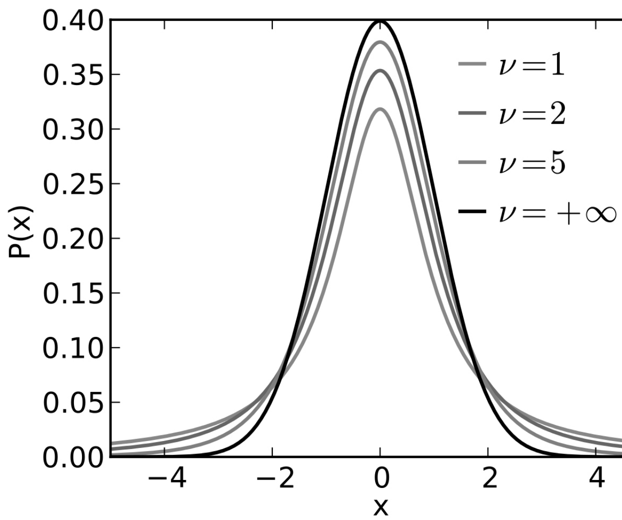
M3.5.1 Statistiques

Corrigé Examen 2016-2017



v0.2

24 mars 2017

Exercice 1. Un analyse financier étudie les comptes de 200 clients ayant souscrit un emprunt. À partir d'un échantillon de 20 comptes, il a établi les données réduites suivantes (exprimées en dirhams)

$$\bar{x} = 1514.69 \text{ et } \sum_{i=1}^{20}(x_i - \bar{x})^2 = 3904811.97$$

(a) En supposant que les données sont gaussiennes, donner une estimation ponctuelle de m, solde moyen d'un compte et de son écart-type σ.

(b) Donner des intervalles de confiance de 95% de m et σ.

Exercice 2. On désire estimer le nombre N d'individus d'une espèce animale vivant sur une île. Pour cela, on capture 800 individus ; ces individus sont marqués, puis relâchés. Ensuite, on re-capture ultérieurement 1000 animaux parmi lesquels on dénombre 250 animaux marqués. En déduire un intervalle de confiance à 95% de N.

Exercice 3. L'enquête annuelle d'un cabinet de conseil sur la qualité des automobiles a révélé que le nombre moyen de défauts par voiture neuve est de 1.07. Supposons qu'un échantillon de 36 voitures neuves d'un fabricant particulier fournisse les données suivantes, concernant le nombre de défauts par voiture

0	1	1	2	1	0	2	3	2	1	0	4	3	1	0	1	2	0
0	2	0	0	2	3	0	2	0	2	0	3	1	0	2	2	0	1

(a) Utiliser ces données pour déterminer le nombre moyen de défauts par voiture dans l'échantillon.

(b) Quel est l'écart-type de l'échantillon ?

(c) Construire un intervalle de confiance à 95% pour estimer le nombre moyen de défauts par voiture pour la population des voitures construites par ce fabricant.

(d) Après avoir pris connaissance de l'estimation par intervalle de confiance effectuée à la question (c), un statisticien suggère de tester un plus grand nombre de voitures neuves avant de conclure quant à la qualité de ces voitures, en la comparant à la moyenne des défauts par voiture au niveau de l'industrie (1.07) établie par le bureau de conseil. Êtes-vous d'accord ? Pourquoi ?

Exercice 1. Un analyse financier étudie les comptes de 200 clients ayant souscrit un emprunt. À partir d'un échantillon de 20 comptes, il a établi les données réduites suivantes (exprimées en dirhams)

$$\bar{x} = 1514.69 \text{ et } \sum_{i=1}^{20}(x_i - \bar{x})^2 = 3904811.97$$

(a) En supposant que les données sont gaussiennes, donner une estimation ponctuelle de m, solde moyen d'un compte et de son écart-type σ.

Solution.

Un estimateur de m est

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

donc une estimation ponctuelle de m est

$$\boxed{\bar{x} = 1514.69}$$

Un estimateur de σ est S avec

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

donc une estimation ponctuelle de σ est $s = \sqrt{s^2}$ avec

$$s^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = \frac{1}{20} \times 3904811.97 = 195240.60$$

$$\boxed{s = 441.86}$$

(b) Donner des intervalles de confiance de 95% de m et σ.

Solution.

