# CORRIGE DES EXERCICES : Distributions d'échantillonnage - Intervalles de variation

## Exercice 1

$\mathcal{P}$={élèves du secondaire}
X= résultat de fluidité au test de pensée Créative de Torrance, variable quantitative de moyenne connue $\mu$ =20, et d'écart-type connu $\sigma$ =6,5 dans $\mathcal{P}$.
Echantillons de taille n de X issu de $\mathcal{P}$ pour lesquels $\bar{x}$ , $s$ et $s^*$ ne sont pas calculés.

1) On peut prévoir le résultat moyen observé $\bar{x}$ pour chaque échantillon par la moyenne de la moyenne empirique $\overline{X}_n$ qui est égale à $\mu$ puisque $\overline{X}_n$ est un estimateur sans biais de $\mu$ : cette prévision est constante pour tous les échantillons et vaut $\mu$ =20.

2) On peut calculer la variance (écart-type) du résultat moyen par la variance (écart-type) de la moyenne empirique $\overline{X}_n$ qui est égale à $\frac{\sigma^2}{n}$ (égal à $\frac{\sigma}{\sqrt{n}}$) qui varie avec la taille de l'échantillon : plus la taille de l'échantillon est grande plus la variance (écart-type) est faible d'où une plus grande précision dans l'estimation (cf tableau ci-dessous colonnes 3 et 4).

| taille | distribution de la moyenne empirique $\overline{X}_n$ | | | distribution de la variance empirique | | distribution de l'écart-type empirique | |
|---|---|---|---|---|---|---|---|
| | | | | sans biais $S_n^{2*}$ | biaisée $S_n^2$ | sans biais $S_n^*$ | biaisé $S_n$ |
| n | moyenne $\mu$ | variance $\frac{\sigma^2}{n}$ | écart-type $\frac{\sigma}{\sqrt{n}}$ | moyenne $\sigma^2$ | moyenne $\left(\frac{n-1}{n}\right)\sigma^2$ | moyenne $\sigma$ | moyenne $\sqrt{\frac{n-1}{n}}\,\sigma$ |
| **1** | **20** | **42,25** | **6,5** | **42,25** | | **6,5** | |
| 20 | 20 | 2,1125 | 1,4534 | 42,25 | 40,1375 | 6,5 | 6,3354 |
| 50 | 20 | 0,845 | 0,9192 | 42,25 | 41,4050 | 6,5 | 6,4347 |
| 100 | 20 | 0,4225 | 0,65 | 42,25 | 41,8275 | 6,5 | 6,4674 |

*remarque* : on pourra affiner la prévision du résultat moyen observé en calculant un intervalle de variation au risque $\alpha$ (par exemple $\alpha$=5%) de la moyenne empirique $\overline{X}_n$ en utilisant l'approximation normale sur $\overline{X}_n$ pour les deux échantillons de taille 50 et 100 (n≥30), qui prédira le résultat moyen observé avec un risque d'erreur de $\alpha$ ($\alpha$=5%) en faisant intervenir sa moyenne $\mu$ et son écart-type $\frac{\sigma}{\sqrt{n}}$ : $I_{95\%}(\overline{X}_n) \approx \left[\mu \pm z_{0,975}\frac{\sigma}{\sqrt{n}}\right]$

pour n=50 $I_{95\%}(\overline{X}_n) \approx [20 \pm 1{,}96 \times 0{,}9192] = [20 \pm 1{,}8] = [18{,}2\ ;\ 21{,}8]$

pour n=100 $I_{95\%}(\overline{X}_n) \approx [20 \pm 1{,}96 \times 0{,}65] = [20 \pm 1{,}274] \approx [20 \pm 1{,}3] = [18{,}7\ ;\ 21{,}3]$

3) On peut prévoir la variance observée du résultat, biaisée $s^2$ ou sans biais $s^{*2}$ pour chaque échantillon, par la moyenne de la variance empirique biaisée $S_n^2$ ou sans biais $S_n^{2*}$.

- la moyenne de $S_n^2$ est égale à $\left(\frac{n-1}{n}\right)\sigma^2$ : $S_n^2$ est un estimateur biaisé de $\sigma^2$ qui sous estime toujours $\sigma^2$. Cette prévision varie avec la taille de l'échantillon : plus la taille de l'échantillon est grande plus le biais est faible, d'où une prévision qui se rapproche de $\sigma^2$ (cf tableau ci-dessus colonne 6).
- la moyenne de $S_n^{2*}$ est égale à $\sigma^2$ puisque $S_n^{2*}$ est un estimateur sans biais de $\sigma^2$ : cette prévision est constante pour tous les échantillons et vaut $\sigma^2 = 6{,}5^2 = 42{,}25$ (cf tableau ci-dessus colonne 5).

On peut prévoir l'écart-type observé du résultat, biaisé $s$ ou sans biais $s^*$ pour chaque échantillon, par la moyenne de l'écart-type empirique biaisé $S_n$ ou sans biais $S_n^*$.

- la moyenne de $S_n$ est égale à $\sqrt{\frac{n-1}{n}}\,\sigma$ : $S_n$ est un estimateur biaisé de $\sigma$ qui sous estime toujours $\sigma$. Cette prévision varie avec la taille de l'échantillon : plus la taille de l'échantillon est grande plus le biais est faible, d'où une prévision qui se rapproche de $\sigma$ = 6,5 (cf tableau ci-dessus colonne 8).
- la moyenne de $S_n^*$ est égale à $\sigma$ puisque $S_n^*$ est un estimateur sans biais de $\sigma$ : cette prévision est constante pour tous les échantillons et vaut $\sigma$ =6,5 (cf tableau ci-dessus colonne 7).

**Exercice 2**

$\mathcal{P}$={français recensés en 1999}

X= âge, variable quantitative X~$\mathcal{N}(\mu=39, \sigma=23)$ dans $\mathcal{P}$.

Echantillons de taille n=25 de X issu de $\mathcal{P}$

1) La moyenne empirique de l'âge, $\overline{X}_{25}$ a une distribution normale de moyenne $\mu$=39, de variance $\frac{\sigma^2}{n}=\frac{23^2}{25}=21{,}16$ et d'écart-type $\frac{\sigma}{\sqrt{n}}=\frac{23}{\sqrt{25}}=4{,}6$ puisque X a une distribution normale de moyenne $\mu$=39 et d'écart-type $\sigma$=23 dans $\mathcal{P}$.

2) $P(\overline{X}_{25} \le 20)=P\left(Z \le \frac{20-39}{4{,}6}\right)=F(-4{,}13)=1-F(4{,}13)\approx 1-F(4{,}1)=1-0{,}999979=0{,}000021$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ quasiment aucun des échantillons de taille 25 de X issus de $\mathcal{P}$ ont un âge moyen observé inférieur à 20 ans.

3) $P(20 \le \overline{X}_{25} \le 60)=P(\overline{X}_{25} \le 60)-P(\overline{X}_{25} \le 20)$ avec $P(\overline{X}_{25} \le 20)=0{,}000021$ et

$P(\overline{X}_{25} \le 60)=P\left(Z \le \frac{60-39}{4{,}6}\right)=F(4{,}565)\approx F(4{,}6)=0{,}999998$ d'où $P(20 \le \overline{X}_{25} \le 60)=0{,}999998-0{,}000021=0{,}999977$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ quasiment tous les échantillons de taille 25 de X issus de $\mathcal{P}$ ont un âge moyen observé compris entre 20 et 60 ans.

