Université Lyon 2
Licence 3 MIASHS

Statistique Inférentielle

TD1 : Estimation

Exercice 1 En 2011 en France, la durée moyenne des périodes de chômage était de 14 mois. En supposant que la durée d'une période de chômage peut-être modélisée par une loi normale, de moyenne 14 et de variance 36, répondez aux questions suivantes :

(i) quelles sont les limites de cette modélisation ?
Réponse : Une loi normale $\mathcal{N}(\mu, \sigma^2)$ modélise la distribution d'une variable aléatoire X qui peut *a priori* prendre des valeurs négatives, ce qui n'est pas le cas ici.

(ii) quelle est la probabilité qu'une période de chômage dure plus de 2 ans ?
Réponse : On introduit la variable centrée-réduite associée

$$Z = \frac{X - \mu}{\sqrt{\sigma^2}} = \frac{X - 14}{\sqrt{36}}.$$

Cette variable aléatoire suit une loi normale $\mathcal{N}(0, 1)$. On cherche[1] $\mathbb{P}(X > 24)$. On peut

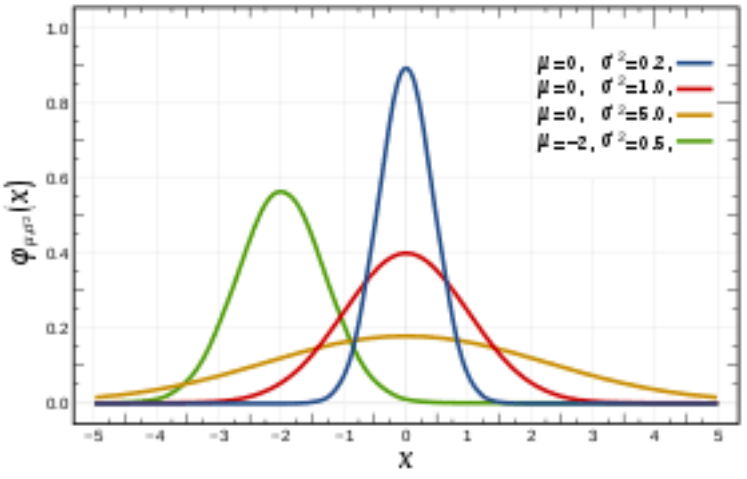

Figure 1 – Loi normale

réexprimer cette quantité en fonction de Z, dont on connaît la loi, de la manière suivante :

$$\begin{aligned}
\mathbb{P}(X > 24) &= \mathbb{P}\left(\frac{(X - \mu)}{\sqrt{\sigma^2}} > \frac{(24 - \mu)}{\sqrt{\sigma^2}}\right), \\
&= \mathbb{P}\left(Z > \frac{24 - 14}{6}\right), \\
&= \mathbb{P}\left(Z > 1.66666\right).
\end{aligned}$$

1. sachant que 2 ans font 24 mois

La table de la loi normale centrée réduite consigne les valeurs de la fonction de répartition de Z, c'est à dire les valeurs de $\mathbb{P}(Z \geq x)$ pour certaines valeur de x. Comme

$$\mathbb{P}\left(Z > 1.66666\right) = 1 - \mathbb{P}\left(Z \leq 1.66666\right)$$

on peut trouver dans la Table 1.3 du polycopié que $\mathbb{P}\left(Z \leq 1.66666\right) = 0.9515$ et donc $1 - \mathbb{P}\left(Z \leq 1.66666\right) = 0.0485$, ce qui donne la réponse à la question $\mathbb{P}(X > 24) = 0.0485$.

(iii) quelle est la probabilité qu'une période de chômage dure moins de 6 mois ?

Réponse : la méthode est la même. On cherche $\mathbb{P}(X \leq 6)$. On peut réexprimer cette quantité en fonction de Z, dont on connaît la loi, de la manière suivante [2] :

$$\begin{aligned}\mathbb{P}(X < 6) &= \mathbb{P}\left(\frac{(X-\mu)}{\sqrt{\sigma^2}} < \frac{(6-\mu)}{\sqrt{\sigma^2}}\right), \\ &= \mathbb{P}\left(Z < \frac{6-14}{6}\right), \\ &= \mathbb{P}\left(Z < -1.33333\right).\end{aligned}$$

Comme les valeurs négatives ne se trouvent pas dans la table, il faut faire une transformation simple, utilisant la symétrie de la Gaussienne centrée-réduite :

$$\begin{aligned}\mathbb{P}\left(Z < -1.33333\right) &= \mathbb{P}\left(Z > 1.33333\right) \\ &= 1 - \mathbb{P}\left(Z \leq 1.33333\right)\end{aligned}$$

on peut trouver dans la Table 1.3 du polycopié que $\mathbb{P}\left(Z \leq 1.33333\right) = 0.9082$ et donc $1 - \mathbb{P}\left(Z \leq 1.333333\right) = 0.0918$, ce qui donne la réponse à la question $\mathbb{P}(X > 6) = 0.0918$.

