Correction du TD 5 : Estimation par Intervalle de Confiance

Exercice 1 : Intervalle de confiance pour une proportion

Deux candidats, Ségolène et Nicolas, sont en présence lors du deuxième tour d'une élection présidentielle au cours de laquelle 40 millions électeurs sont amenés à s'exprimer.

n personnes sont tirées au hasard parmi ces électeurs et interrogées sur leurs intentions de vote (on suppose qu'à ce moment tous les électeurs ont fixé leur choix et n'en changeront pas au moment du vote). 52% des électeurs interrogés annoncent qu'ils sont partisans de Ségolène.

1. **Calculer, pour n = 100, une borne inférieure de confiance 95% pour le pourcentage d'électeurs favorables à Ségolène dans la population totale.**

   Modèle : Soit X indicatrice de "favorable à Ségolène", et p la proportion d'électeurs favorables à Ségolène parmi les 40.000 électeurs. Le nombre n d'électeurs interrogés étant très petit devant la taille de la population, on peut considérer que les n variables tirées sont indépendantes et identiquement distribuées, comme si le tirage avait été fait "avec remise". Le modèle que nous utilisons est donc : $X_1, ..., X_n \sim i.i.d.B(1,p)$ (n-échantillon d'une loi de Bernoulli de paramètre p).

   Pour construire un intervalle de confiance, nous devons utiliser une "fonction pivotale", c'est-à-dire une variable fonction uniquement des observations et du paramètre étudié, dont la loi soit entièrement connue. La fonction pivotale utilisée ici est construite à partir de la proportion observée d'électeurs favorables à Ségolène parmi les n interrogés.

$$\sum_{i=1}^{n} X_i \sim B(n,p)$$

   Nous considérons que n est assez grand pour pouvoir utiliser l'approximation normale de la loi de cette proportion, notée F_n :

$$B(n,p) \approx N(np, np(1-p))$$

$$F_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(p, \frac{p(1-p)}{n}\right) \Leftrightarrow U_n = \frac{F_n - p}{\sqrt{\frac{F_n(1-F_n)}{n}}} \approx N(0;1)$$

$$\text{Lecture dans la loi } N[0;1] : P\{U \leq 1,645\} = 0,95$$

   On en déduit pour F_n que $P\left\{\frac{F_n - p}{\sqrt{\frac{F_n(1-F_n)}{n}}} \leq 1,645\right\} \simeq 0,95$

$$P\left\{F_n - p \leq 1,645\sqrt{\frac{F_n(1-F_n)}{n}}\right\} \simeq 0,95$$

$$P\left\{p \geq F_n - 1,645\sqrt{\frac{F_n(1-F_n)}{n}}\right\} \simeq 0,95$$

La borne inférieure de confiance 95% est donc égale à $F_n - 1,645\sqrt{\frac{F_n(1-F_n)}{n}}$.

La proportion observée de $f_n = 0,52$ conduit à $p \geq 0,4378$.

Rappel : il n'est pas question d'écrire la probabilité que p soit supérieur à 0,4378, car il n'y a plus rien d'aléatoire dans l'inégalité; une fois l'observation faite, la proposition est soit vraie (proba 1) soit fausse (proba 0).

Remarque : nous cherchions un intervalle unilatéral de la forme $p \geq A(F_n)$. Nous avons donc calculé un intervalle de probabilité unilatéral également.

L'approximation de la loi de U_n par une loi normale est valable pour tout p tel que $np(1-p) \geq 15$, c'est-à-dire, pour $n = 100$: $(p-0.5)^2 - 0,25 + 0,015 \leq 0$, ou encore $0,184 \leq p \leq 0,816$. Nous pouvons considérer que p satisfait ces conditions. Nous vérifions que la borne inférieure trouvée est bien supérieure à 0,184.

2. **Que devient cette borne inférieure de confiance 95% pour les valeurs suivantes de n :**
   (a) **n = 1000 ?**
   (b) **n = 2000 ?**

   Tableau des résultats numériques pour les différentes valeurs de n, avec l'observation de $f_n = 0,52$.

	$p \geq$
n=1000	49,40%
n=2000	50,16%

   Nous vérifions que lorsque n augmente, la borne inférieure de p se rapproche de F_n.