4) Intervalle de variation à 90% (au risque $\alpha$=10%) de l'âge moyen sur les échantillons de taille 25 de X issus de $\mathcal{P}$ :

$I_{90\%}(\overline{X}_n)=[Q_{0{,}05}\,;Q_{0{,}95}]=[39\pm 4{,}6\times z_{0{,}95}]=[39\pm 4{,}6\times 1{,}645]=[39\pm 7{,}567]\approx[39\pm 7{,}6]=[31{,}4\,;46{,}6]$

car $z_{1-(\alpha/2)}=z_{0{,}95}=1{,}645$ est le quantile d'ordre 0,95 de la loi $\mathcal{N}(0,1)$.

➡ 90% des échantillons de taille 25 de X issus de $\mathcal{P}$ ont un âge moyen compris entre 31,4 et 46,6 ans.

Intervalle de variation à 95% (au risque $\alpha$=5%) de l'âge moyen sur les échantillons de taille 25 de X issus de $\mathcal{P}$ :

$I_{95\%}(\overline{X}_n)=[Q_{0{,}025}\,;Q_{0{,}975}]=[39\pm 4{,}6\times z_{0{,}975}]=[39\pm 4{,}6\times 1{,}96]=[39\pm 9{,}016]\approx[39\pm 9]=[30\,;48]$

car $z_{1-(\alpha/2)}=z_{0{,}975}=1{,}96$ est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 95% des échantillons de taille 25 de X issus de $\mathcal{P}$ ont un âge moyen compris entre 30 et 48 ans.

➡ la valeur de la borne inférieure de cet intervalle de variation à 95% ne peut plus remettre en cause l'hypothèse de normalité faite sur la variable moyenne empirique de l'âge sur les échantillons de taille 25.

5) On observe un âge moyen de 35 ans, alors qu'on s'attendait "raisonnablement" (dans 95% des cas) à observer un âge moyen compris entre 30 et 48 ans, ce qui n'est pas surprenant : on ne peut donc pas mettre en cause la représentativité de l'échantillon pour la variable âge dans la population des femmes françaises du recensement de 1990.

6) La demi-longueur de l'intervalle de variation à 95% de l'âge moyen $I_{95\%}(\overline{X}_n)$ est d'environ 9 ans (cf question 4) ; pour obtenir une demi-longueur plus faible, de 2 ans maximum, il faudrait donc plus de 25 femmes. Pour n inconnu, $\sigma$=23 et $\alpha$=5% connus, la demi-longueur de l'intervalle $I_{95\%}(\overline{X}_n)$ s'écrit : $z_{0{,}975}\frac{\sigma}{\sqrt{n}}=1{,}96\frac{23}{\sqrt{n}}$.

On cherche n tel que : $1{,}96\frac{23}{\sqrt{n}} \le 2$ c'est à dire $\frac{1{,}96\times 23}{2} \le \sqrt{n}$ d'où $n \ge \left(\frac{1{,}96\times 23}{2}\right)^2=22{,}54^2=508{,}05$

➡ on choisirait donc une taille d'échantillon au moins égale à 509 pour que la demi-longueur de l'intervalle de pari à 95% soit inférieure à 2 ans. On aurait donc une marge d'erreur à 95% d'au plus 2 ans dans l'estimation de la moyenne d'âge dans $\mathcal{P}$, c'est à dire dans l'intervalle [39 ± 2] soit entre 37 et 41 ans, pour 95% des échantillons de taille n = 509.

**Exercice 3**

$\mathcal{P}$={enfants de 12 ans}

X= résultat au test de richesse et de précision du vocabulaire, variable quantitative de moyenne connue $\mu$ =60, et d'écart-type connu $\sigma$ =10 dans $\mathcal{P}$.

Echantillons de taille n de X issu de $\mathcal{P}$ pour lesquels $\bar{x}$ , $s$ et $s^*$ ne sont pas calculés.

1) La moyenne empirique $\overline{X}_n$ a pour moyenne $\mu$ =60 (puisque $\overline{X}_n$ est un estimateur sans biais de $\mu$) : cette moyenne est constante quelle que soit la taille des échantillons (cf tableau ci-dessous colonne 2).

   La moyenne empirique $\overline{X}_n$ a pour variance $\frac{\sigma^2}{n}$ et pour écart-type $\frac{\sigma}{\sqrt{n}}$ qui varient avec la taille des échantillons : plus la taille de l'échantillon est grande plus variance et écart-type sont faibles, d'où une plus grande précision dans l'estimation (cf tableau ci-dessous colonnes 3 et 4).

   La forme de la distribution de la moyenne empirique $\overline{X}_n$ est inconnue (quelconque) tant que la taille de l'échantillon est faible (n<30) puisque la distribution de X est inconnue (quelconque). Lorsque la taille de l'échantillon est suffisamment grande (n≥30), on peut considérer, d'après le théorème central-limite, que la distribution de $\overline{X}_n$ est approximativement normale $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (cf tableau ci-dessous colonne 5).

| taille | distribution de la moyenne empirique $\overline{X}_n$ | | | | |
|---|---|---|---|---|---|
| n | moyenne | variance | écart-type | forme | $P(\overline{X}_n > 63)$ |
| **1** | **60** | **100** | **10** | **inconnue** | inconnue |
| 4 | 60 | 25 | 10/2=5 | inconnue | inconnue |
| 8 | 60 | 12,5 | 10/2√2=3,54 | inconnue | inconnue |
| 16 | 60 | 6, | 10/4=2,5 | inconnue | inconnue |
| 32 | 60 | 3,125 | 10/4√2=1,77 | approx. normale | ≈ 0,04460 |
| 64 | 60 | 1,5625 | 10/8=1,25 | approx. normale | ≈ 0,00820 |
| 100 | 60 | 1 | 10/10=1 | approx. normale | ≈ 0,00135 |

2) Pour un échantillon de taille n=16 la forme de la distribution de la moyenne empirique $\overline{X}_n$ est inconnue (quelconque) puisque n<30. Il est donc impossible de calculer cette probabilité.

3) Pour un échantillon de taille n=32 la forme de la distribution de la moyenne empirique $\overline{X}_n$ est approximativement normale puisque n≥30. Il est donc possible de calculer cette probabilité de manière approchée :

$$P(\overline{X}_{32} > 63) = 1 - P(\overline{X}_{32} \le 63) \approx 1 - P\left(Z \le \frac{63-60}{1{,}768}\right) \approx 1 - F(1{,}7) = 1 - 0{,}9554 = 0{,}0446$$

   où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

   ➨ environ 4,46% des échantillons de taille 32 de X issus de $\mathcal{P}$ ont un résultat moyen supérieur à 63.

4) Cette probabilité diminue avec la taille de l'échantillon puisque l'écart-type de la moyenne empirique $\overline{X}_n$ diminue (cf tableau ci-dessus colonne 4) : la distribution de $\overline{X}_n$ étant donc plus concentrée autour de sa moyenne $\mu$ =60, la probabilité $P(\overline{X}_n > 63)$ représentée par la surface à droite de la valeur 63 sous la densité de la loi de $\overline{X}_n$ (approximativement normale pour n≥30), sera plus petite.