(iv) quelle est la probabilité qu'une période de chômage dure entre 6 mois et 1 an ?

Réponse : on cherche $\mathbb{P}(6 < X \leq 12)$. On peut réexprimer cette quantité en fonction de Z, dont on connaît la loi, de la manière suivante [3] :

$$\begin{aligned}\mathbb{P}(6 < X \leq 12) &= \mathbb{P}\left(\frac{(6-\mu)}{\sqrt{\sigma^2}} < \frac{(X-\mu)}{\sqrt{\sigma^2}} \leq \frac{(12-\mu)}{\sqrt{\sigma^2}}\right), \\ &= \mathbb{P}\left(\frac{6-14}{6} < Z \leq \frac{12-14}{6}\right), \\ &= \mathbb{P}\left(-1.33333 < Z \leq -.33333\right).\end{aligned}$$

Notons que l'événement $\{-1.33333 < Z \leq -.33333\}$ est égal à l'événement $\{Z \leq -.33333\}$ auquel on enlève l'événement $\{Z < -1.33333\}$. On obtient donc

$$\mathbb{P}(6 < X \leq 12) = \mathbb{P}\left(Z \leq -.33333\right) - \mathbb{P}\left(Z < -1.33333\right).$$

Comme les valeurs négatives ne se trouvent pas dans la table, il faut faire une transformation simple, utilisant la symétrie de la Gaussienne centrée-réduite :

$$\begin{aligned}\mathbb{P}\left(Z < -1.33333\right) &= \mathbb{P}\left(Z > 1.33333\right) \\ &= 1 - \mathbb{P}\left(Z \leq 1.33333\right)\end{aligned}$$

2. sachant que 1 an fait 12 mois
3. sachant que 2 ans font 24 mois

On a déjà calculé que $\mathbb{P}(Z < -1.33333) = 0.0918$. De la même manière, on calcule que $\mathbb{P}(Z < -0.33333) = 1 - \mathbb{P}(Z \leq 0.33333) = 1 - 0.6293 = 0.3707$. Ainsi,

$$\mathbb{P}(6 < X \leq 12) = 0.3707 - 0.0918 = 0.2789.$$

Exercice 2 Mon opérateur de téléphonie mobile m'assure que 95% des SMS que j'envoie seront transmis en moins d'une minute.

(i) Quelle est la probabilité qu'un SMS envoyé arrive en moins d'une minute ?
Réponse : On définit une variable aléatoire X qui modélise le temps mis par un SMS pour être transmis. Dans ce modèle, on peut interpréter l'information donnée comme

$$\mathbb{P}\left(X \leq 1\right) = .95.$$

(ii) J'envoie chaque jour 2 SMS. Quelle est la probabilité que le nombre de SMS arrivés en moins d'une minute soit : 0, 1, 2 ?
Réponse : pour chaque SMS envoyé, on définit la variable de Benoulli $\mathcal{B}(.95)$ Z_i, $i = 1, 2$ qui prend la valeur 0 si le SMS numéro i a été envoyé après 1 minute et la valeur 1 sinon. Le nombre de SMS envoyés en moins d'une minute est donc $Z = Z_1 + Z_2$. Il s'agit d'une variable Binomiale $\mathcal{B}(2, .95)$. Pour une variable binomiale générale $\mathcal{B}(n, p)$, Z, on a

$$\begin{aligned}
\mathbb{P}\left(Z = 0\right) &= C_n^0\ p^0(1-p)^{n-0} \\
\mathbb{P}\left(Z = 1\right) &= C_n^1\ p^1(1-p)^{n-1} \\
\vdots &\qquad\quad \vdots \\
\mathbb{P}\left(Z = n\right) &= C_n^n\ p^n(1-p)^{n-n}
\end{aligned}$$

Dans notre cas on a $n = 2$ et $p = .95$ et donc

$$\begin{aligned}
\mathbb{P}\left(Z = 0\right) &= (1 - .95)^2 = .0025 \\
\mathbb{P}\left(Z = 1\right) &= 2\ p(1 - .95) = .095 \\
\mathbb{P}\left(Z = 2\right) &= p^2 = .9025.
\end{aligned}$$

(iii) Le week-end, j'envoie cette fois 20 SMS par jour. Proposez une modélisation pour le nombre de SMS arrivés en moins d'une minute.
Réponse : même modélisation que pour l'exercice précédant, mais avec cette cette fois $n = 20$ et $Z = Z_1 + Z_2 + \ldots + Z_{20} \sim \mathcal{B}(20, .95)$.

(iv) Quelle est la probabilité pour que le dimanche, au moins la moitié de mes SMS arrive en moins d'une minute ?

Réponse : on veut calculer $\mathbb{P}(Z > 10)$.

Solution 1. $\mathbb{P}(Z \geq 10) = P(Z' \leq 9)$ où $Z' \sim \mathcal{B}(1, .05)$ est le nombre de SMS arrivés en plus d'une minute. D'après les tables statistiques disponibles à la fin du manuscrit, $P(Z' \leq 9) = 1$.

