**TD 2**

**Exercice 7 :**

On fait l'hypothèse que le taux de salaire horaire est distribué selon une loi normale $N(\mu, \sigma^2)$ avec $\mu = 24,07$ dollars et $\sigma = 4,80$ dollars (une loi log-normale serait sans doute plus appropriée mais les calculs seraient beaucoup plus compliqués).

Soit $X_i$ la variable aléatoire représentant le salaire de l'individu $i$ de l'échantillon de taille $n = 120$. Le tirage se fait sans remise mais il est raisonnable de penser que la taille de la population est très large devant la taille de l'échantillon donc on peut supposer que les $X_i$ sont indépendants et identiquement distribués selon la loi $N(\mu, \sigma^2)$ (comme si le tirage avait été fait avec remise).

On pose $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ .

$\bar{X}_n$ est une combinaison linéaire de variables aléatoires **indépendantes** suivant une loi normale donc $\bar{X}_n$ suit une une loi normale. Pour déterminer les paramètres de cette loi, on calcule l'espérance et la variance de $\bar{X}_n$. Par linéarité de l'espérance :

$$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu$$

et par indépendance des $X_i$

$$V(\bar{X}_n) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

d'où le résultat (classique) :

$$\bar{X}_n \rightsquigarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

On cherche à déterminer la probabilité de l'évènement $-\beta \leq \bar{X}_n - \mu \leq \beta$ pour deux valeurs de $\beta$ ($\beta = 0,5$ et $\beta = 1$).

Puisque $\bar{X}_n \rightsquigarrow N\left(\mu, \frac{\sigma^2}{n}\right)$ alors

$$U_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

On a : $P\left(-\beta \leq \bar{X}_n - \mu \leq \beta\right) = P\left(-\frac{\beta}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\beta}{\frac{\sigma}{\sqrt{n}}}\right) = P(-\frac{\beta\sqrt{n}}{\sigma} \leq U_n \leq \frac{\beta\sqrt{n}}{\sigma}) = P(|U_n| \leq \frac{\beta\sqrt{n}}{\sigma})$.

Notons $F$ la fonction de répartition de $N(0,1)$. On a établi dans un exercice précédent que si $U \rightsquigarrow N(0,1)$ alors pour tout $u \geq 0$, $P(|U| \leq u) = 2F(u) - 1$ donc:

$$P\left(-\beta \leq \bar{X}_n - \mu \leq \beta\right) = 2F(\frac{\beta\sqrt{n}}{\sigma}) - 1$$

Sachant que $n = 120$ et $\sigma = 4,8$ on a :

a) pour $\beta = 0,5$ : $\frac{\beta\sqrt{n}}{\sigma} \simeq 1.14$ d'où $P\left(-0,5 \leq \bar{X}_n - \mu \leq 0,5\right) = 2F(1.14) - 1$, or $F(1.14) = 0.8729$ (Cf. table statistique de $N(0,1)$) donc

$$P\left(-0,5 \leq \bar{X}_n - \mu \leq 0,5\right) = 0.7458$$

b) pour $\beta = 1$ : $\frac{\beta\sqrt{n}}{\sigma} \simeq 2.28$ d'où $P\left(-1 \leq \bar{X}_n - \mu \leq 1\right) = 2F(2.28) - 1$ or $F(2.28) = 0.9887$ donc

$$P\left(-1 \leq \bar{X}_n - \mu \leq 1\right) = 0.9774$$

**Exercice** 8 :

A tout individu $i$ de l'échantillon on associe la variable aléatoire $X_i$ prenant la valeur 1 si l'individu $i$ coopère et 0 s'il ne coopère pas. On adopte le modèle statistique suivant: les $X_i$ sont indépendantes et identiquement distribuées selon une loi de Bernouilli de paramètre $p = 0.4$. Le fait de supposer que les $X_i$ sont i.i.d peut être justifié par les mêmes raisons que celles citées à l'exercice précédent.

Posons $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ($\hat{p}_n$ est la proportion empirique des individus qui coopérent).

On a $n\hat{p}_n = \sum_{i=1}^{n} X_i \rightsquigarrow B(n,p)$ car les $X_i$ suivent la même loi $B(1,p)$ et sont indépendantes. On peut effectuer l'approximation normale de la loi binômiale (approcher $B(n,p)$ par $N\left(np, np(1-p)\right)$) si la condition $np(1-p) > 15$ est vérifiée.

