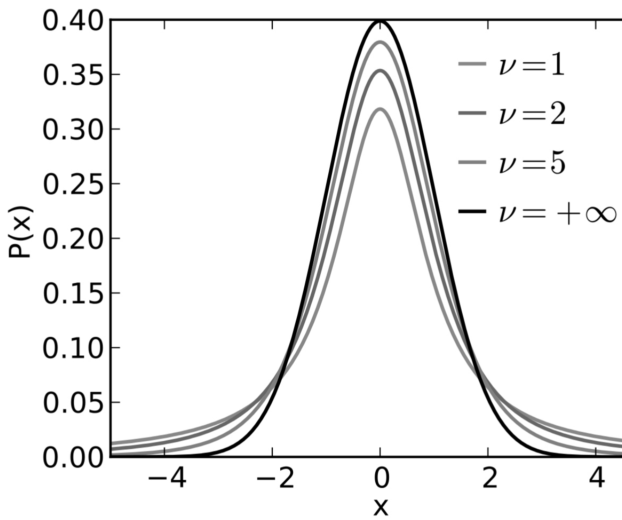
M3.5.1 Statistiques

Corrigé Examen 2015-2016



v0.2

24 mars 2017

Exercice 1. Un contrôle final a été effectué sur une production de 5000 lampes fluorescentes, 14 jours après la fabrication pour en vérifier des fuites possibles. 100 lampes on été choisies au hasard. Sur les 100 lampes observées, 8 ont présenté une fuite de gaz.

(a) Quelle est l'estimation ponctuelle de la vraie proportion de lampes de la production qui présentent une fuite de gaz ?

(b) Estimer par intervalle de confiance la proportion de lampes de toute la production qui présentent une fuite de gaz au niveau de confiance de 95%.

Exercice 2. La durée de vie d'une batterie est une variable aléatoire X de densité

$$f_\theta(x) = \begin{cases} \theta^{-1}\, \mathrm{e}^{-\frac{x}{\theta}} & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

(a) Montrer que le temps moyen de bon fonctionnement d'une batterie est égal à θ.

(b) Afin d'établir une durée de garantie, le fabriquant cherche à estimer le temps moyen de bon fonctionnement. Étudier l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ.

Exercice 3. Un bureau d'étude conseille ces clients sur les techniques d'échantillonnage et les procédures statistiques à utiliser pour contrôler un processus de production. Dans un cas particulier, un client a fourni au bureau d'étude un échantillon de 800 observations sélectionnées à un moment donné, au cours duquel le processus de production était satisfaisant. L'écart-type de l'échantillon était égal à 0.21 ; par conséquent, l'écart-type de la production était supposé égal à 0.21. Le bureau d'étude suggéra alors que des échantillons aléatoires de taille 30 soient sélectionnés périodiquement pour contrôler le processus en cours. En analysant les nouveaux échantillons, le client pourra apprendre rapidement si le processus est toujours satisfaisant.

Lorsque le processus ne fonctionne pas de manière satisfaisante, des actions correctives doivent être prises pour résoudre le problème. La spécification indiquait que la moyenne du processus devait être égale à 12. Le test d'hypothèses suggéré par le bureau d'étude est le suivant :

$$H_0 : \mu = 12$$
$$H_1 : \mu \neq 12$$

Une action corrective devra être prise à chaque fois que H_0 est rejetée.

L'échantillon suivant a été collecté au cours du premier jour d'exploitation de la nouvelle procédure de contrôle statistique.

11.55	11.62	11.52	11.75	11.90	11.64
11.80	12.03	11.94	11.92	12.13	12.09
11.93	12.21	12.32	11.95	11.85	11.76
12.16	11.77	12.00	12.04	11.98	12.30
12.18	11.97	12.17	11.85	12.30	12.15

(a) Effectuer un test d'hypothèse au seuil de signification de 0.01 et déterminer quelle action doit être prise.

(b) Discuter les implications d'une augmentation du seuil de signification. Quelle erreur peut augmenter si le seuil de signification est modifié ?

(c) Quelle est la probabilité de conclure que le procédé opère correctement alors qu'en réalité il est centré à 11.90 ? Comment appelle-t-on ce risque ?

(d) Donner une estimation par intervalle de confiance au niveau 98% de la moyenne théorique μ.