Solution 2. On peut faire une approximation Gaussienne en utilisant le fait que $\mathbb{E}[Z] = 20 * .95 = 19.1$ et $\mathrm{Var}(Z) = 20 \cdot .95 \cdot (1 - .95) = .95$. Ainsi,

$$\begin{aligned}\mathbb{P}(Z \geq 10) &= \mathbb{P}\left(\frac{Z - 19.1}{\sqrt{.95}} > \frac{10 - 19.1}{\sqrt{.95} = -9.34}\right) \\ &\approx \mathbb{P}(W > -9.34)\,.\end{aligned}$$

où W est une variable Gaussienne centrée réduite $\mathcal{N}(0,1)$. On procède alors comme dans l'exercice précédent. Dans la table, on trouve que $\mathbb{P}(W > -9.34) \approx 0$ et donc la probabilité d'envoyer plus de 10 messages, chacun partant en moins d'une minute est très proche de 1. On aurait pu aussi utiliser une approximation par la loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda = \mathbb{E}[Z] = 19.1$. Dans la Table 1.2, on trouve

$$\mathbb{P}(Z \leq 9) = 0.0089,$$

ce qui nous donne que

$$\begin{aligned}\mathbb{P}(Z \geq 9) &= 1 - \mathbb{P}(Z \leq 9) \\ &= .9911.\end{aligned}$$

Dans le cas présent, en supposant que la loi de Poisson est une approximation plus précise que celle par la loi normale, on prendra le résultat que nous venons de trouver.

Exercice 3 Sur mon ordinateur, sous le logiciel R, j'ai tapé la commande suivante :
$round(rnorm(12, 10, 1), 1)$
J'ai alors obtenu le résultat suivant : 10.7 11.7 11.6 9.1 11.1 10.7 10.7 11.1 9.5 10.6 12.0 11.1

(i) Estimer l'espérance et la variance de la population dont sont issues ces observations.
L'espérance est estimée par la moyenne empirique, 10.825, et la variance par la variance empirique débiaisée 0.709318.

(ii) Que fait ce code R ?
Ce code génère une suite de 12 tirages d'une variable aléatoire Gaussienne $(10, 1)$ dont les valeurs sont arrondies au dixième.

Exercice 4 Une société de vente à distance demande à l'un de ses ingénieurs marketing de modéliser le nombre d'appels téléphoniques par heure reçus sur le standard dédié aux commandes, dans le but d'optimiser la taille de celui-ci. Les nombres d'appels, relevés sur une période de 53 heures, ont été les suivants :

Nb d'appels x_i	0	1	2	3	4	5	6	7	8	9
Occurence N_i	1	4	7	11	10	9	5	3	2	1

(i) Estimer le mode, la moyenne et la variance du nombre d'appels.
Le mode est la valeur qui correspond au plus grand nombre d'appels. C'est donc 3. La moyenne est

$$\begin{aligned}\bar{x} &= \frac{1 \cdot 0 + 4 \cdot 1 + 7 \cdot 2 + 11 \cdot 3 + 10 \cdot 4 + 9 \cdot 5 + 5 \cdot 6 + 3 \cdot 7 + 2 \cdot 8 + 1 \cdot 9}{1 + 4 + 7 + 11 + 10 + 9 + 5 + 3 + 2 + 1} \\ &= \frac{212}{53} \\ &= 4.\end{aligned}$$

L'estimation de μ est donc : $\hat{\mu} = 4$.
La variance empirique est

$$\begin{aligned} v_x^2 &= \frac{1 \cdot 0^2 + 4 \cdot 1^2 + 7 \cdot 2^2 + 11 \cdot 3^2 + 10 \cdot 4^2 + 9 \cdot 5^2 + 5 \cdot 6^2 + 3 \cdot 7^2 + 2 \cdot 8^2 + 1 \cdot 9^2}{53} - \bar{x}^2 \\ &= \frac{212}{53} - 4^2 \\ &= \frac{1052}{53} - 16 \\ &= 19.85 - 16 \approx 3.85. \end{aligned}$$

Or, on sait que cet estimateur est biaisé. On le débiaise en multipliant par $n/(n-1)$:

$$s_x^2 = \frac{n}{n-1} v_x^2 \approx \frac{53}{52} 3.85 \approx 3.92$$

L'estimation de σ^2 est donc : $\hat{\sigma}^2 = 3.92$

(ii) Quelle type de loi proposez-vous pour décrire ce nombre d'appels ?
Réponse : la loi de Poisson $\mathcal{P}(4)$ semble appropriée car pour la loi de Poisson, l'espérance et la variance sont égales, ce qui est approximativement le cas ici. De plus, la loi de Poisson est bien adaptée pour modéliser un nombre d'événement indépendants dans un intervalle de temps fixé.

Exercice 5 Une société de vente à distance de matériel informatique s'intéresse au nombre journalier de connexions sur son site internet. Sur une période de 10 jours, les nombres suivants ont été relevés :

759 750 755 756 761 765 770 752 760 767

On suppose que ces résultats sont indépendants et identiquement distribués selon une loi normale d'espérance μ et de variance σ^2.

