UNIVERSITÉ NATIONALE DES SCIENCES, TECHNOLOGIES, INGÉNIERIE ET MATHÉMATIQUES (UNSTIM)

ÉCOLE NATIONALE SUPÉRIEURE DE GÉNIE MATHÉMATIQUE ET MODÉLISATION (ENSGMM)

****

TRAVAUX PRATIQUES

Unité d'enseignement : STATISTIQUE INFERENTIELLE

Membres :

- AKAKPO Expéra
- HOUNSOUNOU Marius
- SOGBEDJI Prosper

Sous la supervision :
Dr Nicodème ATCHADE

2025-2026

Table des matières

SECTION A

1 A1.

On a obtenu un échantillon de taille $n = 20$ d'une distribution particulière. Les données ont été saisies dans R. Une sortie R est enregistrée ci-dessous.

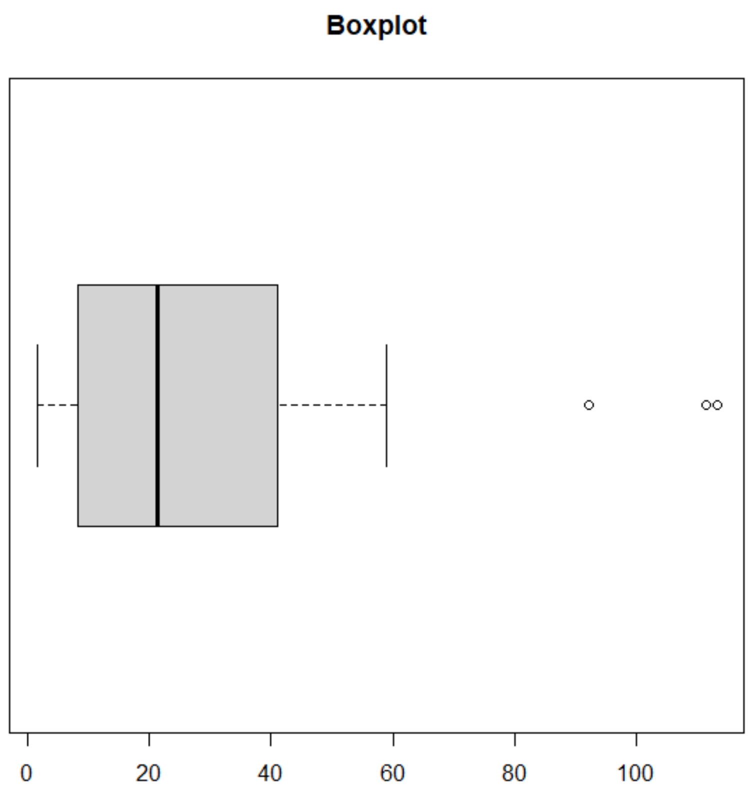
```
> x
[1]  1.49  1.67  2.20  3.23  7.32  9.28 10.35 11.85 13.67 19.79
[11] 22.95 29.07 36.80 36.81 38.20 43.81 59.00 92.17 111.62 113.32

> quantile(x, type=6)
     0%      25%      50%      75%     100%
  1.490   7.810   21.370   42.4075 113.320
```

(a)

(i) Boxplot des données.

```
boxplot(x, horizontal=TRUE, main="Boxplot")
```



(ii) Commentaire sur la forme de la distribution et sur tout autre caractéristique des données.

Le box-plot révèle une distribution fortement asymétrique à droite, avec une queue droite très longue et lourde. La médiane (21,37) est nettement plus proche du premier quartile (7,81) que du troisième quartile (42,4075), ce qui confirme un coefficient d'asymétrie positif marqué. L'écart interquartile