Estimation par intervalle de confiance de m :
On cherche ε tel que

$$1-\alpha = P(|\bar{X}-m| \leq \varepsilon)$$

avec $\alpha = 5\%$.
Soit

$$T = \sqrt{n-1}\,\frac{\bar{X}-m}{S}$$

Les X_i sont supposées indépendantes et identiquement distribuées selon la loi normale $\mathcal{N}(m,\sigma)$. Alors avec le théorème de Fisher, T suit la loi de Student à $n-1$ degrés de liberté : $T \hookrightarrow \mathrm{St}(n-1)$.
On a

$$\begin{aligned}1-\alpha &= P(|\bar{X}-m| \leq \varepsilon)\\ &= P(\left|\sqrt{n-1}\,\frac{\bar{X}-m}{S}\right| \leq \sqrt{n-1}\,\frac{\varepsilon}{S})\\ &= P(|T| \leq \sqrt{n-1}\,\frac{\varepsilon}{S})\end{aligned}$$

À partir de la table de la loi de Student, on trouve $t_{n-1,\alpha} = 2.0930$ correspondant à $n-1 = 19$ et $1-\alpha = 0.95$.
On a $t_{n-1,\alpha} = \sqrt{n-1}\,\frac{\varepsilon}{s}$, donc $\varepsilon = \frac{s}{\sqrt{n-1}} t_{n-1,\alpha}$.
A.N. :

$$\varepsilon = \frac{441.86}{\sqrt{19}} \times 2.0930 = 212.17$$

donc un intervalle de confiance à 95% pour m est

$$\boxed{[\bar{x}-\varepsilon, \bar{x}+\varepsilon] = [1302.52, 1726.86]}$$

Estimation par intervalle de confiance de σ :
On a avec le théorème de Fisher, $Z = n\frac{S^2}{\sigma^2}$ suit la loi de khi-deux à $n-1$ degrés de liberté : $Z \hookrightarrow \chi^2_{n-1}$.
On a $\forall\, 0 < a < b$, d'une part

$$P(a \leq Z \leq b) = F_{\chi^2_{n-1}}(b) - F_{\chi^2_{n-1}}(a)$$

avec $F_{\chi^2_{n-1}}$ la fonction de répartition de la loi de khi-deux à $n-1$ degrés de liberté.
Et d'autre part,

$$\begin{aligned}P(a \leq Z \leq b) &= P(a \leq n\frac{S^2}{\sigma^2} \leq b)\\ &= P(\frac{n}{b}S^2 \leq \sigma^2 \leq \frac{n}{a}S^2)\\ &= P(\sqrt{\frac{n}{b}}S \leq \sigma \leq \sqrt{\frac{n}{a}}S)\end{aligned}$$

On équilibre les risques en choisissant a et b tels que

$$F_{\chi^2_{n-1}}(b) = 1 - \frac{\alpha}{2} \text{ et } F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$$

pour avoir

$$P(\sqrt{\frac{n}{b}}S \leq \sigma \leq \sqrt{\frac{n}{a}}S) = (1 - \frac{\alpha}{2}) - \frac{\alpha}{2} = 1 - \alpha$$

donc $\left[\sqrt{\frac{n}{b}}s, \sqrt{\frac{n}{a}}s\right]$ est un intervalle de confiance de σ au niveau de confiance de 95%. On calcule a et b à partir de la table de la loi de khi-deux avec $n - 1 = 19$. On trouve

$$b = 32.85 \text{ correspondant à } 1 - \frac{\alpha}{2} = 0.975$$
$$\text{et } a = 8.91 \text{ correspondant à } \frac{\alpha}{2} = 0.025$$

A.N. :

$$\sqrt{\frac{n}{b}}s = \sqrt{\frac{20}{32.85}} \times 441.86 = 344.77$$

$$\sqrt{\frac{n}{a}}s = \sqrt{\frac{20}{8.91}} \times 441.86 = 662.00$$

donc un intervalle de confiance à 95% pour σ est

$$\boxed{[344.77, 662.00]}$$

Exercice 2. On désire estimer le nombre N d'individus d'une espèce animale vivant sur une île. Pour cela, on capture 800 individus ; ces individus sont marqués, puis relâchés. Ensuite, on re-capture ultérieurement 1000 animaux parmi lesquels on dénombre 250 animaux marqués. En déduire un intervalle de confiance à 95% de N.

Solution.

Soit

$$X_i = \begin{cases} 1 & \text{si le } i^e \text{ animal de l'échantillon est marqué} \\ 0 & \text{sinon} \end{cases}$$

la variable aléatoire qui suit la loi de Bernoulli de paramètre p, avec $1 \le i \le n = 1000$.
On a $p = \frac{800}{N}$, et on cherche à estimer $N = \frac{800}{p}$.
L'ESBVM de p est

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

donc l'estimation ponctuelle de p dans l'échantillon est $\hat{p} = \frac{250}{1000} = 0.25$.
Soit

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

Les X_i sont supposés indépendants et $n \ge 30$, 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

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

avec $\alpha = 5\%$.
Pour trouver l'intervalle de confiance recherché, il suffit d'écrire $|Z| \le u_\alpha$ sous la forme $r_1 \le p \le r_2$.