(i) Donner une estimation ponctuelle de l'espérance μ et de la variance σ^2 du nombre journalier de connexions. Réponse : la moyenne empirique est de $\bar{x} = 759.5$ et la variance empirique débiaisée est $s_x^2 = 42.056$. On peut prendre ces résultats comme estimations respectives de l'espérance et de la variance.

(ii) Construire un intervalle de confiance pour μ avec les niveaux de confiance 0.90 et 0.99.
Réponses : on est dans la situation où la variance est inconnue. On proc'ede donc de la manière suivante. On note X la variable "nombre de connexions". C'est une variable discrète, mais on va faire l'approximation qu'elle est Gaussienne $\mathcal{N}(\mu, \sigma^2)$ car il n'y a pas de valeurs répétées, (ce qui nous incite a choisir un modèle de variable continue) et que la loi Gaussienne est bien pratique. On a alors que

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \approx \mathcal{N}(0, 1).$$

et en remplacant σ^2 par son estimateur non-biaisé, on obtient

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{S_x^2}{n}}} \approx t_{n-1}.$$

où t_d représente la loi de Student à d degrés de liberté et $t_{d,\beta}$ représente le quantile de student au niveau β. On obtient alors sur la Table de la loi de Student les quantiles $t_{1-\frac{\alpha}{2}} = 1.833$ et par symétrie, $t_{\frac{\alpha}{2}} = -1.833$ et comme

$$\mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} \leq Z \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

on obtient en réarrangeant les termes que

$$\mathbb{P}\left(\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{S_x^2}{n}} \leq \mu \leq \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{S_x^2}{n}}\right) = 1-\alpha$$

et donc l'intervalle de confiance

$$I_\alpha = \left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{S_x^2}{n}}\ ;\ \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{S_x^2}{n}}\right]$$

ce qui donne après calculs :

$$\begin{aligned} i_{.90} &= \left[759.5 - 1.833\sqrt{\frac{42.06}{10}}\ ;\ 759.5 + 1.833\sqrt{\frac{42.06}{10}}\right] \\ &= [755.74; 763.26]. \end{aligned}$$

Au niveau de confiance .99, seule la valeur de $t_{n-1,1-\frac{\alpha}{2}}$ change : 3.250. On obtient alors

$$\begin{aligned} i_{.95} &= \left[759.5 - 3.250\sqrt{\frac{42.06}{10}}\ ;\ 759.5 + 3.250\sqrt{\frac{42.06}{10}}\right] \\ &= [752.83; 766.16]. \end{aligned}$$

Notez qu'il est normal qu'un intervalle associé à une confiance plus grande soit plus large. Pour une confiance de 1, on aurait l'intervalle $i_1 = [-\infty, +\infty]$. Il est remarquable que pour une confiance de .90 ou .99, les intervalles soient de taille raisonnable, ce qui reflète que les fluctuations sont bien concentrées autour de l'espérance.

(iii) Quel niveau de confiance choisir pour avoir un intervalle de confiance deux fois plus étroit que celui obtenu avec une confiance de 0.9 ?
Pour avoir un intervalle deux fois plus petit, il faut que $t_{n-1,1-\frac{\alpha}{2}}$ soit deux fois plus petit, c'est à dire .9165. On regarde dans la table et on trouve deux valeurs qui encadrent .9165, et qui sont 0.703 et 1.383. Par interpolation linéaire, on peut prendre une valeur de niveau égale à $.75 + \beta \cdot (.9 - .75)$ où $\beta = (.9165 - .703)/(1.383 - .703)$ et on obtient $1 - \dfrac{\alpha}{2} = .797$.On en déduit $\alpha = .41$. Donc un niveau de confiance de 59%

(iv) Sur combien de jours aurait-on dû relever le nombre de connexions pour que la longueur de l'intervalle de confiance, de niveau de 95%, n'excède pas 1 (en supposant que les estimations des moyennes et variances ne changent pas).
Réponse : il faut que la longuer de i_c n'excède pas 1, c'est à dire,

$$2\ t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n}} \leq 1.$$

Cela donne :

$$n \geq 2^2 \cdot t^2_{n-1,1-\frac{\alpha}{2}} \; s_x^2$$

Or, le quantile $t_{n-1,1-\frac{\alpha}{2}}$ dépend de n. Mais comme n sera grand, on peut l'approcher par le quantile de la $\mathcal{N}(0,1)$:

$$\begin{aligned} n &\geq 4 \cdot u^2_{1-\frac{\alpha}{2}} \; s_x^2 \\ n &\geq 4 \cdot 1.96^2 \cdot 46.73 \\ n &\geq 674.78 \end{aligned}$$

(v) Supposons maintenant que l'on connaisse la variance, donnée par le constructeur : $\sigma^2 = 42$ (en pratique, cela n'arrive jamais...). Cela aurait-il un impact sur vos intervalles de confiances ? Réponse : si la variance est connue, on peut utiliser la loi normale :