$$\mathrm{IQR} = 42{,}4075 - 7{,}81 = 34{,}5975$$

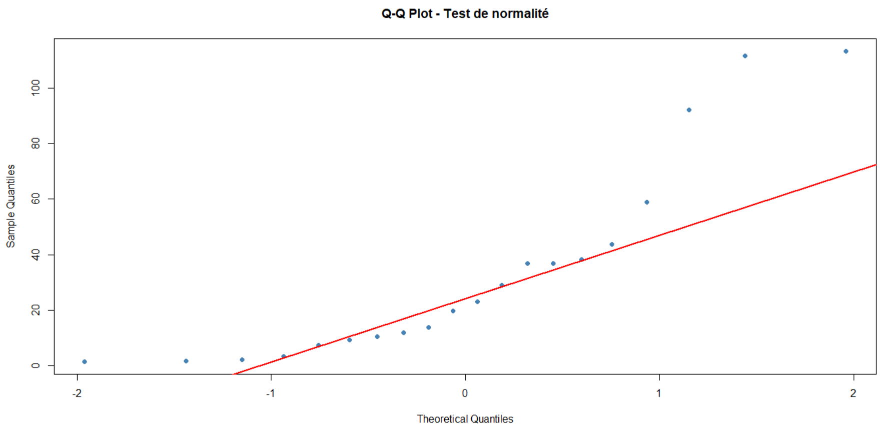
est élevé, traduisant une dispersion déjà très importante des 50 % centraux des observations. Enfin, des valeurs aberrantes (92.17 ; 111.62 et 113.32) sont détectées loin à droite, bien au-delà de la borne supérieure de Tukey renforçant le caractère lourd de la queue droite.

(iii) Modèle pour ajuster une distribution normale des données.

Il n'est pas approprié d'ajuster un modèle $N(\mu,\sigma^2)$ à ces données car la distribution est clairement asymétrique et contient des valeurs aberrantes.

Comme le montre le Q-Q Plot ci-après, les données ne sont pas proches de la droite, donc on ne peut pas ajuster un modèle normal à ces données.

```
qqnorm(x, pch=19, col="steelblue", main="Q-Q Plot -Test de normalite")
qqline(x, col="red", lwd=2)
```



Faisons un test de Shapiro-Wilk pour confirmer cela.

```
shapiro.test(x)$p.value
```

Output :

```
[1] 0.001097833
```

On constate que la *p-value* est inférieure à 0.05, l'hypothèse de normalité des données est donc rejetée. Une transformation couramment utilisée pour traiter ce type de distribution est la transformation **log**. Nous pouvons donc ajuster **un modèle log-normal** à ces données en transformant d'abord les données avec la fonction log() puis en ajustant un modèle normal aux données transformées.

En refaisant un test de Shapiro-Wilk,

```
shapiro.test(log(x))$p.value
```

Output :

```
[1]  0.2644672
```

On constate que la *p-value* est bien supérieure à 0.05. Ce qui nous permet d'affirmer que les données sont distribuées normalement lorsqu'on transforme les données avec la fonction log()

La transformation qui peut permettre d'ajuster une distribution normale est donc la transformation *log-normale*.

(b)

Supposons maintenant que l'intervalle $[0, 120]$ est divisé en classes de longueur égale $h = 20$, qui sont utilisées pour créer un histogramme de densité.

(i) Calcul de la valeur de la fonction $\mathrm{Hist}(x)$ qui définit la hauteur de l'histogramme, à $x = 9$.

La hauteur d'un histogramme de densité est donnée par la formule :

$$\mathrm{Hist}(x) = \frac{n_i}{n\,h},$$

où n_i est l'effectif de la classe contenant x, n le nombre total d'observations et h la largeur des classes.

Comme $x = 9$, il appartient à la classe :

$$[0, 20[.$$

En comptant les valeurs dans cette classe, on obtient :

$$n_i = 10.$$

De plus, on a $n = 20$ et $h = 20$. Ainsi :

$$\mathrm{Hist}(9) = \frac{10}{20 \times 20} = 0.025.$$

$$\boxed{\mathrm{Hist}(9) = 0.025}$$

2 A2.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de variables aléatoires telles que

$$X_i \sim \mathcal{P}(\lambda).$$

(i) Montrons que la fonction de vraisemblance pour l'échantillon peut s'écrire

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}$$

$$X_i \sim \mathcal{P}(\lambda).$$

La fonction de probabilité de la loi de Poisson est, pour tout $k \in \mathbb{N}$,

$$\mathbb{P}(X_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

La fonction de vraisemblance pour l'échantillon $(X_1, \dots, X_n)$ est définie comme le produit des probabilités ponctuelles évaluées en $X_i = x_i$:

$$L(\lambda) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} && (1)\\
&= \frac{\displaystyle\prod_{i=1}^{n} e^{-\lambda} \prod_{i=1}^{n} \lambda^{x_i}}{\displaystyle\prod_{i=1}^{n} x_i!} && (2)\\
&= \frac{e^{-\sum_{i=1}^{n}\lambda}\ \lambda^{\sum_{i=1}^{n} x_i}}{\displaystyle\prod_{i=1}^{n} x_i!} && (3)\\
&= \frac{e^{-n\lambda}\ \lambda^{\sum_{i=1}^{n} x_i}}{\displaystyle\prod_{i=1}^{n} x_i!}. && (4)
\end{aligned}$$

Ainsi la fonction de vraisemblance peut s'écrire sous la forme

$$\boxed{L(\lambda) = \frac{e^{-n\lambda}\ \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}}.$$

(ii) Montrons que l'estimateur du maximum de vraisemblance de λ est $\hat{\lambda} = \bar{X}$.