Dans ce cas $np(1-p) = 400 \times 0.4 \times 0.6 = 96 > 15$ donc on peut affirmer qu'approximativement :

$$n\hat{p}_n \rightsquigarrow N\left(np, np(1-p)\right)$$

d'où

$$\hat{p}_n \rightsquigarrow N\left(p, \frac{p(1-p)}{n}\right)$$

ou encore

$$U_n = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

On cherche à déterminer $P(\hat{p}_n \geq \bar{p})$ où $\bar{p} = 0,375$. On a:

$$\hat{p}_n \geq \bar{p} \Longleftrightarrow U_n = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

donc

$$P(\hat{p}_n \geq \bar{p}) = P\left(U_n \geq \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Avec les valeurs $\bar{p} = 0,375$ , $p = 0.4$ on trouve $\frac{\bar{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \simeq -1,02$ d'où

$$P(\hat{p}_n \geq 0.375) = P\left(U_n \geq -1.02\right) = 1 - F(-1,02) = F(1.02) = 0.8461$$

où $F$ désigne la fonction de répartition de $N(0,1)$.

**Exercice** 9

1- Résultat classique : si les $X_i \overset{i.i.d}{\rightsquigarrow} N(m,\sigma^2)$ alors la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et efficace de $m$ (sa variance est égale à la borne de Fréchet-Darmois-Cramer-Rao, Cf cours). Il s'agit donc du "meilleur" estimateur de $m$ dans la classe des estimateurs sans biais.

Cet estimateur suit la loi $N\left(m, \frac{\sigma^2}{n}\right)$ (démonstration dans l'exercice 7), c'est-à-dire dans ce cas particulier $N\left(m, \frac{4}{n}\right)$

2- La meilleure prévision de $X_{n+1}$ est la variable aléatoire $\hat{x}$ qui ne dépend que de $X_1, X_2, .., X_n$ et qui minimise l'erreur quadratique, c'est-à-dire telle que

$$MSE = E\left[(X_{n+1} - \hat{x})^2\right] \text{ soit minimale}$$

On a :

$$\begin{aligned} MSE &= E\left[(X_{n+1} - \hat{x})^2\right] \\ &= E\left[(X_{n+1} - m + m - \hat{x})^2\right] \\ &= E\left[(X_{n+1} - m)^2 + 2(X_{n+1} - m)(m - \hat{x}) + (m - \hat{x})^2\right] \\ &= E\left[(X_{n+1} - m)^2\right] + 2E\left[(X_{n+1} - m)(m - \hat{x})\right] + E\left[(m - \hat{x})^2\right] \end{aligned}$$

Puisque $\hat{x}$ ne dépend que de $X_1, X_2, .., X_n$ qui sont toutes des variables indépendantes de $X_{n+1}$, alors $\hat{x}$ et $X_{n+1}$ sont des variables indépendantes et par conséquent $X_{n+1} - m$ et $m - \hat{x}$ sont des variables indépendantes. Or on sait que si $U$ et $V$ sont des variables aléatoires indépendantes alors $E(UV) = E(U)E(V)$ (en effet on a alors $0 = cov(U,V) = E(UV) - E(U)E(V)$) donc $E\left[(X_{n+1} - m)(m - \hat{x})\right] = E(X_{n+1} - m)E\left(m - \hat{x}\right)$. Mais $E(X_{n+1}) = m$ donc $E(X_{n+1} - m) = 0$ d'où $E\left[(X_{n+1} - m)(m - \hat{x})\right] = 0$. Il s'ensuit que l'erreur quadratique moyenne se réduit à :

$$\begin{aligned} MSE &= E\left[(X_{n+1} - m)^2\right] + E\left[(m - \hat{x})^2\right] \\ &= V(X_{n+1}) + E\left[(m - \hat{x})^2\right] \\ &= \sigma^2 + E\left[(m - \hat{x})^2\right] \end{aligned}$$

Par conséquent minimiser l'erreur quadratique moyenne $MSE$ revient à minimiser $E\left[(m - \hat{x})^2\right]$. Or minimiser $E\left[(m - \hat{x})^2\right]$ dans l'ensemble des $\hat{x}$ ne dépendant que de

$X_1, X_2, .., X_n$ revient à chercher le meilleur estimateur de $m$. Si on se restreint aux estimateurs non biaisés alors on sait d'après la question 1 que $\bar{X}_n$ est le meilleur estimateur de $m$ et par conséquent $\hat{x} = \bar{X}_n$.