   Pour n≥30 $\overline{X}_n$ est approximativement normale, il est donc possible de calculer cette probabilité de manière approchée :

   • pour n=64 $P(\overline{X}_{64} > 63) = 1 - P(\overline{X}_{64} \le 63) \approx 1 - P\left(Z \le \frac{63-60}{1{,}25}\right) = 1 - F(2{,}4) = 1 - 0{,}9918 = 0{,}0082$

   où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

   ➨ environ 0,82% des échantillons de taille 64 de X issus de $\mathcal{P}$ ont un résultat moyen supérieur à 63.

   • pour n=100 $P(\overline{X}_{100} > 63) = 1 - P(\overline{X}_{100} \le 63) \approx 1 - P\left(Z \le \frac{63-60}{1}\right) = 1 - F(3) = 1 - 0{,}99865 = 0{,}00135$

   où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

   ➨ environ 0,135% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont un résultat moyen supérieur à 63.

   ➨ quand la taille de l'échantillon augmente, il est de moins en moins probable d'observer un résultat moyen supérieur à 63 lorsque la vraie moyenne est égale à 60 (cf tableau ci-dessus colonne 6).

**Exercice 4**

$\mathcal{P}$={enfants âgés de 7 ans}

X= quotient intellectuel QI, variable quantitative X~$\mathcal{N}(\mu=100, \sigma=10)$ dans $\mathcal{P}$.

Echantillon de taille n=16 de X issu de $\mathcal{P}$ pour lequel $\bar{x} = 106$ et $s = 13$.

1)

| | population | échantillon |
|---|---|---|
| taille | N= ? | n = 16 |
| moyenne | $\mu = 100$ | $\bar{x} = 106$ |
| variance | $\sigma^2 = 10^2 = 100$ | $s^2 = 13^2 = 169$ |

2) La moyenne empirique du score sur les échantillons de taille 16, $\overline{X}_{16}$ a une distribution normale de moyenne $\mu$=100, de variance $\frac{\sigma^2}{n} = \frac{100}{16} = 6{,}25$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{16}} = 2{,}5$ puisque X a une distribution normale de moyenne $\mu$=100 et d'écart-type $\sigma$=10 dans $\mathcal{P}$.

3) $P(\overline{X}_{16} > \bar{x}) = P(\overline{X}_{16} > 106) = 1 - P(\overline{X}_{16} \le 106) = 1 - P\left(Z \le \frac{106-100}{2{,}5}\right) = 1 - F(2{,}4) = 1 - 0{,}9918 = 0{,}0082$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ 0,82% des échantillons de taille 16 de X issus de $\mathcal{P}$ ont un QI moyen supérieur à 106, QI moyen observé.

4) 2,5% des échantillons de taille 16 ont un QI moyen supérieur au QI cherché, donc 97,5% des échantillons de taille 16 ont un QI moyen inférieur au QI cherché, qui est donc par définition le quantile d'ordre 0,975 de $\overline{X}_{16}$ :

$Q_{0,975} = 100 + (2{,}5 \times z_{0,975}) = 100 + (2{,}5 \times 1{,}96) = 100 + 4{,}9 = 104{,}9$

car $z_{0,975}$=1,96 est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 2,5% des échantillons de taille 16 de X issus de $\mathcal{P}$ ont un QI moyen supérieur à 104,9.

5) Intervalle de variation à 95% (au risque $\alpha$=5%) du QI moyen sur les échantillons de taille 16 de X issus de $\mathcal{P}$ :

❶ $I_{95\%}(\overline{X}_n) = [Q_{0,025}; Q_{0,975}] = [100 \pm 4{,}9] = [95{,}1; 104{,}9]$

car $Q_{0,025}$ est symétrique de $Q_{0,975}$ par rapport à $\mu$ =100.

❷ $I_{95\%}(\overline{X}_n) = [Q_{0,025}; Q_{0,975}] = [100 \pm 2{,}5 \times z_{0,975}] = [100 \pm 2{,}5 \times 1{,}96] = [100 \pm 4{,}9] = [95{,}1\ ;\ 104{,}9]$

car $z_{1-(\alpha/2)}=z_{0,975}$=1,96 est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 95% des échantillons de taille 16 de X issus de $\mathcal{P}$ ont un QI moyen compris entre 95,1 et 104,9.

6) La moyenne empirique du score sur les échantillons de taille 100, $\overline{X}_{100}$ a une distribution normale de moyenne $\mu$=100, de variance $\frac{\sigma^2}{n} = \frac{100}{100} = 1$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$ puisque X a une distribution normale de moyenne $\mu$=100 et d'écart-type $\sigma$=10 dans $\mathcal{P}$.

$P(\overline{X}_{100} > \bar{x}) = P(\overline{X}_{100} > 106) = 1 - P(\overline{X}_{100} \le 106) = 1 - P\left(Z \le \frac{106-100}{1}\right) = 1 - F(6) \approx 1 - 1 = 0$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ quasiment aucun échantillon de taille 100 de X issus de $\mathcal{P}$ n'a un QI moyen supérieur à 106, QI moyen observé.

Le QI cherché, est par définition le quantile d'ordre 0,975 de $\overline{X}_{100}$ :

$Q_{0,975} = 100 + (1 \times z_{0,975}) = 100 + (1 \times 1{,}96) = 100 + 1{,}96 = 101{,}96$

car $z_{0,975}$=1,96 est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 2,5% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont un QI moyen supérieur à 101,96.

Intervalle de variation à 95% (au risque $\alpha$=5%) du QI moyen sur les échantillons de taille 100 de X issus de $\mathcal{P}$ :

❶ $I_{95\%}(\overline{X}_n) = [Q_{0,025}; Q_{0,975}] = [100 \pm 1{,}96] = [98{,}04; 101{,}96]$

car $Q_{0,025}$ est symétrique de $Q_{0,975}$ par rapport à $\mu$ =100.

❷ $I_{95\%}(\overline{X}_n) = [Q_{0,025}; Q_{0,975}] = [100 \pm 1 \times z_{0,975}] = [100 \pm 1{,}96] = [98{,}04\ ;\ 101{,}96]$

car $z_{1-(\alpha/2)}=z_{0,975}$=1,96 est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 95% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont un QI moyen compris entre 98,04 et 101,96.

7) La demi-longueur de l'intervalle de variation à 95% du QI moyen $I_{95\%}(\overline{X}_n)$ est d'environ 4,9 points pour les échantillons de taille n=16 et est d'environ 1,96 point pour les échantillons de taille n=100 ; pour obtenir une demi-longueur plus faible, de 1 point maximum, il faudrait donc plus de 100 enfants. La demi-longueur de l'intervalle $I_{95\%}(\overline{X}_n)$ s'écrit : $z_{0,975} \times \frac{\sigma}{\sqrt{n}}$ avec n inconnu, σ=10 et α=5% connus et pour qu'elle n'excède pas 1 point, il faut que : $z_{0,975} \times \frac{10}{\sqrt{n}} \leq 1$ donc que $1{,}96 \times 10 \leq \sqrt{n}$ c'est à dire que $n \geq 19{,}6^2 = 384{,}16$ donc que $n \geq 385$

➨ on choisirait donc une taille d'échantillon au moins égale à 385 pour que la demi-longueur de l'intervalle de pari à 95% n'excède pas 1 point. On aurait donc une marge d'erreur à 95% maximum de 1 point dans l'estimation du QI moyen dans $\mathcal{P}$, c'est à dire dans l'intervalle [100 ± 1] soit entre 99 et 101 points de QI, pour 95% des échantillons de taille 385.

