sigma(\bar{X}_{50}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} = \frac{144}{\sqrt{50}} \sqrt{1 - \frac{50}{N}}$$

Pour la population A :

$$\sigma(\bar{X}_{50}) = \frac{144}{\sqrt{50}} \sqrt{0.975} = 20.11$$

Pour la population B :

$$\sigma(\bar{X}_{50}) = \frac{144}{\sqrt{50}} \sqrt{0.99} = 20.26$$

Pour la population C :

$$\sigma(\bar{X}_{50}) = \frac{144}{\sqrt{50}} \sqrt{0.995} = 20.31$$

Pour un échantillon ESAR l'erreur-type est égale à $\sigma(\bar{X}_n) = 20.36$.

Quand N augmente, les deux échantillons ESSR et ESAR fournissent presque la même erreur-type.

2. On doit calculer $P(|\bar{X}_{50} - \mu| \leq 25)$

$$P(|\bar{X}_{50} - \mu| \leq 25) = P\left(\frac{\sqrt{50}}{144} \frac{|\bar{X}_{50} - \mu|}{\sqrt{1 - \frac{50}{N}}} \leq \frac{\sqrt{50}}{144} \frac{25}{\sqrt{1 - \frac{50}{N}}}\right) \approx P\left(|Z| \leq \frac{1.2276}{\sqrt{1 - \frac{50}{N}}}\right)$$

($n = 50 > 30$)

Pour la population A, $N_A = 2000$:

$$P(|Z| \leq 1.243) = 2\phi(1.24) - 1 = 2 * 0.8925 - 1 = 0.785$$

Pour la population B, $N_B = 5000$:

$$P(|Z| \leq 1.234) = 2\phi(1.23) - 1 = 2 * 0.8907 - 1 = 0.781$$

Pour la population C, $N_C = 10000$:

$$P(|Z| \leq 1.231) = 2\phi(1.23) - 1 = 2 * 0.8907 - 1 = 0.781$$

Pour un ESAR, on a :

$$\begin{aligned} P(|\bar{X}_{50} - \mu| \leq 25) &= P\left(\frac{\sqrt{50}}{144} \frac{|\bar{X}_{50} - \mu|}{\sqrt{1 - \frac{50}{N}}} \leq \frac{\sqrt{50} * 25}{144}\right) \cong P(|Z| \leq 1.228) = 2\phi(1.23) - 1 \\ &= 2 * 0.8907 - 1 = 0.781 \end{aligned}$$

Quand N augmente, la probabilité calculée pour un ESSR se confond avec la probabilité calculée pour un ESAR.

Exercice 5 : Selon la magazine *USA Today*, le nombre moyen des jours par an passés sur les routes pour un représentant commercial est égal à 115. L'écart type est de 60 jours par an. Supposez que ces résultats soient associés à la population des représentants commerciaux et qu'un échantillon de 100 représentants soit sélectionné.

1. Quelle est la valeur de l'écart type de la moyenne ?
2. Quelle est la probabilité que la moyenne d'échantillon soit supérieur à 115 jours par an ?
3. Quelle est la probabilité que la moyenne d'échantillon s'écarte au plus de ± 5 jours de la moyenne de la population ?
4. Quelle serait la probabilité de la question 3. si la taille d'échantillon était 200 ?

Solution de l'exercice 5

1. La v.a. moyenne $\bar{X}$ de l'échantillon est de moyenne $E(\bar{X}) = \mu = 115$ et de variance σ^2/n .

L'écart-type de $\bar{X}$ sera $\sigma/\sqrt{n}$. Avec σ l'écart-type de la population qui est égale à 60 jours. Pour un échantillon de taille $n = 100$, l'écart-type de la moyenne de l'échantillon est égal à

$$\frac{60}{\sqrt{100}} = 6 \text{ jours.}$$

2. La loi de $\bar{X}$ n'est pas connue, mais en utilisant le TCL on peut approcher la loi de

$$\sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma} \right) = \frac{\bar{X} - 115}{6}$$

par la loi $N(0, 1)$; $n = 100 > 30$, on aura une bonne approximation.

$$P(\bar{X} \geq 115) = P\left(\frac{\bar{X} - 115}{6} \geq \frac{115 - 115}{6}\right) \approx P(Z \geq 0) = 1 - \phi(0) = 0.5$$

3. On doit calculer $P(|\bar{X} - \mu| \leq 5)$

$$P(|\bar{X} - \mu| \leq 5) = P\left(\frac{|\bar{X} - 115|}{6} \leq \frac{5}{6}\right) \approx P(|Z| \leq 0.833) = 2\phi(0.833) - 1 = 2 * 0.7967 - 1$$

$$P(|\bar{X} - \mu| \leq 5) = 0.5934 = 59.34\%$$

4. On prévoit que la probabilité augmente,

$$P(|\bar{X} - \mu| \leq 5) = P\left(\sqrt{200} \frac{|\bar{X} - 115|}{60} \leq \sqrt{2} \frac{5}{6}\right) \approx P(|Z| \leq 1.1785) = 2\phi(1.18) - 1 = 2 * 0.881 - 1$$

$$P(|\bar{X} - \mu| \leq 5) = 0.762 = 76.2\%$$

En effet, quand on a augmenté la taille de l'échantillon, la probabilité d'être au voisinage de la moyenne de la population accroît. La loi faible des grands nombres stipule que cette probabilité va converger vers 1 si la taille de l'échantillon devient très grande.

Exercice 6 : Des briques sont fabriquées selon une technique qui assure un poids normalement distribué d'espérance 1.6 kg et d'écart type 30 g. On prélève un échantillon aléatoire simple avec remise, de moyenne $\bar{X}$.