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

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

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

donc

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

En négligeant u_α devant n, on a

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

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. :

$$\begin{aligned} r_1 &= 0.25 - 1.96 \sqrt{\frac{0.25(1-0.25)}{1000}} = 0.223 \\ r_2 &= 0.25 + 1.96 \sqrt{\frac{0.25(1-0.25)}{1000}} = 0.277 \end{aligned}$$

On a donc

$$\begin{aligned} 1 - \alpha &= P(r_1 \leq p \leq r_2) \\ &= P(r_1 \leq \frac{800}{N} \leq r_2) \\ &= P(\frac{800}{r_2} \leq N \leq \frac{800}{r_1}) \end{aligned}$$

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

$$\boxed{\left[\frac{800}{r_2}, \frac{800}{r_1}\right] = [2890, 3585]}$$

Exercice 3. L'enquête annuelle d'un cabinet de conseil sur la qualité des automobiles a révélé que le nombre moyen de défauts par voiture neuve est de 1.07. Supposons qu'un échantillon de 36 voitures neuves d'un fabricant particulier fournisse les données suivantes, concernant le nombre de défauts par voiture

0	1	1	2	1	0	2	3	2	1	0	4	3	1	0	1	2	0
0	2	0	0	2	3	0	2	0	2	0	3	1	0	2	2	0	1

(a) Utiliser ces données pour déterminer le nombre moyen de défauts par voiture dans l'échantillon.

Solution.

Le nombre moyen de défauts par voiture dans l'échantillon est

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{44}{36} = 1.22$$

(b) Quel est l'écart-type de l'échantillon ?

Solution.

La variance de l'échantillon est

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{100}{44} - 1.49 = 0.779$$

donc l'écart-type de l'échantillon est

$$s = \sqrt{s^2} = 0.883$$

(c) Construire un intervalle de confiance à 95% pour estimer le nombre moyen de défauts par voiture pour la population des voitures construites par ce fabricant.

Solution.

Soit m le nombre moyen réel de défauts par voiture pour la population des voitures construites par le fabricant. On cherche ε tel que

$$1 - \alpha = P(|\bar{X} - m| \leq \varepsilon)$$

avec $\alpha = 0.05$.
Soit

$$T = \sqrt{n-1}\,\frac{\bar{X} - m}{S}$$

Les X_i sont supposées indépendantes et identiquement distribuées, et $n = 36 \geq 30$. Alors avec le TCL et le théorème de Fisher, T suit approximativement la loi de Student à $n-1$ degrés de liberté : $T \hookrightarrow \mathrm{St}(n-1)$.

On a

$$\begin{aligned} 1-\alpha &= P(|\bar{X}-m| \leq \varepsilon) \\ &= P\left(\left|\sqrt{n-1}\,\frac{\bar{X}-m}{S}\right| \leq \sqrt{n-1}\,\frac{\varepsilon}{S}\right) \\ &= P(|T| \leq \sqrt{n-1}\,\frac{\varepsilon}{S}) \end{aligned}$$

À partir de la table de la loi de Student, pour trouver $t_{n-1,\alpha}$ correspondant à $n-1=35$ et $1-\alpha=0.95$, on fait une interpolation linéaire : $\frac{t_{n-1,\alpha}-2.0423}{35-30} = \frac{2.0211-2.0423}{40-30}$ qui donne $t_{n-1,\alpha} = 2.0317$.

On a $t_{n-1,\alpha} = \sqrt{n-1}\,\dfrac{\varepsilon}{s}$, donc $\varepsilon = \dfrac{s}{\sqrt{n-1}} t_{n-1,\alpha}$.

A.N. :

$$\varepsilon = \frac{0.883}{\sqrt{35}} \times 2.0317 = 0.303$$

donc un intervalle de confiance à 95% pour m est

$$\boxed{[\bar{x}-\varepsilon, \bar{x}+\varepsilon] = [0.917, 1.523]}$$

(d) Après avoir pris connaissance de l'estimation par intervalle de confiance effectuée à la question (c), un statisticien suggère de tester un plus grand nombre de voitures neuves avant de conclure quant à la qualité de ces voitures, en la comparant à la moyenne des défauts par voiture au niveau de l'industrie (1.07) établie par le bureau de conseil. Êtes-vous d'accord ? Pourquoi ?

Solution.

Un échantillon plus grand (augmentation de n) donnera un intervalle de confiance plus précis (ε plus petit).

La moyenne au niveau de l'industrie est 1.07. Cette valeur est incluse dans l'intervalle de confiance, mais elle est beaucoup plus proche de la borne inférieure. Il se pourrait bien qu'un test sur un plus grand nombre de voitures donne un intervalle de confiance qui exclue cette moyenne.

Le fabricant devrait donc faire un test plus grand.