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \approx \mathcal{N}(0,1).$$

Donc, on a

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq u_{1-\frac{\alpha}{2}} \right) = 1 - \alpha.$$

Donc en isolant μ, on obtient

$$\mathbb{P}\left(\bar{X} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right) = 1 - \alpha.$$

L'intervalle de confiance est alors donné par

$$I_\alpha = \left[\bar{X} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \; ; \; \bar{X} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

et donc

$$i_\alpha = \left[\bar{x} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \; ; \; \bar{x} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

où on remplace $t_{n-1,1-\frac{\alpha}{2}}$ par $u_{1-\frac{\alpha}{2}}$. Par exemple $u_{1-\frac{.1}{2}} = 1.64$ et $u_{1-\frac{.05}{2}} = 1.96$. Dans le premier cas, i.e. $\alpha = .1$,

$$i_\alpha = \left[759.5 - 1.64 \sqrt{\frac{42}{10}} \; ; \; \bar{x} + 1.64 \sqrt{\frac{42}{10}} \right]$$

ce qui donne

$$i_\alpha = [756.139\ ;\ 762.861]$$

Dans le deuxième cas, i.e. $\alpha = .05$, on a

$$i_\alpha = \left[759.5 - 1.96\sqrt{\frac{42}{10}}\ ;\ \bar{x} + 1.96\sqrt{\frac{42}{10}}\right]$$

ce qui donne

$$i_\alpha = [755.5\ ;\ 763.52]\,.$$

Exercice 6 Une firme d'expertises en contrôle des matériaux a été mandatée par une société de gérance de projets de construction pour évaluer la qualité d'un mélange bitumineux provenant de deux usines. Il a été convenu d'effectuer une vérification par 115 mètres cubes de béton et d'évaluer la résistance à la compression, à l'âge de 3 jours, sur des cylindres standards. Les résultats de la résistance à la compression en $kg = cm^2$ pour les deux usines se résument comme suit.

	Usine 1	Usine 2
nombre de cylindres	$n_1 = 25$	$n_2 = 23$
moyenne empirique de la résistance	$\bar{x}_1 = 90.6$	$\bar{x}_2 = 94.4$
variance empirique de la résistance	$v_1^2 = 65.42$	$v_2^2 = 58.24$
estimateur débiasé de la variance de la résistance	$s_1^2 = 68.14$	$s_2^2 = 60.89$

On suppose que la résistance à la compression est distribuée normalement quelque soit l'usine de fabrication.

(i) Construire un intervalle de confiance pour la variabilité de la résistance à la compression du béton provenant de chaque usine, au niveau de confiance 0.95.
Réponse : la variabilité est une autre dénomination plus populaire pour la notion de fluctuation. On peut choisir pour répondre à cette question de considèrer un intervalle de confiance sur la variance. Pour cela, on sait que $(n_1 - 1)S_1^2/\sigma_1^2 \sim \chi^2_{n_1-1}$, et donc

$$\mathbb{P}\left(k_{\frac{\alpha}{2}} \leq \frac{(n_1-1)S_1^2}{\sigma_1^2} \leq k_{1-\frac{\alpha}{2}}\right) = 1 - \alpha. \tag{1}$$

où $k_{\frac{\alpha}{2}}$ dénote ici le quantile de la loi du $\chi^2_{n_1-1}$ d'ordre $\dfrac{\alpha}{2}$.
On en déduit que

$$\mathbb{P}\left(\frac{(n_1-1)S_1^2}{k_{1-\frac{\alpha}{2}}} \leq \sigma_1^2 \leq \frac{(n_1-1)S_1^2}{k_{\frac{\alpha}{2}}}\right) = 1 - \alpha. \tag{2}$$

et cela donne un intervalle de confiance au niveau $1 - \alpha$

$$I_c = \left[\frac{(n_1-1)S_1^2}{k_{n_1-1,1-\frac{\alpha}{2}}}; \frac{(n_1-1)S_1^2}{k_{n_1-1,\frac{\alpha}{2}}}\right]$$

On cherche alors dans la table du $\chi^2_{n_1-1}$ et on trouve $k_{\frac{\alpha}{2}} = 12.4$ et $k_{1-\frac{\alpha}{2}} = 39.4$. On en déduit le résultat numérique

$$\begin{aligned} i_c &= \left[\frac{(n_1-1)s_1^2}{k_{n_1-1,1-\frac{\alpha}{2}}}; \frac{(n_1-1)s_1^2}{k_{n_1-1,\frac{\alpha}{2}}}\right] \\ &= \left[\frac{24 \cdot 68.14}{39.4}; \frac{24 \cdot 68.14}{12.4}\right] \\ &= [41.51; 131.88]. \end{aligned}$$

Dans le cas de la seconde usine, la même procédure donne

$$\begin{aligned} i_c &= \left[\frac{(n_2-1)s_2^2}{k_{n_2-1,1-\frac{\alpha}{2}}}; \frac{(n_2-1)s_2^2}{k_{n_2-1,\frac{\alpha}{2}}}\right] \\ &= \left[\frac{22 \cdot 60.89}{36.8}; \frac{22 \cdot 60.89}{11}\right] \\ &= [36.40; 121.78]. \end{aligned}$$

(ii) Peut-on en déduire que la variabilité de la résistance à la compression du béton provenant de chaque usine est différente ?
Réponse : les intervalles se chevauchent substantiellement et donc on ne peut pas considèrer que les variances sont différentes.