3. **A partir de quelle taille n du sondage effectué, le pourcentage observé de 52% d'électeurs favorables à Ségolène conduirait-il celui-ci à accorder une confiance de 95% au fait d'être élu (c'est-à-dire que la borne inférieure de confiance 95% serait supérieure ou égale à 0.50) ?**

   Nous devons résoudre l'inégalité $0,52 - 1,645\sqrt{\frac{0,52*0,48}{n}} \geq 0,5$. Cela conduit à $n \geq 0,52*0,48\left(\frac{1,645}{0.02}\right)^2 = 1688,6$

   La proportion observée de 52% électeurs favorables pourrait donc conduire Albert à accorder une confiance de 95% au fait d'être élu si cette proportion était observée sur au moins 1689 électeurs.

Exercice 2 : Intervalle de confiance pour l'espérance d'une variable Normale

Le coût d'un certain type de sinistre peut être considéré comme une variable aléatoire X suivant une loi Normale $\mathcal{N}(m, \sigma^2)$. On observe, dans une compagnie d'assurance, n dossiers de sinistres indépendants.

1. **On suppose que l'écart-type σ est connu, égal à 15 Euros.**
   Modèle : $X_1, ..., X_N \approx i.i.d.\ \mathcal{N}(m, \sigma^2)$ avec $\sigma = 15$.

   (a) **Calculer l'intervalle bilatéral de confiance 98% pour m.**
   Application numérique : pour 20 dossiers, la moyenne des coûts observée est 120 Euros : dans quelle fourchette placez-vous m ?
   Si $X_1, ..., X_N \approx i.i.d.\ \mathcal{N}(m, \sigma^2)$, alors $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \approx \mathcal{N}\left(m; \frac{\sigma^2}{N}\right)$ et donc

$$U_N = \frac{\overline{X}_N - m}{15/\sqrt{N}} \approx N(0;1) \quad \text{(même pour les petites valeurs de } N\text{)}$$

Un intervalle bilatéral de confiance 98% correspond à une probabilité 1% de chaque côté.
Lecture dans la table de la loi $N(0;1)$: $P\{U \leq 2,326\} = 0,99$
On en déduit que : $P\{-2,326 \leq U \leq 2,326\} = 0,98$, et en résolvant les inégalités en m , on obtient l'intervalle bilatéral de confiance 98% pour m :

$$P\left\{\overline{X}_N - 2,326\frac{15}{\sqrt{N}} \leq m \leq \overline{X}_N + 2,326\frac{15}{\sqrt{N}}\right\} = 0,98$$

Application numérique : pour $N = 20$, nous obtenons l'intervalle bilatéral de confiance 0,98% suivant :

$$P\left\{\overline{X}_{20} - 7,80 \leq m \leq \overline{X}_{20} + 7,80\right\} = 0,98$$

L'observation de $\overline{x} = 120$ Euros conduit à la fourchette : $112,20 \leq m \leq 127,80$ Euros.

(b) **Combien de dossiers doit-on examiner pour que la longueur de l'intervalle de confiance 98% soit inférieure ou égale à 10 Euros ?**
La longueur de l'intervalle de confiance 98% est égale à $2 \times 2,326\frac{15}{\sqrt{N}}$, soit 15,60 Euros pour 20 dossiers examinés. Pour que cette longueur soit inférieure ou égale à 10 Euros, il faut que $2 \times 2,326\frac{15}{\sqrt{N}} \leq 10$, c'est-à-dire $N \geq \left(15 \times \frac{2\times 2,326}{10}\right)^2 = 48,7$.
Il faudrait donc examiner *au moins 49 dossiers* pour que la longueur de l'intervalle de confiance 98% soit inférieure à 10 Euros.

2. **L'écart-type σ n'est en fait pas connu.**
Si l'écart-type n'est pas connu, on ne peut utiliser la variable normale $\frac{\overline{X}_N - m}{\sigma/\sqrt{N}}$ comme fonction pivotale, car elle contient l'inconnue σ.