**Exercice 5**

$\mathcal{P}$={enfants âgés de 4 à 12 ans}

X= score au questionnaire CBCL, variable quantitative de moyenne μ=20 et d'écart-type σ=14 dans $\mathcal{P}$.

Echantillon de taille n=36 de X issu de $\mathcal{P}$ pour lequel $\overline{x} = 20{,}5$.

1)

| | population | échantillon |
|---|---|---|
| taille | N = ? | n = 36 |
| moyenne | μ = 20 | $\overline{x}$ = 20,5 |
| variance | $\sigma^2 = 14^2 = 196$ | $s^2* = ?$ |

2) Puisque n=36 ≥ 30, la moyenne empirique du score $\overline{X}_{36}$ a une distribution approximativement normale de moyenne μ=20, de variance $\frac{\sigma^2}{n} = \frac{196}{36} \approx 5{,}44$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{36}} \approx \sqrt{5{,}44} \approx 2{,}33$ car X a une distribution de moyenne μ=20 et d'écart-type σ=14.

3) $P(\overline{X}_{36} < \overline{x}) = P(\overline{X}_{36} < 20{,}5) \approx P\left(Z \leq \frac{20{,}5-20}{2{,}33}\right) = F(0{,}214) \approx F(0{,}21) = 0{,}5832$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➨ environ 58,3% des échantillons de taille 36 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 20,5, score moyen observé.

4) $P(\overline{X}_{36} > 25) = 1 - P(\overline{X}_{36} \leq 25) \approx 1 - P\left(Z \leq \frac{25-20}{2{,}33}\right) = 1 - F(2{,}1429) \approx 1 - F(2{,}14) = 1 - 0{,}9838 = 0{,}0162$

➨ environ 1,6% (peu) des échantillons de taille 36 de X issu de $\mathcal{P}$ ont un score moyen supérieur à 25.

5) 99% des échantillons de taille 36 ont un score moyen inférieur au score cherché, qui est donc par définition le quantile d'ordre 0,99 de $\overline{X}_{36}$ :

$Q_{0,99} \approx 20 + (2{,}33 \times z_{0,99}) = 20 + (2{,}33 \times 2{,}325) = 20 + 5{,}425 = 25{,}425 \approx 25{,}4$

car $z_{0,99}$=2,325 est le quantile d'ordre 0,99 de la loi $\mathcal{N}(0,1)$.

➨ 99% des échantillons de taille 36 de $\mathcal{P}$ ont un score moyen inférieur à 25,4 environ.

6) Intervalle de variation à 98% (au risque α=2%) du score moyen sur les échantillons de taille 36 de X issus de $\mathcal{P}$ :

❶ $I_{98\%}(\overline{X}_n) = [Q_{0,01}; Q_{0,99}] \approx [20 \pm 5{,}425] \approx [20 \pm 5{,}4] = [14{,}6;\ 25{,}4]$

car $Q_{0,01}$ est symétrique de $Q_{0,99}$ par rapport à μ=20.

❷ $I_{98\%}(\overline{X}_n) = [Q_{0,01}; Q_{0,99}] \approx [20 \pm 2{,}33 \times z_{0,99}] = [20 \pm 2{,}33 \times 2{,}325] \approx [20 \pm 5{,}4] = [14{,}6;\ 25{,}4]$

car $z_{1-(\alpha/2)} = z_{0,99}$=2,325 est le quantile d'ordre 0,99 de la loi $\mathcal{N}(0,1)$.

➨ 98% des échantillons de taille 36 de X issus de $\mathcal{P}$ ont un score moyen compris entre 14,6 et 25,4 environ.

7) La demi-longueur de l'intervalle de variation à 98% du score moyen $I_{98\%}(\overline{X}_n)$ est d'environ 5,4 points pour les échantillons de taille n=36 ; pour obtenir une demi-longueur plus faible, de 4 points maximum, il faudrait donc plus de 36 enfants. La demi-longueur de l'intervalle $I_{98\%}(\overline{X}_n)$ s'écrit : $z_{0,99} \times \frac{\sigma}{\sqrt{n}}$ avec n inconnu, σ=14 et α=1% connus et

pour qu'elle n'excède pas 4 points, il faut que : $z_{0,99} \times \frac{14}{\sqrt{n}} \leq 4$ donc que $2{,}325 \times \frac{14}{4} \leq \sqrt{n}$ c'est à dire que $n \geq \left(2{,}325 \times \frac{14}{4}\right)^2 = 66{,}5$ donc que $n \geq 67$

➡ on choisirait donc une taille d'échantillon au moins égale à 67 pour que la demi-longueur de l'intervalle de pari à 98% n'excède pas 4 points. On aurait donc une marge d'erreur à 98% d'au plus 4 points dans l'estimation du score moyen dans $\mathcal{P}$, c'est à dire dans l'intervalle [20 ± 4] soit entre 16 et 24 points de QI, pour 98% des échantillons de taille 67.

**Exercice 6**

$\mathcal{P}$={sujets} X= score de résistance au stress (test de Stroop), variable quantitative de moyenne μ=50 et d'écart-type σ=25 dans $\mathcal{P}$. Echantillon de taille n=30 de X issu de $\mathcal{P}$ pour lequel $\bar{x} = 41$.

| | population | échantillon |
|---|---|---|
| taille | N = ? | n = 30 |
| moyenne | μ = 50 | $\bar{x} = 41$ |
| écart-type | σ = 25 | $s^* = ?$ |

1) Puisque n=30 ≥ 30, la moyenne empirique du score $\overline{X}_{30}$ a une distribution approximativement normale de moyenne μ=50, de variance $\frac{\sigma^2}{n} = \frac{25^2}{30} \approx 20{,}83$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{30}} \approx 4{,}564$ car X a une distribution de moyenne μ=50 et d'écart-type σ=25.

2) $P(\overline{X}_{30} < \bar{x}) = P(\overline{X}_{30} < 41) \approx P\left(Z \leq \frac{41-50}{4{,}564}\right) = F(-1{,}9718) \approx F(-1{,}97) = 1 - F(1{,}97) = 1 - 0{,}9756 = 0{,}0244$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ environ 2,5% (peu) des échantillons de taille 30 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 41, score moyen observé.

3) $P(\overline{X}_{30} > \bar{x}) = P(\overline{X}_{30} > 41) \approx 1 - P(\overline{X}_{30} < 41) = 1 - 0{,}0244 = 0{,}9756$ avec F fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ environ 97,5% des échantillons de taille 30 de X issu de $\mathcal{P}$ n'a un score moyen supérieur à 41, score moyen observé.