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}$$

$$\begin{aligned}
\ln L(\lambda) &= \ln\left(e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}\right) - \ln\left(\prod_{i=1}^{n} X_i!\right)\\
&= \ln e^{-n\lambda} + \ln \lambda^{\sum_{i=1}^{n} X_i} - \sum_{i=1}^{n} \ln X_i!\\
&= -n\lambda + \sum_{i=1}^{n} X_i \ln \lambda - \sum_{i=1}^{n} \ln X_i!
\end{aligned}$$

$$\frac{\partial}{\partial\lambda} \ln L(\lambda) = -n + \frac{\sum_{i=1}^{n} X_i}{\lambda}$$

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\ln L(\lambda) = 0 &\implies -n + \frac{\sum_{i=1}^{n} X_i}{\hat{\lambda}} = 0 \\
&\implies \frac{\sum_{i=1}^{n} X_i}{\hat{\lambda}} = n \\
&\implies \hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}
\end{aligned}$$

On a :

$$\frac{\partial^2}{\partial\lambda^2}\ln L(\lambda) = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2} < 0$$

Ainsi, la valeur maximale de la vraisemblance est atteinte pour :

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

D'où l'estimateur du maximum de vraisemblance de λ est :

$$\boxed{\hat{\lambda} = \bar{X}}$$

(iii) Biais et la variance de $\hat{\lambda}$

$$\begin{aligned}
\text{biais}(\hat{\lambda}) &= \mathbb{E}[\hat{\lambda}] - \lambda \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \lambda \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] - \lambda \\
&= \frac{1}{n}\sum_{i=1}^{n} \lambda - \lambda \quad (\text{car } X_i \sim \text{Po}(\lambda)) \\
&= \frac{n\lambda}{n} - \lambda \\
&= \lambda - \lambda \\
\text{biais}(\hat{\lambda}) &= 0
\end{aligned}$$

$$\boxed{\text{biais}(\hat{\lambda}) = 0}$$

$$\begin{aligned}
\text{Var}(\hat{\lambda}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \lambda \quad (\text{car } X_i \sim \text{Po}(\lambda))
\end{aligned}$$

$$= \frac{n\lambda}{n^2}$$
$$\mathrm{Var}(\hat{\lambda}) = \frac{\lambda}{n}$$

$$\boxed{\mathrm{Var}(\hat{\lambda}) = \frac{\lambda}{n}}$$

(iv) Probabilité que $9{,}9 < \bar{X} < 10{,}1$ lorsque $n = 100$ et $\lambda = 10$.

Soit $X_1, \ldots, X_{100}$ un échantillon i.i.d. de loi $\mathcal{P}(\lambda)$ avec $\lambda = 10$. On sait que

$$\mathbb{E}(\bar{X}) = \lambda = 10, \quad \mathrm{Var}(\bar{X}) = \frac{\lambda}{n} = \frac{10}{100} = 0{,}1.$$

Comme $n = 100$ est suffisamment grand , le théorème central limite (TCL) s'applique et l'on a l'approximation suivante :

$$\bar{X} \overset{\text{approx}}{\sim} \mathcal{N}\left(\lambda,\ \frac{\lambda}{n}\right) = \mathcal{N}\left(10,\ 0{,}1\right).$$

On a donc :

$$P(9{,}9 < \bar{X} < 10{,}1) = P\left(\frac{9{,}9-10}{\sqrt{0{,}1}} < \frac{\bar{X}-10}{\sqrt{0{,}1}} < \frac{10{,}1-10}{\sqrt{0{,}1}}\right) = P\left(\frac{-0{,}1}{\sqrt{0{,}1}} < Z < \frac{0{,}1}{\sqrt{0{,}1}}\right)$$

où $Z \sim \mathcal{N}(0,1)$.
donc

$$P(9{,}9 < \bar{X} < 10{,}1) = P(-0{,}316 < Z < 0{,}316) = 2\,\Phi(0{,}316) - 1.$$

$$\Phi(0{,}316) \approx 0{,}62395 \quad \Rightarrow \quad 2\Phi(0{,}316) - 1 \approx 0{,}2479.$$