(iii) Déterminer un intervalle de confiance pour le rapport des deux variances, σ_1^2/σ_2^2, avec un niveau de confiance de 95%.
Réponse : On propose comme estimateur de σ_1^2/σ_2^2 le rapport S_1^2/S_2^2. On sait que le rapport de deux χ^2 divisées par leurs degrés de liberté suit une loi de Fisher. Plus exactement, on a

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

On regarde alors les valeur des quantiles de la loi de Fisher au niveau .025 et .975 : $f_{24,22,.975} = 2.33$ et $f_{24,22,.025} = 1/f_{22,24,.975}$. Comme $f_{22,24,.975}$ n'est pas disponible dans la table, on l'approxime par la moyenne de $f_{20,24,.975} = 2.33$ et $f_{24,24,.975} = 2.27$ soit $f_{24,22,.975} \simeq 2.3$ et $f_{24,22,.025} = 1/2.3$.
On obtient donc

$$\mathbb{P}\left(f_{24,22,.025} \leq \frac{S_1^2}{S_2^2}\,\frac{\sigma_2^2}{\sigma_1^2} \leq f_{24,22,.975}\right) = 1-\alpha.$$

et ainsi,

$$\mathbb{P}\left(f_{24,22,.025}\,\frac{S_2^2}{S_1^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_{24,22,.975}\,\frac{S_2^2}{S_1^2}\right) = 1-\alpha.$$

d'où

$$\mathbb{P}\left(\frac{1}{f_{24,22,.975}}\ \frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{f_{24,22,.025}}\ \frac{S_1^2}{S_2^2}\right) = 1-\alpha.$$

On choisit donc un intervalle de confiance

$$\begin{aligned} I_c = i_c &= \left[\frac{1}{f_{24,22,.975}}\ \frac{s_1^2/\nu_1}{s_2^2/\nu 2}; \frac{1}{f_{24,22,.025}}\ \frac{s_1^2}{s_2^2}\right] \\ &= [\frac{1}{2.33}\cdot 1.12\ ;\ \frac{1}{(1/2.3)}\cdot\ 1.12] \\ &= [.48; 2.58]. \end{aligned}$$

Exercice 7 Lors d'un sondage précédant les élections présidentielles, 500 personnes ont été interrogées. Bien que ce ne soit pas le cas en pratique, on suppose pour simplifier les calculs que les 500 personnes constituent un échantillon indépendant et identiquement distribué de la population française. Sur les 500 personnes, 150 ont répondu vouloir voter pour le candidat C1, et 140 pour le candidat C2.

(i) Donner une estimation ponctuelle des intentions de votes pour chaque candidat, sous la forme d'un pourcentage.
Réponse : Pour le candidat numéro 1, on construit les variables aléatoires indépendantes et identiquement distribuées de Bernoulli $\mathcal{B}(1, p_1)$, $X_1, \ldots, X_n$. On sait que $S_1 = X_1 + \cdots + X_n$ suit une loi Binomiale $\mathcal{B}(n, p_1)$ d'espérance $n\ p_1$ et donc que

$$\mathbb{E}\left[\frac{S_1}{n}\right] = p_1$$

Il est donc raisonnable de choisir la proportion des gens ayant voté pour le candidat 1, i.e. $F_1 = S_1/n$ comme estimateur de p_1. Numériquement, on obtient

$$f_1 = 150/500 = .3.$$

De même, on a

$$f_2 = 140/500 = .28.$$

pour l'estimation de p_2.

(ii) Donner un intervalle de confiance à 95% pour chacun des deux intentions de votes.
Réponse : Notons

$$F_1 = \frac{1}{n}\ (X_1 + \cdots + X_n) = \bar{X}. \tag{3}$$

Cette fréquence aléatoire F_1 est donc une moyenne et donc a pour loi une loi proche de la loi Gaussienne par le théorème central limite. On connaît son espérance

$$\mathbb{E}\left[F_1\right] = p_1. \tag{4}$$

On note que sa variance est donnée par

$$\begin{aligned}\mathrm{Var}(F_1) &= \frac{n\ p_1(1-p_1)}{n^2}\\ &= \frac{p_1(1-p_1)}{n_1}.\end{aligned}$$

En faisant l'approximation Gaussienne (puisque n est ici grand) disant que

$$F_1 \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right).$$

on peut obtenir un intervalle de confiance de la manière suivante. On regarde la table de la loi normale et on otient $u_{1-\frac{\alpha}{2}} = 1.96$. On a donc par symétrie