(a) **Comment l'intervalle est-il modifié ?**
Application numérique : pour 20 dossiers, la moyenne des coûts observés est 120 Euros, et l'estimation sans biais de σ^2 est égale à $(15)^2$.
Propriété utilisée : soient $X_1, ..., X_N \approx i.i.d.N\left(m, \sigma^2\right)$: nous notons $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ et $S^2 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1}$.
Alors : $Z_N = \frac{\overline{X} - m}{S/\sqrt{N}} \approx T_{(N-1)}$, loi indépendante de m et de σ^2.
C'est cette variable Z_N que nous devons utiliser pour construire l'intervalle de confiance pour m (la variable S^2 est l'estimation sans biais de σ^2).
Lecture dans la table de la loi $T_{(20-1)}$: $P\{|Z| \leq 2,540\} = 0,98$
On en déduit : $P\left\{-2,540 \leq \frac{\overline{X} - m}{S/\sqrt{N}} \leq 2,540\right\} = 0,98$
et donc : $P\left\{\ \overline{X} - 2,540\frac{S}{\sqrt{N}} \leq m \leq\ \overline{X} + 2,540\frac{S}{\sqrt{N}}\right\} = 0,98$
Pour $N = 20$, $\frac{2,540}{\sqrt{20}} = 0,5680$:

$$P\left\{\ \overline{X} - 0,5680.S \leq m \leq\ \overline{X} + 0,5680.S\right\} = 0,98$$

avec $\overline{x} = 120$ et $S = 15$, on obtient : $111,48 \leq m \leq 128,52$ Euros.
Rappel : ne pas écrire la probabilité que m soit compris entre 111,48 et 128,52, car il n'y a plus rien d'aléatoire dans l'inégalité une fois l'observation faite : l'événement correspondant est soit réalisé, soit non-réalisé.

(b) **Vérifier que la longueur de l'intervalle de confiance est aléatoire. Croyez-vous pouvoir calculer facilement sa valeur moyenne ?**
La longueur de l'intervalle est égale à $2 \times 0,5680 \times S = 1,136 \times S$, où S est aléatoire. Nous savons que $E\left(S^2\right) = \sigma^2$. Cela n'entraîne pas que $E\left(S\right)$ soit calculable facilement.
Remarque supplémentaire : En fait (inégalité de Jensen) : $\sqrt{u}$ est une fonction concave, cela entraîne que :

$$E\left(S\right) = E\left(\sqrt{S^2}\right) < \sqrt{E\left(S^2\right)} = \sigma$$

Exercice 3 :

Soient (Y_1, Y_2,...Y_{25}) des variables aléatoires indépendantes et identiquement distribuées selon une loi Normale d'espérance m et de variance σ^2 :

$$Y_1, Y_2, ..., Y_{25} \approx \mathcal{N}(m, \sigma^2)$$

Soient (Y_1, Y_2,...Y_{25}) des variables aléatoires indépendantes et identiquement distribuées selon une loi Normale d'espérance m et de variance σ^2 :

$$Y_1, Y_2, ..., Y_{25} \approx i.i.d.\ \mathcal{N}(m, \sigma^2)$$

Rappels : Si X et Y sont des variables aléatoires et a une constante, alors on a
- *E(X + a)=E(X) + a*
- *E(X + Y)=E(X)+E(Y)*
- *$V(aX) = a^2V(X)$*
- *V(X+Y)=V(X)+V(Y) si et seulement si X et Y sont indépendants.*

1. **Quel estimateur proposez vous pour l'espérance m ?**
Le meilleur estimateur de m est la moyenne empirique $\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i$.

2. **Quelle est la loi de cet estimateur ?**
Cet estimateur suit une loi normale de moyenne l'espérance de $\bar{y}$, $E(\bar{y})$ et de variance $V\left(\bar{y}\right)$.

$$E(\bar{y}) = \frac{1}{25}\sum_{i=1}^{25} E(y_i) = \frac{1}{25}25 * m = m$$

$$V\left(\bar{y}\right) = \left(\frac{1}{25}\right)^2 \sum_{i=1}^{25} V(y_i) = \left(\frac{1}{25}\right)^2 25\sigma^2 = \frac{\sigma^2}{25}$$

D'où $\bar{y} \approx \mathcal{N}\left(m, \frac{\sigma^2}{25}\right)$

3. **Si nous pouvons assurer que la variance est connue, $\sigma^2 = 2,25$, calculez un intervalle de confiance à 95 % pour *m*.**
Rappel : Un intervalle de confiance pour le paramètre θ de niveau de confiance $1 - \alpha$ est un intervalle qui a la probabilité $1 - \alpha$ de contenir la vraie valeur de θ.
En observant la table de la loi normale centrée réduite, $U \approx N(0, 1)$ on peut en déduire un intervalle de confiance à 95 % :