Finalement

$$\boxed{P(9{,}9 < \bar{X} < 10{,}1) \approx 0{,}2479}$$

3 A3.

Soient

$$X_{1i} \sim \mathcal{N}\left(\mu_1, \sigma^2\right)$$
$$X_{2j} \sim \mathcal{N}\left(\mu_2, \sigma^2\right)$$

(i) Trouvons un estimateur sans biais $\hat{\sigma}^2$ de la variance commune σ^2

Méthode du maximum de vraissemblance

Pour la variable aléatoire suivant la loi $\mathcal{N}(\mu_1, \sigma^2)$, considérons S_1^2 un estimateur sans biais de σ^2. On a :

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_n\right)^2 \tag{5}$$

Pour la variable aléatoire suivant la loi $\mathcal{N}(\mu_2, \sigma^2)$, considérons S_2^2 unbiased l'un estimateur sans biais de σ^2

$$S_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(X_{2j} - \bar{X}_m\right)^2 \tag{6}$$

Donc un estimateur sans biais $\hat{\sigma}^2$ de la variance commune σ^2, est :

$$\boxed{\hat{\sigma}^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}} \tag{7}$$

<u>Distribution d'une version mise à l'échelle $\hat{\sigma}^2$</u>

$$\boxed{(n+m-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n+m-2}} \tag{8}$$

(ii) <u>Statistique de test et loi d'échantillonnage sous H_0</u>
Sous H_0, $\bar{X}_n - \bar{X}_m \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$ indépendement de $(n+m-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n+m-2}$ alors une statistique appropriée (test de Student bilatéral, variances égales) est :

$$T = \frac{\bar{X}_n - \bar{X}_m}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \tag{9}$$

(iii) <u>Région de rejet approprié pour ce test</u>
Pour un test bilatéral au niveau α la région de rejet est

$$\boxed{W = \{|T| > t_{n+m-2,\frac{\alpha}{2}}\}} \tag{10}$$

où α est le risque d'erreur et $t_{n+m-2,\frac{\alpha}{2}}$ est le quantile α de la loi de student de dégré de liberté $n+m-2$

(iv) <u>Application numérique</u> $n = m = 10$, $\bar{x}_1 = 46.0$ $\bar{x}_2 = 48.1$ $s_1^2 = 2.04^2$ $s_2^2 = 4.92^2$

$$\hat{\sigma}^2 = \frac{9[2.04^2 + 4.92^2]}{18} \tag{11}$$

$$\hat{\sigma}^2 = 3.924 \tag{12}$$

$$T = \frac{46.0 - 48.1}{\sqrt{3.924 \times \frac{2}{10}}} \tag{13}$$

$$= -2.37 \tag{14}$$

$$\boxed{t_{\text{obs}} = -2.37} \tag{15}$$

- Pour $\alpha = 0.05$
  $t_{18,0.025} = 2.1009$, $|T| > t_{18,0.025}$. donc on rejette H_0 au niveau $\alpha = 0.05$.
- Pour $\alpha = 0.01$
  $t_{18,0.005} = 2.8784$, $|T| < t_{18,0.005}$. donc on ne rejette pas H_0 au niveau $\alpha = 0.01$.

Avec les données fournies, il y a une différence significative entre μ_1 et μ_2 au niveau $\alpha = 0.05$, mais pas au niveau $\alpha = 0.01$.

4 A4.

Echantillon indépendant de taille $n = 1000$; proportion observée soutenant Labour $p = 0.30$. On suppose les observations Bernoulli i.i.d.

(i) Formule générale d'un intervalle de confiance approximatif à $100(1-\alpha)\%$ pour p

Pour un échantillon de taille n et proportion empirique $\hat{p}$, un intervalle de confiance approché (par la CLT) à niveau $100(1-\alpha)\%$ est

$$\boxed{IC_p = \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]} \tag{16}$$

où $z_{1-\alpha/2}$ est le quantile $1-\alpha/2$ de la loi normale standard

Commentaire distributionnel : cette construction repose sur l'approximation par la loi normale de la variable $\hat{p}$ (par le théorème central limite) : $\hat{p} \approx N(p, p(1-p)/n)$ pour n grand. L'approximation est raisonnable si np et $n(1-p)$ sont suffisamment grands (classiquement ≥ 5 ou mieux ≥ 10).