Exercice 1. Un contrôle final a été effectué sur une production de 5000 lampes fluorescentes, 14 jours après la fabrication pour en vérifier des fuites possibles. 100 lampes on été choisies au hasard. Sur les 100 lampes observées, 8 ont présenté une fuite de gaz.

(a) Quelle est l'estimation ponctuelle de la vraie proportion de lampes de la production qui présentent une fuite de gaz ?

Solution.

On a un échantillon de $n = 100$ lampes.
Soit

$$X_i = \begin{cases} 1 & \text{si la } i^e \text{ lampe présente une fuite} \\ 0 & \text{sinon} \end{cases}$$

la variable aléatoire qui suit la loi de Bernoulli de paramètre p, avec $1 \leq i \leq n$.
L'ESBVM de p est

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{8}{100}$$

donc l'estimation ponctuelle de la vraie proportion de lampes de production qui présentent une fuite de gaz est

$$\boxed{\hat{p} = 0.08}$$

(b) Estimer par intervalle de confiance la proportion de lampes de toute la production qui présentent une fuite de gaz au niveau de confiance de 95%.

Solution.

On a $\alpha = 5\%$.
Soit

$$Z = \sqrt{n}\frac{\bar{X} - p}{\sqrt{p(1-p)}}$$

Les X_i sont indépendants et $n \geq 30$, donc avec le TCL, Z suit approximativement la loi normale centrée réduite $\mathcal{N}(0,1)$.
En posant

$$T = n\bar{X} = n\hat{p}$$

on a

$$Z = \frac{T - np}{\sqrt{np(1-p)}}$$

Soit u_α le réel tel que

$$P(|Z| \leq u_\alpha) = 1 - \alpha$$

Pour trouver l'intervalle de confiance recherché, il suffit d'écrire $|Z| \leq u_\alpha$ sous la forme $r_1 \leq p \leq r_2$.

$$
\begin{aligned}
|Z| \leq u_\alpha &\iff \left| \frac{T - np}{\sqrt{np(1-p)}} \right| \leq u_\alpha \\
&\iff \frac{(T-np)^2}{np(1-p)} \leq u_\alpha^2 \\
&\iff \dots \\
&\iff (n + u_\alpha^2)\, p^2 - (2T + u_\alpha^2)\, p + \frac{T^2}{n} \leq 0 \\
&\iff r_1 \leq p \leq r_2
\end{aligned}
$$

avec r_1 et r_2 les racines de l'équation quadratique en p

$$(n + u_\alpha^2)\, p^2 - (2T + u_\alpha^2)\, p + \frac{T^2}{n} = 0$$

On a

$$\Delta = \dots = u_\alpha^2 \left(u_\alpha^2 + 4T \left(1 - \frac{T}{n} \right) \right)$$

donc

$$r_{1,2} = \dots = \frac{\frac{T}{n} + \frac{u_\alpha^2}{2n} \pm u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T}{n^2} \left(1 - \frac{T}{n} \right)}}{1 + \frac{u_\alpha^2}{n}}$$

En négligeant u_α devant n, on a

$$
\begin{aligned}
r_{1,2} &= \frac{T}{n} \pm u_\alpha \sqrt{\frac{T}{n^2} \left(1 - \frac{T}{n} \right)} \\
&= \hat{p} \pm u_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\end{aligned}
$$

donc un intervalle de confiance (asymptotique) de p au niveau de confiance de 95% est

$$[r_1, r_2] = \left[\hat{p} - u_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \,,\, \hat{p} + u_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Calculons u_α. On a

$$P(|Z| \leq u_\alpha) = 1 - \alpha \iff P(Z \leq u_\alpha) = 1 - \frac{\alpha}{2}$$

On a $1 - \frac{\alpha}{2} = 0.975$, donc à partir de la table de la loi normale centrée réduite, on lit $u_\alpha = 1.96$. A.N. :

$$
\begin{aligned}
r_1 = 0.08 - 1.96 \sqrt{\frac{0.08(1-0.08)}{100}} = 0.0722 \\
r_2 = 0.08 + 1.96 \sqrt{\frac{0.08(1-0.08)}{100}} = 0.0878
\end{aligned}
$$

donc un intervalle de confiance (asymptotique) de p à 95% est

$$\boxed{[0.0722, 0.0878]}$$

Exercice 2. La durée de vie d'une batterie est une variable aléatoire X de densité

$$f_\theta(x) = \begin{cases} \theta^{-1}\,\mathrm{e}^{-\frac{x}{\theta}} & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

(a) Montrer que le temps moyen de bon fonctionnement d'une batterie est égal à θ.