4) 5% des échantillons de taille 30 ont un score moyen supérieur au score cherché donc 95% des échantillons de taille 30 ont un score moyen inférieur au score cherché, qui est donc par définition le quantile d'ordre 0,95 de $\overline{X}_{30}$ : $Q_{0,95} \approx 50 + (4{,}564 \times z_{0,95}) = 50 + (4{,}564 \times 1{,}645) \approx 50 + 7{,}5 \approx 57{,}5$
car $z_{0,95}$=1,645 est le quantile d'ordre 0,95 de la loi $\mathcal{N}(0,1)$.

➡ 95% des échantillons de taille 30 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 57,5 environ.

5) Intervalle de variation à 90% (au risque α=10%) du score moyen sur les échantillons de taille 30 de X issus de $\mathcal{P}$ :

❶ $I_{90\%}(\overline{X}_n) = [Q_{0,05}\,;\, Q_{0,95}] \approx [50 \pm 7{,}5] \approx [50 \pm 7{,}5] = [42{,}5\,;\, 57{,}5]$
car $Q_{0,05}$ est symétrique de $Q_{0,95}$ par rapport à μ=50.

❷ $I_{90\%}(\overline{X}_n) = [Q_{0,05}\,;\, Q_{0,95}] \approx [50 \pm 4{,}564 \times z_{0,95}] = [50 \pm 4{,}564 \times 1{,}645] \approx [50 \pm 7{,}5] = [42{,}5\,;\, 57{,}5]$
car $z_{1-(\alpha/2)} = z_{0,95}$=1,645 est le quantile d'ordre 0,95 de la loi $\mathcal{N}(0,1)$.

➡ 90% des échantillons de taille 30 de X issus de $\mathcal{P}$ ont un score moyen compris entre 42,5 et 57,5 environ.

**Exercice 7**

$\mathcal{P}$={fumeurs} X= score au test de dépendance tabagique de Fagerström, variable quantitative de moyenne μ=5 et d'écart-type σ=4,5 dans $\mathcal{P}$. Echantillon de taille n=45 de X issu de $\mathcal{P}$.

| | population | échantillon |
|---|---|---|
| taille | N = ? | n = 45 |
| moyenne | μ = 5 | $\bar{x} = ?$ |
| écart-type | σ = 4,5 | $s^* = ?$ |

1) Puisque n=45 ≥ 30, la moyenne empirique du score $\overline{X}_{45}$ a une distribution approximativement normale de moyenne μ=5, de variance $\frac{\sigma^2}{n}=\frac{4{,}5^2}{45}=0{,}45$ et d'écart-type $\frac{\sigma}{\sqrt{n}}=\frac{4{,}5}{\sqrt{45}}\approx 0{,}67$ car X a une distribution de moyenne μ=5 et d'écart-type σ=4,5.

2) $P(\overline{X}_{45}<6)\approx P\left(Z\le\frac{6-5}{0{,}67}\right)\approx F(1{,}49)=0{,}9319$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➨ environ 93% des échantillons de taille 45 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 6.

3) $P(\overline{X}_{45}<3)\approx P\left(Z\le\frac{3-5}{0{,}67}\right)\approx F(-2{,}9851)\approx 1-F(2{,}99)=1-0{,}9986=0{,}0014$

➨ environ 0,14% des échantillons de taille 45 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 3.

4) $P(3<\overline{X}_{45}<6)=P(\overline{X}_{45}\le 6)-P(\overline{X}_{45}\le 3)\approx 0{,}9319-0{,}0014=0{,}9305$

➨ environ 93% des échantillons de taille 45 de X issus de $\mathcal{P}$ ont un score moyen compris entre 3 et 6.

5) $P(\overline{X}_{45}>9)=1-P(\overline{X}_{45}\le 9)\approx 1-P\left(Z\le\frac{9-5}{0{,}67}\right)\approx 1-F(4{,}97)\approx 1-F(5)\approx 1-1=0$

➨ quasiment aucun échantillon de taille 45 de X issu de $\mathcal{P}$ n'a un score moyen supérieur à 9.

6) 90% des échantillons de taille 45 ont un score moyen inférieur au score cherché, qui est donc par définition le quantile d'ordre 0,9 de $\overline{X}_{45}$ :

$Q_{0,9}\approx 5+(0{,}67\times z_{0,9})=5+(0{,}67\times 1{,}28)=5+0{,}8576=5{,}8576\approx 5{,}86$

car $z_{0,9}$=1,28 est le quantile d'ordre 0,9 de la loi $\mathcal{N}(0,1)$.

➨ 90% des échantillons de taille 45 de X issus de $\mathcal{P}$ ont un score moyen inférieur à 5,86 environ.

7) Intervalle de variation à 80% (au risque α=20%) du score moyen sur les échantillons de taille 45 de X issus de $\mathcal{P}$ :

❶ $I_{80\%}(\overline{X}_n)=[Q_{0,1};Q_{0,9}]\approx[5\pm 0{,}8576]\approx[5\pm 0{,}86]=[4{,}14\,;5{,}86]$
car $Q_{0,1}$ est symétrique de $Q_{0,9}$ par rapport à μ=5.

❷ $I_{80\%}(\overline{X}_n)=[Q_{0,1};Q_{0,9}]\approx[5\pm 0{,}67\times z_{0,9}]=[5\pm 0{,}67\times 1{,}28]\approx[5\pm 0{,}86]=[4{,}14\,;5{,}86]$
car $z_{1-(\alpha/2)}=z_{0,9}$=1,28 est le quantile d'ordre 0,9 de la loi $\mathcal{N}(0,1)$.

➨ 80% des échantillons de taille 45 de X issus de $\mathcal{P}$ ont un score moyen compris entre 4,14 et 5,86 environ.

**Exercice 8**

$\mathcal{P}$={consommateurs}

X= influence de la marque de commerce lors de l'achat d'un produit, variable qualitative dichotomique : oui, non

p = proportion de consommateurs influencés par la marque de commerce dans $\mathcal{P}$, p = 25% = 0,25

Echantillon de taille n=100 de X issu de $\mathcal{P}$ sur lequel lesquels ni le nombre observé de consommateurs influencés, ni la fréquence (proportion) observée de consommateurs influencés $f$ ne sont donnés.

1)

| | population | échantillon |
|---|---|---|
| taille | N = ? | n = 100 |
| proportion | p = 0,25 | $f$ = ? |

2) Puisque n =100 ≥ 30, np = 100×0,25 = 25 ≥ 5 d'où n(1–p) = 100×0,75 ≥ 5, la fréquence empirique $F_{100}$ a une distribution approximativement normale de moyenne p = 0,25

de variance $\frac{p(1-p)}{n}=\frac{0{,}25\times 0{,}75}{100}=0{,}001875$, et d'écart-type $\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0{,}25\times 0{,}75}{100}}=\sqrt{0{,}001875}\approx 0{,}0433$ car la proportion de consommateurs influencés par la marque de commerce p = 0,25 dans $\mathcal{P}$.

3) Observer au moins 35 consommateurs influencés sur 100 c'est observer une fréquence $f$ de plus de 0,35.