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{F_1 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Comme $p_1 \approx f_1$ par la loi des grands nombres, on peut faire l'approximation

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{F_1 - p_1}{\sqrt{\frac{f_1(1-f_1)}{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Ainsi, on peut isoler p_1 et obtenir

$$\mathbb{P}\left(F_1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}} \leq p_1 \leq F_1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}}\right) = 1-\alpha$$

ce qui nous conduit à l'intervalle de confiance

$$I_c = \left[F_1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}}\ ;\ F_1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}}\right].$$

Numériquement, on obtient

$$\begin{aligned}i_c &= \left[f_1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}}\ ;\ f_1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n}}\right]\\ &= \left[.3 - 1.96\cdot\sqrt{\frac{.3(1-.3)}{500}}\ ;\ .3 + 1.96\cdot\sqrt{\frac{.3(1-.3)}{500}}\right]\\ &= [.2598; .3402]\end{aligned}$$

De même pour le deuxième candidat. On obtient

$$\begin{aligned}i_c &= \left[f_2 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_2(1-f_2)}{n}}\ ;\ f_2 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f_2(1-f_2)}{n}}\right]\\ &= \left[.28 - 1.96\cdot\sqrt{\frac{.28(1-.28)}{500}}\ ;\ .3 + 1.96\cdot\sqrt{\frac{.28(1-.28)}{500}}\right]\\ &= [.2406; .3193]\end{aligned}$$

(iii) Peut-on prédire l'élection d'un candidat ?

Réponse : Les intervalles respectifs de p_1 et p_2 se chevauchant substantiellement, on ne peut pas conclure qu'un candidat va prendre le dessus sur l'autre. On verra que la théorie des tests nous permettra de poser la question de manière plus rigoureuse.

Exercice 8 Nous cherchons à modéliser le nombre de SMS reçus pendant une séance de cours. Nous supposons que ce nombre de SMS suit une loi de Poisson de paramètre λ. Nous cherchons à estimer le paramètre λ en construisant l'estimateur du maximum de vraisemblance.

(i) Que représente λ ?

Réponse : Le paramètre λ est l'espérance de la loi de Poisson $\mathcal{P}(\lambda)$. Pour cette raison, et en utilisant la loi des grands nombres, on peut déjà anticiper que la moyenne sur un échantillon $x_1, \dots, x_n$ devrait nous donner une estimation raisonnable de λ.

(ii) Ecrire la fonction de vraisemblance d'un échantillon $X_1, \dots, X_n$.

Réponse : En supposant l'indépendance des X_i, $i = 1, \dots, n$, la fonction de vraisemblance associée à cet échantillon est donnée par

$$L_{X_1,\dots,X_n}(\lambda) = \prod_{i=1}^{n} f_\lambda(X_i)$$

où $f_(\lambda)$ est la densité de la loi de Poisson, donnée par

$$f_\lambda(x) = e^{-\lambda}\ \frac{\lambda^x}{x!}.$$

La vraisemblance peut donc s'écrire

$$\begin{aligned} L_{X_1,\dots,X_n}(\lambda) &= \prod_{i=1}^{n} \exp(-\lambda)\ \frac{\lambda^{X_i}}{X_i!} \\ &= \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!} \end{aligned}$$

(iii) Un relevé effectué sur 10 étudiants donne les nombres de SMS suivant : 1, 4, 1, 3, 1, 6, 3, 0, 0, 0. Représenter ces données sous la forme d'un histogramme.

Réponse :

Calculer la vraisemblance en fonction de λ pour cet échantillon.

Réponse : Pour cet échantillon, la vraisemblance est donnée par

$$L_{X_1,\dots,X_n}(\lambda) = \left(\exp(-\lambda)\ \frac{\lambda^0}{0!}\right)^3 \cdot \left(\exp(-\lambda)\ \frac{\lambda^1}{1!}\right)^3 \cdot \left(\exp(-\lambda)\ \frac{\lambda^3}{3!}\right)^2 \cdot \left(\exp(-\lambda)\ \frac{\lambda^4}{4!}\right)^1 \cdot \left(\exp(-\lambda)\ \frac{\lambda^6}{6!}\right)^1$$

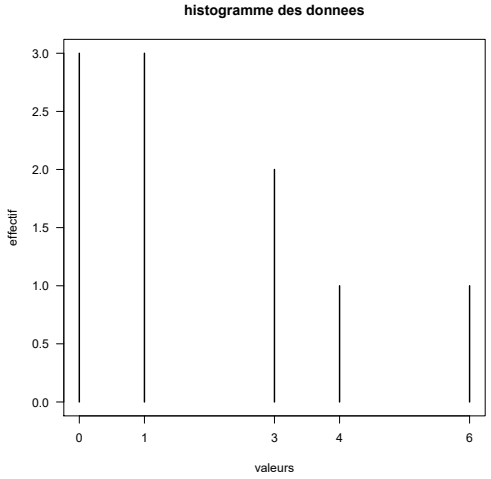

FIGURE 2 – Histogramme pour les SMS. On peut aussi utiliser un diagramme en bâtons.