Solution.

Il s'agit de montrer que $E[X] = \theta$.
On a

$$\begin{aligned} E[X] &= \int_{\mathbb{R}} x f_\theta(x)\,\mathrm{d}x \\ &= \frac{1}{\theta}\int_0^{+\infty} x\,\mathrm{e}^{-\frac{x}{\theta}}\,\mathrm{d}x \end{aligned}$$

On fait une intégration par parties

$$\begin{aligned} \int_0^{+\infty} x\,\mathrm{e}^{-\frac{x}{\theta}}\,\mathrm{d}x &= \left[-\theta x\,\mathrm{e}^{-\frac{x}{\theta}}\right]_0^{+\infty} - \int_0^{+\infty} -\theta\,\mathrm{e}^{-\frac{x}{\theta}}\,\mathrm{d}x \\ &= \theta\int_0^{+\infty} \mathrm{e}^{-\frac{x}{\theta}}\,\mathrm{d}x \\ &= \theta\left[-\theta\,\mathrm{e}^{-\frac{x}{\theta}}\right]_0^{+\infty} \\ &= \theta^2 \end{aligned}$$

donc

$$E[X] = \frac{1}{\theta}\times\theta^2$$

$$\boxed{E[X] = \theta}$$

(b) Afin d'établir une durée de garantie, le fabriquant cherche à estimer le temps moyen de bon fonctionnement. Étudier l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ.

Solution.

Soit X_i la variable aléatoire représentant la durée de vie de la i^e batterie de l'échantillon. Les X_i sont supposées indépendantes et identiquement distribuées.
On a la fonction de vraisemblance pour l'échantillon $(x_1, \dots, x_n)$

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$$

L'estimateur de maximum de vraisemblance (EMV) de θ est la variable aléatoire correspondant à l'estimation $\hat{\theta}$ de θ qui rend maximale la fonction de vraisemblance. On calcule $\hat{\theta}$ en maximisant la log-vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} \ln \mathcal{L}(\theta; x_1, \dots, x_n)$$

On a

$$\begin{aligned}
\ln \mathcal{L}(\theta; x_1, \dots, x_n) &= \ln \prod_{i=1}^{n} f_{X_i}(x_i; \theta) \\
&= \sum_{i=1}^{n} \ln f_{X_i}(x_i; \theta) \\
&= \sum_{i=1}^{n} \ln \left(\frac{1}{\theta} \mathrm{e}^{-\frac{x_i}{\theta}} \right) \\
&= \sum_{i=1}^{n} - \ln \theta - \frac{x_i}{\theta} \\
&= -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i \\
&= -n \ln \theta - n\bar{x}\frac{1}{\theta} \\
&= -n \left(\ln \theta + \bar{x}\frac{1}{\theta} \right)
\end{aligned}$$

$\hat{\theta}$ est solution de l'équation de vraisemblance

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x_1, \dots, x_n) = 0$$

(A priori, cette solution peut être un minimum, mais la nature de la fonction de vraisemblance fait que c'est bien un maximum...)
On a

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x_1, \dots, x_n) = -n \left(\frac{1}{\theta} - \bar{x}\frac{1}{\theta^2} \right)$$

qui s'annule pour $\theta = \bar{x}$, donc l'EMV de θ est $\hat{\theta} = \bar{X}$.
On sait que $\bar{X}$ est l'ESBVM (estimateur sans biais et de variance minimale) de $E[X]$, donc $\hat{\theta}$ est l'ESBVM de θ.

Exercice 3. Un bureau d'étude conseille ces clients sur les techniques d'échantillonnage et les procédures statistiques à utiliser pour contrôler un processus de production. Dans un cas particulier, un client a fourni au bureau d'étude un échantillon de 800 observations sélectionnées à un moment donné, au cours duquel le processus de production était satisfaisant. L'écart-type de l'échantillon était égal à 0.21 ; par conséquent, l'écart-type de la production était supposé égal à 0.21. Le bureau d'étude suggéra alors que des échantillons aléatoires de taille 30 soient sélectionnés périodiquement pour contrôler le processus en cours. En analysant les nouveaux échantillons, le client pourra apprendre rapidement si le processus est toujours satisfaisant.