$P(F_{100}>f)=P(F_{100}>0{,}35)=1-P(F_{100}\le 0{,}35)\approx 1-P\left(Z\le\frac{0{,}35-0{,}25}{0{,}0433}\right)\approx 1-F(2{,}31)=1-0{,}9896=0{,}0104$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ la probabilité d'observer une fréquence de consommateurs influencés par la marque de commerce supérieure à 35% sur un échantillon de taille 100 de X issus de $\mathcal{P}$ (c'est à dire plus de 35 consommateurs influencés par la marque de commerce), est d'environ 1%.
environ 1% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont une fréquence observée de consommateurs influencés par la marque de commerce supérieure à 35%.

4) Observer moins de 20 consommateurs influencés sur 100 c'est observer une fréquence $f$ de moins de 0,2.

$P(F_{100} \le f) = P(F_{100} \le 0{,}2) \approx 1 - P\left(Z \le \frac{0{,}2 - 0{,}25}{0{,}0433}\right) \approx F(-1{,}15) = 1 - F(1{,}15) = 1 - 0{,}8749 = 0{,}1251$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ la probabilité d'observer une fréquence de consommateurs influencés par la marque de commerce inférieure à 20% sur un échantillon de taille 100 de X issus de $\mathcal{P}$ (c'est à dire moins de 20 consommateurs influencés par la marque de commerce), est d'environ 12,5%.
environ 12,5% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont une fréquence observée de consommateurs influencés par la marque de commerce inférieure à 20%.

5) Intervalle de variation à 90% (au risque $\alpha$=10%) de la fréquence observée sur les échantillons de taille 100 de X issus de $\mathcal{P}$ : $I_{90\%}(F_n) = [Q_{0,05}; Q_{0,95}] \approx \left[0{,}25 \pm z_{0,95} \times \sqrt{\frac{0{,}25 \times 0{,}75}{100}}\right] = [0{,}25 \pm 1{,}645 \times 0{,}0433] \approx [0{,}25 \pm 0{,}07]$

$I_{90\%}(F_n) \approx [0{,}18 ; 0{,}32] \approx [18\% ; 32\%]$ car $z_{1-(\alpha/2)}=z_{0,95}=1{,}645$ est le quantile d'ordre 0,95 de la loi $\mathcal{N}(0,1)$.

➡ 90% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont une fréquence observée de consommateurs influencés par la marque de commerce comprise entre environ 18% et 32% (soit environ entre 18 et 32 consommateurs influencés par la marque de commerce).
On peut donc s'attendre à ce que, sur les 100 consommateurs de l'échantillon observé, 18 à 32 d'entre-eux environ soient influencés par la marque de commerce.

6) On observe 31 consommateurs influencés par la marque de commerce sur 100 consommateurs, soit une fréquence observée de 31%, alors qu'on s'attendait "raisonnablement" (dans 90% des cas) à en observer entre 18% et 32%, ce qui n'est pas surprenant : sur cette observation, on ne peut pas mettre en doute la représentativité de l'échantillon.

**Exercice 9**

$\mathcal{P}$={lancers d'une pièce de monnaie équilibrée}

X= tomber sur le côté face, variable qualitative dichotomique : oui, non

p = proportion de lancers tombant sur le côté face dans $\mathcal{P}$, p = 50% = 0,5

Echantillons de taille n de X issu de $\mathcal{P}$ sur lesquels ni le nombre de faces observé, ni la fréquence (proportion) observée de faces $f$ ne sont donnés.

1) La fréquence empirique $F_n$ a pour moyenne p =0,5 (puisque $F_n$ est un estimateur sans biais de p) : cette moyenne est constante quelle que soit la taille des échantillons (cf tableau ci-après colonne 2).

La fréquence empirique $F_n$ a pour variance $\frac{p(1-p)}{n}$ et pour écart-type $\sqrt{\frac{p(1-p)}{n}}$ qui varient avec la taille des échantillons : plus la taille de l'échantillon est grande plus la variance et l'écart-type sont faibles d'où une plus grande précision dans l'estimation (cf tableau ci-après colonnes 3 et 4).
Puisque la distribution de X n'est pas normale ("inconnue"), la forme de la distribution de la fréquence empirique $F_n$ n'est pas normale ("inconnue") tant que la taille de l'échantillon est faible (n<30).
Lorsque la taille de l'échantillon est suffisamment grande (n≥30) et que la proportion p n'est pas trop proche de 0 ou de 1 (np ≥5 et n(1−p) ≥5), on peut considérer que la distribution de $F_n$ est approximativement normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
(cf tableau ci-après colonnes 5 et 6).

| taille | distribution de la fréquence empirique $F_n$ | | | | | |
|---|---|---|---|---|---|---|
| n | moyenne | variance | écart-type | condition | forme | $P(F_n > 0{,}6)$ |
| 5 | 0,5 | 0,05 | 0,2236 | | "inconnue" | "inconnue" |
| 10 | 0,5 | 0,025 | 0,1581 | | "inconnue" | "inconnue" |
| 30 | 0,5 | 0,0083 | 0,0913 | n/2=15 | approx. normale | ≈ 0,1367 |
| 50 | 0,5 | 0,005 | 0,0707 | n/2=25 | approx. normale | ≈ 0,0786 |
| 100 | 0,5 | 0,0025 | 0,05 | n/2=50 | approx. normale | ≈ 0,0228 |

2) Pour un échantillon de taille n=30 la forme de la distribution de la fréquence empirique $F_n$ est approximativement normale puisque n≥30 (np = n(1–p) =30/2=15≥5). Il est possible de calculer cette probabilité de manière approchée, en remarquant préalablement qu'observer 18 faces sur 30 lancers correspond à observer une fréquence de 18/30=0,6 :

$$P(F_{30} > 0{,}6) = 1 - P(F_{30} \le 0{,}6) \approx 1 - P\left(Z \le \frac{0{,}6 - 0{,}5}{0{,}0913}\right) \approx 1 - F(1{,}1) = 1 - 0{,}8643 = 0{,}1357$$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➨ environ 13,6% des échantillons de taille 30 de X issus de $\mathcal{P}$ ont un nombre de faces supérieur à 18 (ou une fréquence de faces supérieure à 0,6).

3) Pour un échantillon de taille n=50 la forme de la distribution de la fréquence empirique $F_n$ est approximativement normale puisque n≥30 et np = n(1–p) =50/2=25≥5. Il est donc possible de calculer cette probabilité de manière approchée :

$$P(F_{50} > 0{,}6) = 1 - P(F_{50} \le 0{,}6) \approx 1 - P\left(Z \le \frac{0{,}6 - 0{,}5}{0{,}0707}\right) \approx 1 - F(1{,}41) = 1 - 0{,}9207 = 0{,}0793$$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➨ environ 7,9% des échantillons de taille 50 de X issus de $\mathcal{P}$ ont une fréquence de faces supérieure à 0,6 (nombre de faces supérieur à 50×0,6=30).