$$p\big(-1,96 < U < 1,96\big) = 0,95$$

Il faut donc transformer notre estimateur $\bar{y}$ en un estimateur centré réduit. Il est facile de vérifier que $\frac{\bar{y}-m}{\sigma/5} \approx N(0,1)$ et dès lors :

$$p\Big(-1,96 < \frac{\bar{y}-m}{\sigma/5} < 1,96\Big) = 0,95$$

$$p\Big(\bar{y} - \frac{1,96}{5}\sigma < m < \bar{y} + \frac{1,96}{5}\sigma\Big) = 0,95$$

$$p\Big(\bar{y} - 0,588 < m < \bar{y} + 0,588\Big) = 0,95$$

4. **La variance est en fait inconnue. Comment est modifié l'intervalle de confiance à 95 % précédent ?**
Lorsque la variance est inconnue, il faut utiliser un estimateur de celle-ci. Le meilleur estimateur est la variance empirique, soit $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ dans le cas général et dans notre cas particulier $S^2 = \frac{1}{24}\sum_{i=1}^{24}(y_i - \bar{y})^2$.

La loi du Chi-deux *Si X_1, X_2... X_n des variables aléatoires indépendantes qui suivent une loi normale centrée réduite, alors*

$$\sum (X_i)^2 \approx \chi^2$$

Application aux échantillons normaux *Si X_1, X_2... X_n des variables aléatoires indépendantes qui suivent une loi normale $N(m,\sigma^2)$, alors*

$$\frac{X_i - m}{\sigma} \approx N(0,1) \quad \text{et} \quad \sum \Big(\frac{X_i - m}{\sigma}\Big)^2 \approx \chi^2_n$$

Si on note en outre $\bar{X}$ la moyenne empirique de m, alors

$$\sum \Big(\frac{X_i - \bar{X}}{\sigma}\Big)^2 \approx \chi^2_{n-1}$$

En effet, ces variables aléatoires suivent bien une loi normale centrée réduite mais vérifient aussi une relation (leur somme est égale à 0) ce qui induit un degré de liberté de moins.

La loi de Student *Si $U \approx N(0,1)$ et $Y \approx \chi^2_n$, alors*

$$\frac{U}{\sqrt{\frac{Y}{n}}} \approx T_n$$

En appliquant les résultats de cours ci-dessus, il est facile de montrer que l'estimateur $\frac{\bar{y}-m}{S/5} \approx T_{24}$. On lit alors dans la table de Student :

$$p\big(-2,064 < \frac{\bar{y}-m}{S/5} < 2,064\big) = 0,95$$

$$p\big(\bar{y} - \frac{2,064}{5}S < m < \bar{y} + \frac{2,064}{5}S\big) = 0,95$$

$$p\big(\bar{y} - 0,4128S < m < \bar{y} + 0,4128S\big) = 0,95$$

5. **Application numérique : donner l'estimation de m et les deux intervalles de confiance correspondant aux observations faites, qui ont pour moyenne et variance empirique :**

$$\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i = 13,5 \text{ et } s^2 = \frac{1}{24}\sum_{i=1}^{25}(y_i - \bar{y})^2 = 1,69$$

Variance connue : $12,912 \leq m \leq 14,088$
Variance inconnue : $12,963 \leq m \leq 14,037$

Exercice 4 :

On observe sur un échantillon E_n de taille $n = 400$, deux variables aléatoires X et Y indépendantes telles que : E(X)=2, V(X)=4, E(Y)=8, V(Y)=10.

1. **Déterminez la valeur de λ pour que $Z = 5X + \lambda Y$ ait une espérance mathématique nulle.**
$E(Z) = 0 \Rightarrow 5E(X) + \lambda E(Y) = 0$

$$\lambda = -\frac{5E(X)}{E(Y)} = -\frac{10}{8} = -1,25$$

2. **Calculez V(Z).**
$V(Z) = 25V(X) + \lambda^2 V(Y) = 100 + \frac{100}{64}10 = 115,625$

3. **En déduire un intervalle I de la forme [-a,a] tel que $P(Z \in I) \geq 0,9$.**
D'après l'inégalité de Bienaymé-Tchebychev (également appelée de Chebychev) :

$$P(|Z - E(Z)| < \varepsilon) = 1 - \frac{var(Z)}{\varepsilon^2}$$

Donc $P(-a < Z < a) = 1 - \frac{115,625}{a^2} \geq 0,9 \Rightarrow a = 34,0036$ On a donc I=(-34,0036 ; 34,0036).