1. Quel est l'écart type de $\bar{X}$, lorsque sa taille vaut 50 et lorsqu'elle vaut 100 ?
2. Quelle est la loi de $\bar{X}$?
3. À partir de quelle taille de l'échantillon on aura plus de 95% de chance, pour que la moyenne $\bar{X}$ s'écartera-t-elle de sa valeur espérée de moins de 20 g ?
4. Supposons qu'on prélève deux échantillons indépendants I et II de tailles respectives 50 et 100 briques. On considère leurs deux moyennes respectives $\bar{X}_I$ et $\bar{X}_{II}$. Quelle est la loi de $\bar{X}_I - \bar{X}_{II}$? Déduisez la probabilité que les moyennes des deux échantillons diffèrent de plus de 10 g ?

Solution de l'exercice 6

La loi de X est exactement une loi normale $X \sim N(1600g, \sigma = 30g)$

1. Pour un échantillon aléatoire simple tiré avec remise, l'écart-type de $\bar{X}$ sera :

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{n}} g$$

Si l'échantillon est de taille $n = 50$, on aura :

$$\sigma(\bar{X}) = \frac{30}{\sqrt{50}} g = 4.243g$$

Si l'échantillon est de taille $n = 100$, l'erreur-type devient,

$$\sigma(\bar{X}) = \frac{30}{10} g = 3g$$

Si on augmente la taille n de l'échantillon, les fluctuations de la moyenne de l'échantillon autour de la moyenne de la population vont diminuer.

2. La loi de $\bar{X}$ est exactement une loi normale,

$$\bar{X} \sim N(1600g, \frac{30}{\sqrt{n}}g)$$

Toute combinaison linéaire de v.a. indépendantes et de lois normales, est une loi normale. En particulier la moyenne $\bar{X}$.

Rappel : soient n v.a. indépendantes $X_i \sim N(\mu_i, \sigma_i)$; $i = 1, \dots, n$ alors toute combinaison linéaire des X_i est une v.a. de loi normale,

$$\sum_{i=1}^n \alpha_i X_i \sim N\left(\sum_{i=1}^n \alpha_i \mu_i, \sqrt{\sum_{i=1}^n \alpha_i^2 \sigma_i^2}\right) \text{ avec } \alpha_i \in \mathbb{R} \text{ pour tout } i = 1, \dots, n$$

En particulier, si $\alpha_i = 1/n$, $\mu_i = \mu$ et $\sigma_i = \sigma$ pour tout $i = 1, \dots, n$; on aura la loi de la moyenne $\bar{X}$.

3. On doit chercher la valeur de n tel que,

$$P(|\bar{X} - 1600| \leq 20) \geq 0.95$$

$$P(|\bar{X} - 1600| \leq 20) = P\left(\sqrt{n} \frac{|\bar{X} - 1600|}{30} \leq \sqrt{n} \frac{20}{30}\right) = P\left(|Z| \leq \sqrt{n} \frac{2}{3}\right) = 2\phi\left(\sqrt{n} \frac{2}{3}\right) - 1$$

$$2\phi\left(\sqrt{n} \frac{2}{3}\right) - 1 \geq 0.95 \Rightarrow \phi\left(\sqrt{n} \frac{2}{3}\right) \geq \frac{1.95}{2} = 0.975 = \phi(1.96)$$

$$\sqrt{n} \frac{2}{3} \geq 1.96 \Rightarrow \sqrt{n} \geq 2.94 \Rightarrow n = 8.64$$

On prend $n = 9$. On a besoin d'un échantillon de taille au moins 9, pour être confiant à plus de 95% que la moyenne $\bar{X}$ s'écartera de la moyenne 1600g de moins de 20g.

4. Les v.a. $\bar{X}_I$ et $\bar{X}_{II}$ sont de lois normales, elles sont indépendantes du fait que les deux échantillons sont indépendants, alors toute combinaison linéaire de $\bar{X}_I$ et $\bar{X}_{II}$ est de loi normale.

$$\bar{X}_I \sim N(1600g, 4.243g) \text{ et } \bar{X}_{II} \sim N(1600g, 3g)$$

$a\bar{X}_I + b\bar{X}_{II}$ suit une loi normale, en particulier $\bar{X}_I - \bar{X}_{II}$, ($a = 1$ et $b = -1$)

Posons $D = \bar{X}_I - \bar{X}_{II}$ alors $D \sim N(E(D), \sigma_D)$ avec

$$E(D) = E(\bar{X}_I - \bar{X}_{II}) = E(\bar{X}_I) - E(\bar{X}_{II}) = 0,$$

$$\sigma_D^2 = \text{Var}(D) = \text{Var}(\bar{X}_I - \bar{X}_{II}) = \text{Var}(\bar{X}_I) + \text{Var}(\bar{X}_{II}) = \sigma^2 \left(\frac{1}{50} + \frac{1}{100}\right) = 900 * \frac{3}{100} = 27,$$

$$\sigma_D = 3\sqrt{3} = 5.20g.$$

$$D \sim N(0, 5.20g)$$

$$P(|D| \geq 10g) = P\left(|Z| \geq \frac{10}{5.20}\right) = 2 - 2 * \phi(1.92) = 2 - 2 * 0.9726 = 0.0548 = 5.48\%$$

Si on contrôlait le poids des briques par des échantillons de taille 100 et on veut réduire les frais de contrôle, pour ne considérer que des échantillons de taille 50, alors il y'aura un risque de 5.5% que le résultat obtenu par l'échantillon I soit différent de plus de 10g que le résultat obtenu par l'échantillon II .