Conclusion: $\bar{X}_n$ est la meilleure prévision de $X_{n+1}$.

3- L'erreur de prévision $Z = X_{n+1} - \bar{X}_n$ suit une loi normale en tant que combinaison linéaire de deux variables indépendantes suivant une loi normale ($\bar{X}_n$ et $X_{n+1}$ sont indépendantes car $\bar{X}_n$ est la combinaison linéaire de variables indépendantes de $X_{n+1}$ à savoir $X_1, X_2, .., X_n$). Son espérance est

$$E(Z) = E(X_{n+1}) - E(\bar{X}_n) = m - m = 0$$

et sa variance est:

$$\begin{aligned} V(Z) &= V(X_{n+1}) + V(\bar{X}_n) \text{ (car } \bar{X}_n \text{ et } X_{n+1} \text{ sont indépendantes)} \\ &= \sigma^2 + \frac{\sigma^2}{n} = \sigma^2 \frac{n+1}{n} = 4\frac{n+1}{n} \end{aligned}$$

Conclusion :

$$Z \rightsquigarrow N\left(0, 4\frac{n+1}{n}\right)$$

4- Déterminer un intervalle de prévision pour $X_{n+1}$ à 80% (centré sur sa meilleure prévision $\bar{X}_n$) revient à trouver un réel $\alpha_n > 0$ tel que

$$P(\bar{X}_n - \alpha_n \leq X_{n+1} \leq \bar{X}_n + \alpha_n) = 0.80$$

c'est-à-dire tel que:

$$P(-\alpha_n \leq X_{n+1} - \bar{X}_n \leq \alpha_n) = 0.80$$

Pour pouvoir exploiter la table statistique de $N(0,1)$, on commence par réduire la variable $X_{n+1} - \bar{X}_n$ (qui est déjà centrée) :

$$U_n = \frac{X_{n+1} - \bar{X}_n}{\sqrt{\sigma^2 \frac{n+1}{n}}} \rightsquigarrow N(0,1)$$

puis on cherche le réel $u > 0$ tel que

$$P(-u \leq U_n \leq u) = 0.80$$

La valeur de $u$ n'est pas donnée directement par la table mais on sait que $P(-u \leq U_n \leq u) = P(|U_n| \leq u) = 2F(u) - 1$ où $F$ est la fonction de répartition de $N(0,1)$ donc $2F(u) - 1 = 0.80$ soit $F(u) = 0,90$ d'où $u \simeq 1,28$ (on trouve dans la table statistique $F(1,28) = 0.8997$; si on veut avoir une valeur de $u$ plus précise, on peut faire une interpolation linéaire entre $1,28$ et $1,29$ et on trouve alors $u \simeq 1,2816$)

On a donc :

$$P(-1,2816 \leq \frac{X_{n+1} - \bar{X}_n}{\sqrt{\sigma^2 \frac{n+1}{n}}} \leq 1,2816) = 0.80$$

d'où

$$P(\bar{X}_n - 1,2816\sqrt{\sigma^2 \frac{n+1}{n}} \leq X_{n+1} \leq \bar{X}_n + 1,2816\sqrt{\sigma^2 \frac{n+1}{n}}) = 0.80$$

Pour $\sigma = 4$ et $n = 24$, on trouve :

$$P(\bar{X}_{24} - 2,616 \leq X_{25} \leq \bar{X}_{24} + 2,616) = 0.80$$

c'est-à-dire que $\left[\bar{X}_{24} - 2,616, \bar{X}_{24} + 2,616\right]$ est un intervalle de prévision à 80% de $X_{25}$.

**Exercice** 10

Enoncé (corrigé) : Pour simuler un tirage d'une loi normale standard $N(0,1)$, on calcule souvent la somme de 12 tirages indépendants d'une loi uniforme $U_{[0,1]}$ auquel on enlève 6. Pouvez-vous justifier cette prodédure ?