4) Cette probabilité diminue avec la taille de l'échantillon puisque l'écart-type de la fréquence empirique $F_n$ diminue (cf tableau ci-dessus colonne 4) : la distribution de $F_n$ étant donc plus concentrée autour de sa moyenne p =0,5, la probabilité $P(F_n>0{,}6)$ représentée par la surface à droite de la valeur 0,6 sous la densité de la loi de $F_n$ (approximativement normale pour n≥30) sera plus petite.
Pour n≥30, $F_n$ est approximativement normale, il est donc possible de calculer cette probabilité de manière approchée :

pour n=100 $P(F_{100} > 0{,}6) = 1 - P(F_{100} \le 0{,}6) \approx 1 - P\left(Z \le \frac{0{,}6 - 0{,}5}{0{,}05}\right) = 1 - F(2) = 1 - 0{,}9772 = 0{,}0228$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➨ environ 2,3% des échantillons de taille 100 de X issus de $\mathcal{P}$ ont une fréquence de faces supérieure à 0,6 (nombre de faces supérieur à 100×0,6=60).

➨ quand la taille de l'échantillon augmente, il est de moins en moins probable d'observer une fréquence de faces supérieure à 0,6 lorsque la vraie proportion est égale à 0,5 (cf tableau ci-dessus colonne 6).

5) On s'attend à observer en moyenne une fréquence de faces égale à p=0,6 c'est à dire 60 faces sur un échantillon de 100 lancers. On peut raffiner cette prévision en donnant l'intervalle de variation, par exemple à 95% (au risque α=5%), de la fréquence empirique de faces sur les échantillons de taille 100 de X issus de $\mathcal{P}$, déduit à partir de l'approximation normale de $F_n$ puisque n=100≥30 (np = n(1–p) =100/2=50≥5) :

$$I_{95\%}(F_n) = \left[Q_{0{,}025}\,;\,Q_{0{,}975}\right] \approx \left[0{,}5 \pm z_{0{,}975} \times \sqrt{\tfrac{0{,}5\times 0{,}5}{100}}\right] = \left[0{,}5 \pm 1{,}96 \times 0{,}05\right] = \left[0{,}5 \pm 0{,}098\right] \approx [0{,}5 \pm 0{,}1] = [0{,}4\ ;\ 0{,}6]$$

$I_{95\%}(F_n)$ ≈ [40% ; 60%] car $z_{1-(\alpha/2)}=z_{0{,}975}=1{,}96$ est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➨ pour 95% des échantillons de taille 100 de X issus de $\mathcal{P}$ la fréquence observée de faces sera comprise entre environ 40% et 60% (soit environ entre 40 et 60 faces sur 100 lancers), ce que l'on s'attend donc "raisonnablement" à observer.

6) On observe 35 faces sur 100 lancers soit une fréquence observée de faces de 35%, alors qu'on s'attendait "raisonnablement" (dans 95% des cas) à en observer entre 40% et 60%, ce qui est surprenant : soit l'échantillon observé fait partie des 5% qui ont, par construction de l'intervalle, une fréquence de faces en dehors de l'intervalle de pari à 95%, soit l'échantillon observé n'est pas représentatif de la variable (par ex, la pièce n'est pas réellement équilibrée donc la proportion de faces dans $\mathcal{P}$ est différente de (inférieure à) 50%). On peut donc douter de la représentativité de l'échantillon.

7) Sur un échantillon de taille n=100, la demi-longueur de l'intervalle de variation à 95% est d'environ 10% (cf question 5) ; pour obtenir une demi-longueur plus faible, de 5%, il faudrait donc plus de 100 lancers. Pour n inconnu, p=0,50 et α=5% connus, la demi-longueur de l'intervalle $I_{95\%}(F_n)$ s'écrit : $z_{0{,}975}\sqrt{\frac{p(1-p)}{n}} = 1{,}96\sqrt{\frac{0{,}5(1-0{,}5)}{n}}$.

On cherche n tel que : $1{,}96\frac{0{,}5}{\sqrt{n}} \le 5\% = 0{,}05$ c'est à dire $\frac{1{,}96}{0{,}05} \times 0{,}5 \le \sqrt{n}$ d'où $n \ge \left(\frac{1{,}96}{0{,}05} \times 0{,}5\right)^2 = (1{,}96 \times 10)^2 = 384{,}16$

➨ on choisirait donc une taille d'échantillon au moins égale à 385 pour que la demi-longueur de l'intervalle de pari à 95% soit inférieure à 5%. On aurait donc une marge d'erreur à 95% d'au plus 5% dans l'estimation de la proportion de faces dans $\mathcal{P}$, c'est à dire dans l'intervalle [0,5 ± 0,05] = [0,45 ; 0,55] soit entre 45% et 55%, pour 95% des échantillons de taille n=385.

**Exercice 10**

$\mathcal{P}$={étudiants d'une université}

X= pratique d'au moins une activité physique par semaine, variable qualitative dichotomique : oui, non

p = proportion d'étudiants pratiquant au moins une activité physique par semaine dans $\mathcal{P}$, p = 45% = 0,45

Echantillon de taille n=400 de X issu de $\mathcal{P}$ sur lequel on observe 205 étudiants pratiquant au moins une activité physique par semaine c'est à dire une fréquence (proportion) observée $f$ = 205/400 = 0,5125.

1)

| | population | échantillon |
|---|---|---|
| taille | N = ? | n = 400 |
| proportion | p = 0,45 | $f$ = 0,5125 |

2) Puisque $n = 400 \geq 30$, $np = 400 \times 0{,}45 = 180 \geq 5$ d'où $n(1-p) = 400 \times 0{,}55 \geq 5$, la fréquence empirique $F_{400}$ a une distribution approximativement normale de moyenne p = 0,45 , de variance $\frac{p(1-p)}{n} = \frac{0{,}45 \times 0{,}55}{400} = 0{,}00061875$ , et d'écart-type $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}45 \times 0{,}55}{400}} = \sqrt{0{,}00061875} \approx 0{,}0249 \approx 0{,}025$ car la proportion d'étudiants pratiquant au moins une activité physique par semaine p = 0,45 dans $\mathcal{P}$.

3) $P(F_{400} > f) = P(F_{400} > 0{,}5125) = 1 - P(F_{400} \leq 0{,}5125) \approx 1 - P\left(Z \leq \frac{0{,}5125 - 0{,}45}{0{,}0249}\right) \approx 1 - F(2{,}51) = 1 - 0{,}994 = 0{,}006$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

➡ environ 0,6% des échantillons de taille 400 de X issus de $\mathcal{P}$ ont une fréquence d'étudiants pratiquant au moins une activité physique par semaine supérieure à 51,25%. L'échantillon observé fait partie de ces 0,6% d'échantillons : on avait très peu de chance d'observer autant d'étudiants (205 ou plus) pratiquant au moins une activité physique par semaine sur un échantillon de taille 400, ce résultat est donc surprenant.

4) Intervalle de variation à 95% (au risque $\alpha$=5%) de la fréquence observée sur les échantillons de taille 400 de X issus de $\mathcal{P}$ : $I_{95\%}(F_n) = [Q_{0,025}; Q_{0,975}] \approx \left[0{,}45 \pm z_{0,975} \times \sqrt{\frac{0{,}45 \times 0{,}55}{400}}\right] = [0{,}45 \pm 1{,}96 \times 0{,}0249] \approx [0{,}45 \pm 0{,}0488]$

$I_{95\%}(F_n) \approx [0{,}45 \pm 0{,}05] \approx [0{,}4 ; 0{,}5] \approx [40\% ; 50\%]$ car $z_{1-(\alpha/2)} = z_{0,975} = 1{,}96$ est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$.