(ii) Intervalle de confiance à 95% pour la proportion

Nous avons

$$\boxed{IC_p = p \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{p\,(1-p)}{n}}}$$

avec $p = 0.3$ et $1 - \frac{\alpha}{2} = 0.975$. $z_{1-\frac{\alpha}{2}} = 1.96$, alors $IC = 0.3 \pm 0.0284$.

$$\boxed{IC_p = [0.2716; 0.3284]}$$

(iii) Probabilité qu'au moins 150 personnes soutiennent Labour dans un nouvel échantillon

Nous avons $X \backsim \mathcal{N}\,(np, np(1-p))$

$$\begin{aligned}
\mathbb{P}\,(X \geq 150) &= \mathbb{P}\left(Z \geq \frac{150 - np}{\sqrt{np(1-p)}}\right) \quad \text{avec } Z \backsim \mathcal{N}\,(0,1) \\
&= \mathbb{P}\left(Z \geq \frac{150 - np}{\sqrt{np(1-p)}}\right) \\
&= \mathbb{P}\,(Z \geq 0.99) \\
&= 1 - \phi\,(0.99) \\
\mathbb{P}\,(X \geq 150) &= 0.1611
\end{aligned}$$

$$\boxed{\mathbb{P}\,(X \geq 150) = 0.1611}$$

SECTION B

5 B5.

(a)

Soit $(X_1, \ldots, X_n)$ un échantillon aléatoire de taille n issu d'une population $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont **tous deux inconnus**. On définit l'estimateur de la variance :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(i) Montrons que $\mathbb{E}[S^2] = \sigma^2$

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right] \\
&= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \\
&= \frac{(n-1)\sigma^2}{n-1} \\
\mathbb{E}[S^2] &= \sigma^2.
\end{aligned}$$

$$\boxed{\mathbb{E}[S^2] = \sigma^2}$$

(ii) Distribution de $(n-1)S^2/\sigma^2$

$X_1, \ldots, X_n \overset{\text{ind}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et S^2 est un estimateur non biaisé de σ^2 , alors :

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}},$$

où χ^2_{n-1} désigne la loi du **khi-deux à** $(n-1)$ **degrés de liberté**.

(b) Données observées

$$n = 10, \quad \sum_{i=1}^{n} x_i = 113{,}20, \quad \sum_{i=1}^{n} x_i^2 = 1474{,}5.$$

(i) Calculons s^2 et l'intervalle de confiance à 99% pour σ^2

Calcul de s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right).$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{113{,}20}{10} = 11{,}32.$$

$$s^2 = \frac{1}{9} \left(1474.5 - 10(11.32)^2 \right)$$

$$\boxed{s^2 = 21{,}45}.$$

Intervalle de confiance à 99% pour σ^2

Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(\mu, \sigma^2)$ est :

$$IC(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \,;\, \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right],$$

où $\chi^2_{n-1,p}$ est le quantile d'ordre p de la loi du khi-deux à $n-1$ degrés de liberté.

On a : $n = 10$, $\alpha = 1\%$; $s^2 = 21{,}45$
$\chi^2_{9,0.005} = 23.589, \quad \chi^2_{9,0.995} = 1{,}734$

$$IC(\sigma^2) = \left[\frac{9 \times 21.45}{\chi^2_{9,\,0.005}} \,;\, \frac{9 \times 21.45}{\chi^2_{9,\,0.995}} \right] = \left[\frac{193.05}{23.589} \,;\, \frac{193.05}{1.734} \right] = [\,8.18 \,;\, 111.33\,].$$

$$\boxed{IC(\sigma^2) = [\,8.18 \,;\, 111.33\,].}$$

(ii) Intervalle de confiance à 99% pour μ

Un intervalle de confiance de seuil α pour le paramètre μ de la loi $\mathcal{N}(\mu, \sigma^2)$ est :

$$IC(\mu) = \left[\bar{X}_n - \frac{s_n}{\sqrt{n}}\, t_{n-1,\,\alpha} \,;\, \bar{X}_n + \frac{s_n}{\sqrt{n}}\, t_{n-1,\,\alpha} \right],$$

où $t_{n-1,\,\alpha}$ est le quantile d'ordre α de la loi de Student à $n-1$ degrés de liberté.
On a : $n = 10$; $\bar{X}_n = 11{,}32$; $s_n = 4{,}63$; $t_{9,0.01} = 3{,}25$.