*Rappels:*

*1- Si la variable aléatoire $Z$ suit une loi uniforme sur l'intervalle $[a,b]$ (notée $U_{[a,b]}$) alors*

$$E(Z) = \frac{a+b}{2}$$

$$V(Z) = \frac{(b-a)^2}{12}$$

*2- Théorème central limite : Si $(X_i)$ est une suite de variables aléatoires indépendantes et de même loi, admettant une espérance $m$ et une variance $\sigma^2$, alors, en posant $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, on a :*

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \underset{Loi}{\rightsquigarrow} N(0,1)$$

*Le théorème implique que si $n$ est "suffisamment grand" alors on peut approcher la loi de $\sqrt{n}\frac{\bar{X}_n - m}{\sigma} = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ par la loi normale standard $N(0,1)$ (cette approximation est particulièrement utile pour le calcul des intervalles de confiance).*

Soit $(X_i)$ une suite de variables aléatoires indépendantes suivant la loi $U_{[0,1]}$. D'après les rappels ci-dessus $E(X_i) = \frac{1}{2}$ et $V(X_i) = \frac{1}{12}$ pour tout $i$, donc d'après le théorème central limite :

$$\sqrt{n}\frac{\bar{X}_n - \frac{1}{2}}{\sqrt{\frac{1}{12}}} = \sqrt{12n}\left(\bar{X}_n - \frac{1}{2}\right) \underset{Loi}{\rightsquigarrow} N(0,1)$$

On peut donc approcher la loi de $\sqrt{12n}\left(\bar{X}_n - \frac{1}{2}\right)$ par $N(0,1)$ pour n "suffisamment grand".

Montrons qu'accepter la procédure de simulation proposée revient à considérer que l'approximation de la loi de $\sqrt{12n}\left(\bar{X}_n - \frac{1}{2}\right)$ par $N(0,1)$ est valable pour $n = 12$. En effet si $n = 12$ :

$$\sqrt{12n}\left(\bar{X}_n - \frac{1}{2}\right) = \sqrt{12^2}\left(\bar{X}_n - \frac{1}{2}\right) = 12\left(\bar{X}_n - \frac{1}{2}\right) = \sum_{i=1}^{12} X_i - 6$$

La procédure proposée consiste à approcher la loi de $\sum_{i=1}^{12} X_i - 6$ par $N(0,1)$ (ou plutôt $N(0,1)$ par la loi de $\sum_{i=1}^{12} X_i - 6$, ce qui, bien entendu, revient au même). Elle consiste donc à approcher la loi de $\sqrt{12n}\left(\bar{X}_n - \frac{1}{2}\right)$ par $N(0,1)$ pour $n = 12$, ce qui revient à considérer que $n = 12$ est "suffisamment grand" pour que cette approximation soit valable. Savoir si cette approximation donne de bons résultats ou pas est une question essentiellement empirique (il semblerait que ce soit le cas puisqu'il s'agit d'une procédure usuelle d'après l'énoncé). Le raisonnement que nous venons de faire permet toutefois de donner un fondement théorique à la procédure.

**Exercice** 11

Supposons qu'on dispose d'un échantillon $X_1, X_2, ..., X_n$.

- La moyenne de l'échantillon est $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Pour calculer la médiane de l'échantillon, on commence par classer les $X_i$ par ordre croissant :

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

- Si $n$ est impair, alors la médiane de l'échantillon est $X_{\left(\frac{n+1}{2}\right)}$
- Si $n$ est pair alors la médiane de l'échantillon est $\frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right)$

- La variance de l'échantillon est donnée par $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ et la variance corrigée de l'échantillon est $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

- L'écart-type de l'échantillon est donné par $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$ et l'écart-type corrigé de l'échantillon est $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$

Remarque : la variance empirique $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ est un estimateur biaisé de la variance (théorique) qu'on note $\mu^2$ alors que la variance empirique corrigée $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ est un estimateur sans biais de $\mu^2$ .

On applique ces formules à l'échantillon dont on dispose :

- La moyenne de l'échantillon est : $\frac{1}{10}\sum_{i=1}^{10}X_i = 1,51$

- Lorsqu'on classe les $X_i$ par ordre croissant, on obtient :

$$0,2;\ 0,4;\ 0,5;\ 1,3;\ 1,3;\ 1,8;\ 1,8;\ 2,1;\ 2,5;\ 3,2$$

La médiane de l'échantillon est donc la moyenne de $1,3$ et $1,8$ c'est-à-dire $1,55$.

- La variance de l'échantillon est $0,8409$ et la variance corrigée de l'échantillon est $0,9343$.

- L'écart-type de l'échantillon est $\sqrt{0,8409} = 0,9170$ et l'écart-type corrigé de l'échantillon est $\sqrt{0,9343} = 0,9666$