Que vaut-elle pour $\lambda = 1$ et $\lambda = 2$?

Réponse : Un code en R est donné ci-dessous pour vous montrer que cela se programme rapidement et facilement

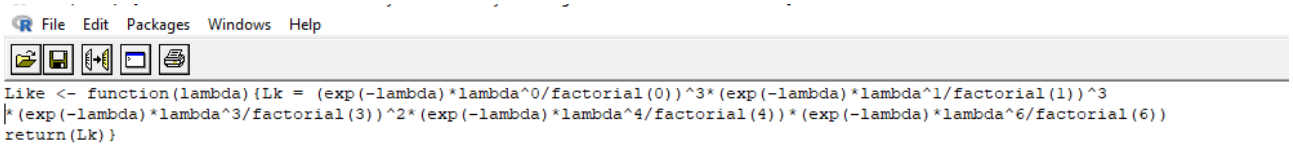

FIGURE 3 – Code en R pour calculer la fonction de vraisemblance.

On trouve en utilisant cette fonction que la valeur de la vraisemblance en $\lambda = 1$ est 7.298085e-11 et avec $\lambda = 2$ on obtient 1.737137e-09.

Quel est la valeur de λ qui maximise la vraisemblance de cet échantillon ?

Réponse : entre ces deux valeurs, $\lambda = 2$ donne la plus grande vraisemblance. On verra dans

la question suivante comment maximiser la vraisemblance pour toutes les valeurs possibles.

(iv) Reprenons l'écriture générique de la vraisemblance d'un échantillon $X_1, \ldots, X_n$. En écrivant le logarithme de la vraisemblance, calculer théoriquement l'estimateur T du maximum de vraisemblance.

Réponse : Le logarithme de la vraisemblance, encore appelé "log-vraisemblance" s'écrit

$$\begin{aligned}\ell_{X_1,\ldots,X_n}(\lambda) &= \sum_{i=1}^{n} -\lambda + \log(\lambda)X_i - \log(X_i!) \\ &= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\log(X_i!).\end{aligned}$$

L'estimateur au sens du maximum de vraisemblance de λ, noté T est la valeur qui maximise la log-vraisemblance. Maximiser la log-vraisemblance ou la vraisemblance est équivalent car la fonction "log" est croissante. Pour l'obtenir, on calcule la dérivée de la log-vraisemblance :

$$\ell'_{X_1,\ldots,X_n}(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i.$$

La valeur de T est celle pour laquelle on obtient $\ell'_{X_1,\ldots,X_n}(T) = 0$, c'est à dire,

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(v) Déduisez-en l'estimation par maximum de vraisemblance de λ.

Réponse : L'estimation est obtenue en remplaçant T par la valeur obtenue en remplaçant les variables aléatoires X_i par les données x_i, $i = 1, \ldots, n$, c'est à dire

$$\begin{aligned}t = \bar{x} &= (3\cdot 0 + 3\cdot 1 + 2\cdot 3 + 4 + 6)/10 \\ &= 1.9.\end{aligned}$$

(vi) Quels sont l'espérance et la variance de T ?

Réponse : On a

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right], &(5)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i], &(6)\\ &= \frac{1}{n}\sum_{i=1}^{n}\lambda, &(7)\\ &= \lambda. &(8)\end{aligned}$$

On constate que l'estimateur T a exactement pour espérance la valeur du paramètre qu'il est censé estimer. On dit alors que cet estimateur est "sans biais", ou encore "non-biaisé". La variance est donnée par

$$\begin{aligned} Var(T) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), && (9)\\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \quad \text{par independance}, && (10)\\ &= \frac{1}{n^2}\sum_{i=1}^{n} \lambda, && (11)\\ &= \frac{1}{n}\,\lambda. && (12) \end{aligned}$$

(vii) Calculer l'information de Fisher apportée sur le paramètre λ par un n-chantillon $X_1, \ldots, X_n$.

Réponse : L'information de Fisher est "moins l'espérance de la dérivée seconde", c'est à dire

$$\begin{aligned} I &= -\mathbb{E}\left[\ell''_{X_1,\ldots,X_n}(\lambda)\right], && (13)\\ &= \mathbb{E}\left[\frac{1}{\lambda^2}\sum_{i=1}^{n} X_i\right], && (14)\\ &= \frac{1}{\lambda^2}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right], && (15)\\ &= \frac{1}{\lambda^2}\sum_{i=1}^{n} \lambda, && (16)\\ &= \frac{n}{\lambda}. && (17) \end{aligned}$$

(viii) En déduire que l'estimateur T est un estimateur sans biais de variance minimale (i.e. efficace) de λ.

Réponse : Le théorème de Cramer-Rao dit que pour un estimateur non-biaisé, sa variance ne peut pas être inférieure a l'inverse de l'information de Fisher. Dans la cas présent, on constate que la variance de T est exactement égale à cette borne inférieure, ce qui nous permet de conclure que T a la variance la plus faible possible, ce qui caractérise le fait que T est efficace.