Lorsque le processus ne fonctionne pas de manière satisfaisante, des actions correctives doivent être prises pour résoudre le problème. La spécification indiquait que la moyenne du processus devait être égale à 12. Le test d'hypothèses suggéré par le bureau d'étude est le suivant :

$$H_0 : \mu = 12$$
$$H_1 : \mu \neq 12$$

Une action corrective devra être prise à chaque fois que H_0 est rejetée.

L'échantillon suivant a été collecté au cours du premier jour d'exploitation de la nouvelle procédure de contrôle statistique.

11.55	11.62	11.52	11.75	11.90	11.64
11.80	12.03	11.94	11.92	12.13	12.09
11.93	12.21	12.32	11.95	11.85	11.76
12.16	11.77	12.00	12.04	11.98	12.30
12.18	11.97	12.17	11.85	12.30	12.15

(a) Effectuer un test d'hypothèse au seuil de signification de 0.01 et déterminer quelle action doit être prise.

Solution.

On suppose que l'écart-type de la production est connu : $\sigma = 0.21$.
On veut tester H_0 : "$\mu = \mu_0$" contre H_1 : "$\mu \neq \mu_0$", avec $\mu_0 = 12$ et $\alpha = 0.01$.
La région critique du test est $W = \{(X_1, \dots, X_n); |\bar{X} - \mu_0| > l_\alpha\}$ avec $l_\alpha \in \mathbb{R}$.
La règle de rejet du test est :

- Si l'échantillon $(X_1, \dots, X_n) \in W$, on rejette H_0.
- Sinon, H_0 n'est pas rejetée.

La fonction puissance du test est

$$\begin{aligned} \beta : \mathbb{R} &\to [0, 1] \\ \mu &\mapsto \beta(\mu) = P((X_1, \dots X_n) \in W; \mu) \end{aligned}$$

($\beta(\mu)$ est la probabilité de rejeter H_0 quand la vraie valeur de la moyenne est μ.)

Le seuil du test α est la probabilité maximale de rejeter H_0 alors que H_0 est vraie :

$$\begin{aligned}
\alpha &= \sup_{H_0} \beta(\mu) \\
&= \sup_{H_0} P((X_1, \dots X_n) \in W; \mu) \\
&= \sup_{\mu=\mu_0} P(|\bar{X} - \mu_0| > l_\alpha; \mu) \\
&= P(|\bar{X} - \mu_0| > l_\alpha; \mu_0) \\
&= P(\left|\sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma}\right| > \frac{l_\alpha}{\sigma}\sqrt{n}; \mu_0) \\
&= P(|Y| > \frac{l_\alpha}{\sigma}\sqrt{n}; \mu_0)
\end{aligned}$$

en posant $Y = \sqrt{n}\,\dfrac{\bar{X} - \mu_0}{\sigma}$. On a $n = 30$ et les X_i sont supposés indépendants, donc avec le TCL, Y suit approximativement $\mathcal{N}(0, 1)$.

$$P(|Y| > \frac{l_\alpha}{\sigma}\sqrt{n}; \mu_0) = \alpha \iff P(Y < \frac{l_\alpha}{\sigma}\sqrt{n}; \mu_0) = 1 - \frac{\alpha}{2}$$

À partir de la table de la loi normale centrée réduite, on trouve $u_\alpha = 2.575$ correspondant à $1 - \dfrac{\alpha}{2} = 0.995$, avec $\dfrac{l_\alpha}{\sigma}\sqrt{n} = u_\alpha$, donc

$$l_\alpha = \frac{\sigma}{\sqrt{n}} u_\alpha$$

A.N. : $l_\alpha = \frac{0.21}{\sqrt{30}} \times 2.575 = 0.09873$.
La région critique du test est $W = \{(X_1, \dots, X_n); |\bar{X} - \mu_0| > l_\alpha\}$.
On calcule $\bar{X} = 11.96$, donc $|\bar{X} - \mu_0| = |11.96 - 12| = 0.04$.
Donc $|\bar{X} - \mu_0| \leq l_\alpha$, donc $(X_1, \dots, X_n) \notin W$, donc H_0 n'est pas rejetée.
Le bureau d'étude n'a pas besoin de prendre d'actions correctives.

(b) Discuter les implications d'une augmentation du seuil de signification. Quelle erreur peut augmenter si le seuil de signification est modifié ?

Solution.

Si on augmente α, u_α diminue, et donc l_α aussi. Il y aura alors plus de chances de rejeter H_0.
C'est l'erreur de première espèce qui va augmenter (c'est-à-dire rejeter à tort H_0).