➡ 95% des échantillons de taille 400 de X issus de $\mathcal{P}$ ont une fréquence observée d'étudiants pratiquant au moins une activité physique par semaine comprise entre environ 40% et 50% (soit environ entre 160 et 200 étudiants pratiquant au moins une activité physique par semaine).

5) La demi-longueur de l'intervalle précédent $I_{95\%}(F_n)$ est d'environ 5% ; pour obtenir une demi-longueur plus faible, de 2%, il faudrait donc plus de 400 étudiants. Pour n inconnu, p=0,45 et $\alpha$=5% connus, la demi-longueur de l'intervalle $I_{95\%}(F_n)$ s'écrit : $z_{0,975}\sqrt{\frac{p(1-p)}{n}} = 1{,}96\sqrt{\frac{0{,}45(1-0{,}55)}{n}}$ . On cherche n tel que : $1{,}96\sqrt{\frac{0{,}45 \times 0{,}55}{n}} \leq 2\% = 0{,}02$ c'est à dire $\frac{1{,}96}{0{,}02}\sqrt{0{,}45 \times 0{,}55} \leq \sqrt{n}$ d'où $n \geq \left(\frac{1{,}96}{0{,}02}\right)^2 \times 0{,}45 \times 0{,}55 = 2376{,}99$

➡ on choisirait donc une taille d'échantillon au moins égale à 2 377 pour que la demi-longueur de l'intervalle de pari à 95% soit inférieure à 2%. On aurait donc une marge d'erreur d'au plus ± 2% dans l'estimation de la proportion d'étudiants pratiquant au moins une activité physique dans $\mathcal{P}$, c'est à dire dans l'intervalle $[0{,}45 \pm 0{,}02]$ = [0,43 ; 0,47] soit entre 43% et 47%, pour 95% des échantillons de taille 2 377.

**Exercice 11**

$\mathcal{P}$ = {français}
X = consommer des antidépresseurs, variable qualitative dichotomique : oui, non
p = proportion de français consommateurs d'antidépresseurs dans $\mathcal{P}$, p connue dans $\mathcal{P}$ : p = 10% = 0,1

1) Pour un échantillon de taille n=30 de X issu de $\mathcal{P}$ n =30 ≥ 30, mais np = 30×0,1 =3 > 5, il n'est donc pas possible de considérer que la fréquence empirique $F_{30}$ a une distribution approximativement normale. En revanche $F_{30}$ a pour moyenne p = 0,1 et pour écart-type $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,1\times 0,9}{30}} = \sqrt{0,003} = 0,05477 \approx 0,055$ car la proportion de français consommateurs d'antidépresseurs dans $\mathcal{P}$ est p = 0,1.

2) Pour un échantillon de taille n=80 de X issu de $\mathcal{P}$, puisque n =80 ≥ 30, np = 80×0,1 =8 ≥ 5 d'où n(1–p) = 80×0,9 ≥ 5, la fréquence empirique $F_{80}$ a une distribution approximativement normale de moyenne p = 0,1 et d'écart-type $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,1\times 0,9}{80}} = \sqrt{0,001125} = 0,03354 \approx 0,034$ car la proportion de français consommateurs d'antidépresseurs dans $\mathcal{P}$ est p = 0,1.

3) Observer 12 consommateurs d'antidépresseurs, ou plus sur n=80 correspond à observer une fréquence (proportion) d'au moins $f$ = 12/80 = 0,15. La probabilité correspondante s'écrit :

$$P(F_{80} > f) = P(F_{80} > 0,15) = 1 - P(F_{80} \le 0,15) \approx 1 - P\left(Z \le \frac{0,15-0,1}{0,034}\right) = 1 - F(1,4706) \approx 1 - F(1,47) = 1 - 0,9292 = 0,0708$$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$ (calcul exact : 1–F(1,4907) ≈ 1–F(1,49)=1–0,9319=0,0681).

➨ environ 7% des échantillons de taille 80 de X issus de $\mathcal{P}$ ont une fréquence de consommateurs d'antidépresseurs supérieure à 15%, soit plus de 12 consommateurs sur 80 français.

4) Observer 2 consommateurs d'antidépresseurs ou moins sur n=80 correspond à observer une fréquence (proportion) d'au plus $f$ = 2/80 = 0,025. La probabilité correspondante s'écrit :

$$P(F_{80} \le f) = P(F_{80} \le 0,025) \approx P\left(Z \le \frac{0,025-0,1}{0,034}\right) \approx F(-2,2059) \approx 1 - F(2,21) = 1 - 0,9864 = 0,0136$$ où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$ (calcul exact : F(–2,2361) ≈ 1–F(2,24) = 1–0,9875=0,0125).

➨ environ 1% des échantillons de taille 80 de X issus de $\mathcal{P}$ ont une fréquence de consommateurs d'antidépresseurs inférieure à 2,5%, soit moins de 2 consommateurs sur 80 français.

5) Intervalle de variation au niveau 95% (au risque $\alpha$=5%) de la fréquence empirique sur les échantillons de taille 80 de X issus de $\mathcal{P}$ : $I_{95\%}(F_{80}) = [Q_{0,025}\,;\,Q_{0,975}] \approx \left[0,1 \pm z_{0,975} \times \sqrt{\frac{0,1\times 0,9}{80}}\right] = [0,1 \pm 1,96 \times 0,034] \approx [0,1 \pm 0,066] = [0,034\,;\,0,166]$

$I_{95\%}(F_{80}) \approx [10\% \pm 6,6\%] \approx [3,4\%\,;\,16,6\%]$ car $z_{1-(\alpha/2)}=z_{0,975}=1,96$ est le quantile d'ordre 0,975 de la loi $\mathcal{N}(0,1)$

➨ 95% des échantillons de taille 80 de X issus de $\mathcal{P}$ ont une fréquence observée de français consommateurs d'antidépresseurs entre 3,4% et 16,6% environ (soit entre 3 et 13 consommateurs environ sur 80 français).

6) Observer 10 consommateurs sur 80 personnes interrogées c'est observer une fréquence de $f$ =10/80=0,125=12,5% plus élevée que la proportion p=10% mais qui appartient à l'intervalle de variation à 95% trouvé précédemment, c'est à dire que la probabilité d'observer cette proportion de 12,5% ou plus est supérieure à 2,5%, elle n'est donc pas très faible.
(Elle est égale à :

$$P(F_{80} > f) \approx P(F_{80} > 0,125) \approx 1 - P\left(Z \le \frac{0,125-0,1}{0,034}\right) = 1 - F(0,7353) \approx 1 - F(0,74) = 1 - 0,7704 = 0,2296$$

donc environ 23% des échantillons de taille 80 de X issus de $\mathcal{P}$ ont une proportion de français consommateurs d'antidépresseurs supérieure à 12,5%)

➨ une proportion non négligeable d'échantillons (plus de 2,5%) de taille 80 de X issus de $\mathcal{P}$ ont une proportion de français consommateurs d'antidépresseurs supérieure à 12,5%, soit plus de 10 consommateurs sur 80. On peut donc en déduire que ce n'est pas surprenant d'observer une telle fréquence de consommateurs et que l'on ne peut pas mettre en doute la représentativité de l'échantillon pour la variable étudiée dans la population étudiée, la consommation d'antidépresseurs chez les français.