$$\boxed{IC(\mu) = [\,6.56 \,;\, 16.08\,]}.$$

(iii) Probabilité estimée que la moyenne d'un futur échantillon de taille $n = 10$ satisfasse $\bar{X} > 11{,}0$

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

$$\begin{aligned}
\mathbb{P}(\bar{X} > 11) &= \mathbb{P}\left(Z > \frac{11 - 11.32}{\sqrt{\frac{21.45}{10}}}\right) \\
&= \mathbb{P}(Z > -0.218) \\
&= 1 - \mathbb{P}(Z \leq -0.218) \\
&= 1 - (1 - \Phi(0.218)) \\
&= \Phi(0.218)
\end{aligned}$$

$$\boxed{\mathbb{P}(\bar{X} > 11) = 0.5865}$$

6 B6.

Un chercheur mène un essai clinique sur $n = 60$ patients afin d'étudier un nouveau traitement contre le rhinovirus. Soit $X_i \overset{\text{ind}}{\sim} \mathcal{N}(6, 2^2)$ la variable aléatoire représentant le délai de guérison (en jours) du i-ème patient non traité, pour $i = 1, \ldots, 60$.

On observe que 52 patients sur 60 se rétablissent dans les 7 jours sous traitement.

Soit p la probabilité qu'un patient se rétablisse dans les 7 jours sous le nouveau traitement et p_0 la probabilité qu'un patient se rétablisse dans les 7 jours sans aucun traitement.

(i) Montrons que $p_0 = \Phi(1/2)$ et calculer sa valeur numérique

$$p_0 = P(X_i \leq 7) = P\left(\frac{X_i - 6}{2} \leq \frac{7 - 6}{2}\right) = P\left(Z \leq \frac{1}{2}\right),$$

où $Z \sim \mathcal{N}(0, 1)$. Ainsi :

$$\boxed{p_0 = \Phi\left(\frac{1}{2}\right)}.$$

D'après la table de la loi normale centrée réduite :

$$\Phi(0.5) \approx 0.6915.$$

Donc :

$$\boxed{p_0 \approx 0.6915}.$$

(ii) Écrivons une statistique de test appropriée pour tester

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p > p_0$$

Les observations sont indépendantes et $n = 60 \geq 30$. De plus :

$$np_0 \approx 60 \times 0.6915 = 41.49 > 5, \quad n(1 - p_0) \approx 18.51 > 5.$$

Les conditions d'application du test binomial approximé par la loi normale sont vérifiées.

Soit $k \sim \mathcal{B}(n, p)$ le nombre de succès (guérison en ≤ 7 jours). L'estimateur $\hat{p} = k/n$ est le maximum de vraisemblance de p. Sous H_0, $k \sim \mathcal{B}(n, p_0)$, donc par le théorème central limite (TCL) :

$$\hat{p} \overset{\text{approx}}{\sim} \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right).$$

On pose la statistique de test :

$$\boxed{Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}}.$$

Il s'agit d'un **test unilatéral à droite** (car $H_1 : p > p_0$).

(iii) Proposons une distribution approximative pour Z sous H_0 .

Sous les hypothèses d'indépendance, de taille d'échantillon suffisante ($n \geq 30$) et de conditions $np_0 > 5$, $n(1-p_0) > 5$, on a :

$$\boxed{Z \overset{\text{approx}}{\sim} \mathcal{N}(0, 1) \quad \text{sous } H_0}.$$

Les conditions numériques sont largement satisfaites, donc l'approximation est **fondée** .

(iv)-Région de rejet appropriée pour le test afin d'atteindre le niveau de significativité $\alpha = 0.05$.

La région de rejet appropriée pour atteindre exactement le niveau α est :

$$\boxed{W = \{Z > z_\alpha\}},$$

où Z est la statistique de test que nous avons définie précédemment et $z_\alpha = 1.645$ à partir de la table des quantiles de la loi normale centrée réduite .

La région de rejet est donc :

$$\boxed{W = \left\{ \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > 1{,}645 \right\}}.$$

Ainsi , H_0 est rejettée si :

$$\hat{p} \geq p_0 + 1.645\sqrt{\frac{p_0(1-p_0)}{n}}$$

- Rejet ou non de H_0 si 52 patients sur 60 se sont rétablis dans les 7 jours

$$\hat{p} = \frac{52}{60} = 0.8667, \quad p_0 + 1.645\sqrt{\frac{p_0(1-p_0)}{n}} = 0.7896$$

donc H_0 **est rejettée**

On peut affirmer au risque de 5 % que la proportion de patients qui guérissent avec le nouveau traitement est plus élevée que la proportion de patients qui guérissent sans traitement.