(c) Quelle est la probabilité de conclure que le procédé opère correctement alors qu'en réalité il est centré à 11.90 ? Comment appelle-t-on ce risque ?

Solution.

On cherche $1-\beta(\mu_1)$ avec $\mu_1 = 11.90$, c'est-à-dire la probabilité de décider que $\mu = \mu_0$ (accepter H_0) alors qu'en réalité $\mu = \mu_1$.

$$\begin{aligned}
1-\beta(\mu_1) &= 1 - P((X_1,\ldots,X_n) \in W; \mu_1) \\
&= P((X_1,\ldots,X_n) \notin W; \mu_1) \\
&= P(|\bar{X} - \mu_0| \leq l_\alpha; \mu_1) \\
&= P(\mu_0 - l_\alpha \leq \bar{X} \leq \mu_0 + l_\alpha; \mu_1) \\
&= P(\frac{\mu_0 - \mu_1 - l_\alpha}{\sigma}\sqrt{n} \leq \sqrt{n}\,\frac{\bar{X}-\mu_1}{\sigma} \leq \frac{\mu_0 - \mu_1 + l_\alpha}{\sigma}\sqrt{n}; \mu_1) \\
&= P(t_1 \leq Z \leq t_2; \mu_1)
\end{aligned}$$

en posant

$$\begin{aligned}
Z &= \sqrt{n}\,\frac{\bar{X}-\mu_1}{\sigma} \\
t_1 &= \frac{\mu_0 - \mu_1 - l_\alpha}{\sigma}\sqrt{n} \\
t_2 &= \frac{\mu_0 - \mu_1 + l_\alpha}{\sigma}\sqrt{n}
\end{aligned}$$

On a $n = 30$ et les X_i sont supposés indépendants, donc avec le TCL, Z suit approximativement $\mathcal{N}(0,1)$. Alors

$$1-\beta(\mu_1) = \Phi(t_2) - \Phi(t_1)$$

A.N. :

$$\begin{aligned}
t_1 &= \frac{12 - 11.90 - 0.09873}{0.21}\sqrt{30} = 0.03312 \\
t_2 &= \frac{12 - 11.90 + 0.09873}{0.21}\sqrt{30} = 5.183
\end{aligned}$$

donc

$$\begin{aligned}
\Phi(t_1) &= \Phi(0.03312) = 0.5132 \\
\Phi(t_2) &= \Phi(5.183) \simeq 1
\end{aligned}$$

donc

$$\begin{aligned}
1-\beta(\mu_1) &= 1 - 0.5132 \\
&\boxed{1-\beta(\mu_1) = 0.4868}
\end{aligned}$$

La probabilité de conclure que le procédé opère correctement alors qu'en réalité il est centré à 11.90 est 48.68% ! Ce risque est appelé risque de deuxième espèce.

(d) Donner une estimation par intervalle de confiance au niveau 98% de la moyenne théorique μ.

Solution.

On pose $\alpha\prime = 0.02$.
Soit ε tel que

$$\begin{aligned} 1 - \alpha\prime &= P(|\bar{X} - \mu| \le \varepsilon) \\ &= P(\left|\sqrt{n}\,\frac{\bar{X} - \mu}{\sigma}\right| \le \frac{\varepsilon}{\sigma}\sqrt{n}) \\ &= P(|Y| \le \frac{\varepsilon}{\sigma}\sqrt{n}) \end{aligned}$$

donc

$$\begin{aligned} 1 - \frac{\alpha\prime}{2} &= P(Y \le \frac{\varepsilon}{\sigma}\sqrt{n}) \\ &= \Phi(\frac{\varepsilon}{\sigma}\sqrt{n}) \end{aligned}$$

À partir de la table de $\mathcal{N}(0,1)$, on trouve $u_{\alpha\prime} = 2.326$ correspondant à $1 - \frac{\alpha\prime}{2} = 0.99$.
Donc $\frac{\varepsilon}{\sigma}\sqrt{n} = u_{\alpha\prime}$ et $\varepsilon = \frac{\sigma}{\sqrt{n}} u_{\alpha\prime}$.
A.N. : $\varepsilon = \frac{0.21}{\sqrt{30}} \times 2.326 = 0.0892$
Un intervalle de confiance au niveau 98% de la moyenne théorique μ est donc

$$\boxed{[\bar{x} - \varepsilon, \bar{x} + \varepsilon] = [11.811, 11.989]}$$