(v)- Probabilité de rétablissement sous le nouveau traitement

$X_i \overset{\text{ind}}{\sim} \mathcal{N}(5, 2^2)$ la variable aléatoire représentant le temps de guérison (en jours) du i-ème patient **sous le nouveau traitement** , pour $i = 1, \ldots, 60$

Démonstration.

$$p = P(X_i \leq 7) = P\left(\frac{X_i - 5}{2} \leq \frac{7-5}{2}\right) = P\left(Z \leq 1\right),$$

où $Z \sim \mathcal{N}(0, 1)$ par standardisation. Ainsi :

$$p = \Phi\left(1\right)$$

Donc :

$$\boxed{p = 0.8413}.$$

□

- Probabilité approximative de rejetter l'hypothèse nullle

$P(\hat{p} \geq 0.7896) = P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0.7896-0.8413}{\sqrt{\frac{0.8413\times0.1587}{60}}}\right)$
$P(\hat{p} \geq 0.7896) = 1 - \phi(-1.10)$
Donc :

$$\boxed{P(\hat{p} \geq 0.7896) = 0.86}$$

La probabilité approximative de rejetter l'hypothèse nullle est de 0.86.

7 B7.

(a)

Soit $(X_1, \ldots, X_n)$ un échantillon aléatoire de taille n issu de $\mathcal{N}(\mu, \sigma^2)$. On considère deux estimateurs de μ :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \tilde{\mu} = \frac{1}{2n}\sum_{i=1}^{n} X_i.$$

Soit $\varepsilon > 0$ une quantité fixée par l'expérimentateur. On suppose que, inconnu de l'expérimentateur $\mu = 0{,}2\sigma$ et $\varepsilon = 0{,}1\sigma$.

(i) Biais et variance de $\hat{\mu}$

Biais de $\hat{\mu}$

$$\begin{aligned}
\text{biais}(\hat{\mu}) &= \mathbb{E}(\hat{\mu}) - \mu \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) - \mu \\
&= \frac{1}{n}\, n\, \mathbb{E}(X) - \mu \\
&= \mu - \mu \qquad \text{car } X \sim \mathcal{N}(\mu, \sigma^2) \text{ et } \mathbb{E}(X) = \mu \\
&= 0
\end{aligned}.$$

$$\boxed{\text{biais}(\hat{\mu}) = 0}.$$

Variance de $\hat{\mu}$

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) \qquad \text{(car les } X_i \text{ sont indépendants)} \\
&= \frac{1}{n^2}\, n\ \text{Var}(X) \qquad \text{(car les } X_i \text{ sont i.i.d.)} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

$$\boxed{\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}}.$$

(ii) Biais et variance de $\tilde{\mu}$

Biais de $\tilde{\mu}$

$$\begin{aligned}
\operatorname{Biais}(\tilde{\mu}) &= \mathbb{E}[\tilde{\mu}] - \mu \\
&= \mathbb{E}\left[\frac{1}{2n}\sum_{i=1}^{n} X_i\right] - \mu \\
&= \frac{1}{2n}\sum_{i=1}^{n} \mathbb{E}[X_i] - \mu \\
&= \frac{1}{2n} \cdot n \cdot \mathbb{E}[X] - \mu \\
&= \frac{1}{2}\mu - \mu \\
&= -\frac{1}{2}\mu.
\end{aligned}$$

$$\boxed{\operatorname{Biais}(\tilde{\mu}) = -\frac{1}{2}\mu}$$

Variance de $\tilde{\mu}$

$$\begin{aligned}
\operatorname{Var}(\tilde{\mu}) &= \operatorname{Var}\left(\frac{1}{2n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{4n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i) \\
&= \frac{1}{4n^2} \cdot n \cdot \operatorname{Var}(X) \\
&= \frac{1}{4n}\sigma^2.
\end{aligned}$$

$$\boxed{\operatorname{Var}(\tilde{\mu}) = \frac{\sigma^2}{4n}}$$

(b)

Soit ε une quantité spécifiée par l'expérimentateur. Supposons de plus que $\mu = 0{,}2\sigma$ et $\varepsilon = 0{,}1\sigma$.

(i) Montrons que $\hat{\mu}$ est à moins de ε de la vraie valeur de μ avec une probabilité $p_1(n) = 2\Phi(0{,}1\sqrt{n}) - 1$.

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Ainsi :

$$\begin{aligned} P(|\hat{\mu}-\mu|<\varepsilon) &= P\left(-\varepsilon<\hat{\mu}-\mu<\varepsilon\right) \\ &= P\left(-\frac{\varepsilon\sqrt{n}}{\sigma}<\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}<\frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)-1. \end{aligned}$$

Avec $\varepsilon = 0{,}1\sigma$, on obtient :

$$\boxed{P(|\hat{\mu}-\mu|<\varepsilon) = 2\Phi(0{,}1\sqrt{n})-1}.$$

(ii) Montrons que $\tilde{\mu}$ est à moins de ε de la vraie valeur de μ avec une probabilité $p_2(n) = \Phi(0{,}4\sqrt{n}) - 0{,}5$.

$$\tilde{\mu} \sim \mathcal{N}\left(\frac{\mu}{2}, \frac{\sigma^2}{4n}\right).$$

Ainsi :

$$\begin{aligned} P(|\tilde{\mu}-\mu|<\varepsilon) &= P\left(-\varepsilon<\tilde{\mu}-\mu<\varepsilon\right) \\ &= P\left(-\varepsilon+\frac{\mu}{2}<\tilde{\mu}-\frac{\mu}{2}<\varepsilon+\frac{\mu}{2}\right) \\ &= P\left(\frac{-\varepsilon+\mu/2}{\sigma/(2\sqrt{n})}<\frac{\tilde{\mu}-\mu/2}{\sigma/(2\sqrt{n})}<\frac{\varepsilon+\mu/2}{\sigma/(2\sqrt{n})}\right). \end{aligned}$$

Avec $\mu = 0{,}2\sigma$ et $\varepsilon = 0{,}1\sigma$, on a :

$$-\varepsilon+\frac{\mu}{2} = 0, \quad \varepsilon+\frac{\mu}{2} = 0{,}2\sigma,$$

alors :

$$\begin{aligned} P(|\tilde{\mu}-\mu|<\varepsilon) &= P\left(0<Z<0{,}4\sqrt{n}\right) \\ &= \Phi(0{,}4\sqrt{n})-\Phi(0) \\ &= \Phi(0{,}4\sqrt{n})-0{,}5 \end{aligned}$$

donc :

$$\boxed{P(|\tilde{\mu}-\mu|<\varepsilon) = \Phi(0{,}4\sqrt{n})-0{,}5}.$$

(iii) Estimateur qui a la plus grande probabilité d'être à moins de ε de la vraie valeur de μ quand $n = 10$.

Pour $n = 10$, on a :

$$p_1(10) = 2\Phi(0{,}1\sqrt{10})-1 \approx 0{,}2481$$

et

$$p_2(10) = \Phi(0{,}4\sqrt{10})-0{,}5 \approx 0{,}3970.$$

On peut constater que $p_2(10) > p_1(10)$. Cela signifie que l'estimateur $\tilde{\mu}$ a une plus grande probabilité d'être à moins de ε de la vraie valeur de μ lorsque $n = 10$ que l'estimateur $\hat{\mu}$.

(iv) L'investigateur décide que l'expérience sera considérée comme un succès si et seulement si l'estimation de μ est à moins de ε de la vraie valeur.

Choix de l'estimateur : La probabilité d'un succès pour l'expérience est donc la probabilité que l'estimation de μ soit à moins de ε de la vraie valeur, c'est-à-dire $p_1(n)$ ou $p_2(n)$ selon l'estimateur choisi.

D'après le graphe donné en indice, on peut constater que $p_2(n)$ est plus grande que $p_1(n)$ pour des valeurs plus petites de n ($n < 50$). Ainsi, $\tilde{\mu}$ est préférable à $\hat{\mu}$ en termes de probabilité d'un succès pour l'expérience dans ce cas.
Mais lorsque la taille de l'échantillon est assez grande, le choix est orienté vers l'estimateur $\hat{\mu}$ car $p_1(n)$ est plus grande que $p_2(n)$ pour n assez grand ($n > 50$).

En résumé, l'estimateur $\tilde{\mu}$ est préférable pour des tailles d'échantillon assez petites ($n < 50$), tandis que l'estimateur $\hat{\mu}$ est préférable pour des tailles d'échantillon assez grandes ($n > 50$).