M1 - UE8-2 - MÉTHODES AVANCÉES EN STATISTIQUE INFÉRENTIELLE

LOI D'ÉCHANTILLONNAGE

RECONNAISSANCE DE LOIS

EXERCICE 1.
Soient X_1, X_2 et X_3 3 variables aléatoires indépendantes de lois respectives $\mathcal{N}(\mu_1\ ;\ \sigma_1)$, $\mathcal{N}(\mu_2\ ;\ \sigma_2)$ et $\mathcal{N}(\mu_3\ ;\ \sigma_3)$. Donner les lois des variables aléatoires suivantes :

$$T_1 = 2X_1, \qquad T_2 = X_1 + X_2, \qquad T_3 = \frac{1}{4}(X_1 - X_2), \qquad T_4 = \frac{X_1 - \mu_2}{\sigma_1}, \qquad T_5 = \frac{X_3 - \mu_3}{\sigma_3},$$
$$T_6 = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2, \qquad T_7 = \frac{T_5}{\sqrt{\frac{1}{2}T_6}}.$$

$T_1 \sim \mathcal{N}(2\mu_1\ ;\ 2\sigma_1)$; $T_2 \sim \mathcal{N}\left(\mu_1 + \mu_2\ ;\ \sqrt{\sigma_1^2 + \sigma_2^2}\right)$; $T_3 \sim \mathcal{N}\left(\frac{\mu_1 - \mu_2}{4}\ ;\ \frac{1}{4}\sqrt{\sigma_1^2 + \sigma_2^2}\right)$; $T_4 \sim \mathcal{N}\left(\frac{\mu_1 - \mu_2}{\sigma_1}\ ;\ 1\right)$; $T_5 \sim \mathcal{N}(0\ ;\ 1)$; $T_6 \sim \chi^2(2)$; $T_7 \sim t(2)$.

EXERCICE 2.
On considère deux n-échantillons $X_1, X_2, \ldots, X_n$ et $Y_1, \ldots, Y_n$ indépendants entre eux de loi mère pour le premier $\mathcal{N}(3; 4)$ et pour le second $\mathcal{N}(2; 3)$.
On pose $n = 25$. Calculer les probabilités suivantes.

1. $\mathbb{P}\left(\overline{X} > 4, \overline{Y} < 1\right)$;

   On utilise l'indépendance mutuelles entre $X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_n$.

   $$\mathbb{P}(\overline{X} > 4, \overline{Y} < 1) = \mathbb{P}(\overline{X} > 4) \times \mathbb{P}(\overline{Y} < 1) = \mathbb{P}\left(\frac{\overline{X} - 3}{4/5} > \frac{4-3}{4/5}\right) \times \mathbb{P}\left(\frac{\overline{Y} - 2}{3/5} < \frac{1-2}{3/5}\right)$$
   $$= (1 - F(1.25)) \times F\left(-\frac{5}{3}\right) = (1 - F(1.25)) \times \left(1 - F\left(\frac{5}{3}\right)\right)$$
   $$\simeq (1 - 0.8944) \times (1 - 0.9525) \simeq 0.005$$

2. $\mathbb{P}\left(S^2_{X,cor} > 26.25\right)$;

   On a $\frac{24 S^2_{X,cor}}{16} \sim \chi^2(24)$.

   $$\mathbb{P}(S^2_{X,cor} > 26.25) = \mathbb{P}\left(\frac{24 S^2_{X,cor}}{16} > \frac{24 \times 26.25}{16}\right) = \mathbb{P}\left(\frac{24 S^2_{X,cor}}{16} > 39.375\right) \simeq 0.025$$

3. $\mathbb{P}\left(\overline{Y} > 0.5, S^2_{Y,cor} < 12.45\right)$;

   On a vu que si la loi mère est normale alors l'espérance empirique et la variance empirique corrigée sont indépendantes.

   $$\mathbb{P}\left(\overline{Y} > 0.5, S^2_{Y,cor} < 12.45\right) = \mathbb{P}\left(\overline{Y} > 0.5\right) \times \mathbb{P}\left(S^2_{Y,cor} < 12.45\right)$$
   $$= \mathbb{P}\left(\frac{\overline{Y} - 2}{3/5} > \frac{0.5 - 2}{3/5}\right) \times \mathbb{P}\left(\frac{24 S^2_{Y,cor}}{9} < \frac{24 \times 12.45}{9}\right)$$
   $$= F(2.5) \times \mathbb{P}\left(\frac{24 S^2_{Y,cor}}{9} < 33.2\right) = 0.9946 \times 0.9 \simeq 0.895$$

4. $\mathbb{P}\left(\overline{X} < 3 - 0.137 S_{X,cor}\right)$.

   On a $\dfrac{\overline{X} - 3}{S_{X,cor}/\sqrt{25}} \sim t(24)$.

$$\mathbb{P}\left(\overline{X} < 3 - 0.137 S_{X,cor}\right) = \mathbb{P}\left(\frac{\overline{X} - 3}{S_{X,cor}/5} < -0.137 \times 5\right) = \mathbb{P}\left(\frac{\overline{X} - 3}{S_{X,cor}/5} < -0.685\right)$$
$$= \mathbb{P}\left(\frac{\overline{X} - 3}{S_{X,cor}/5} > 0.685\right) = 0.25$$

Exercices avec contexte

Exercice 3.

Sur l'ensemble des stations-service d'une région, le prix moyen d'un litre d'essence est de 1.40 euros et son écart type est de 0.15. De plus, on suppose que les prix suivent une loi normale.
Un échantillon aléatoire de 20 stations-service est sélectionné et on considère les événements suivants :

A : « le prix observé dans la première station-service est supérieur à 1.45 euros » ;

B : « le prix moyen sur les 20 stations étudiées est supérieur à 1.45 euros » ;

C : « le prix le plus faible observé sur les 20 stations est supérieur à 1.25 euros » ;

D : « la variance empirique corrigée du prix moyen sur les 20 stations étudiées est inférieur à 0.036 ».

Pour chacun de ces événements :

1. donner la statistique permettant de décrire l'événement ;
2. si c'est possible, donner la loi de cette statistique (famille de loi et paramètre(s)) ;
3. calculer la probabilité de l'événement.

A : « le prix observé dans la première station-service est supérieur à 1.45 euros » ;

La statistique à considérer est X_1. Notons que $X_1 \sim \mathcal{N}(1.40; 0.15)$. On a alors

$$\mathbb{P}(X_1 \geqslant 1.45) = 1 - F\left(\frac{1.45 - 1.4}{0.15}\right) = 1 - F\left(\frac{1}{3}\right) \simeq 1 - 0.63 \simeq 0.37$$

Il y a environ 37% de chance que la première station-service sélectionnée affiche un prix supérieur à 1.45 euros.

B : « le prix moyen sur les 20 stations étudiées est supérieur à 1.45 euros » ;

La statistique à considérer est la moyenne empirique $\overline{X}_{20}$. On a $\overline{X}_{20} \sim \mathcal{N}\left(1.40; \dfrac{0.15}{\sqrt{20}}\right)$. On a alors

$$\mathbb{P}\left(\overline{X}_{20} \geqslant 1.45\right) = 1 - F\left(\frac{1.45 - 1.4}{\frac{0.15}{\sqrt{20}}}\right) = 1 - F(1.49) = 1 - 0.9319 = 0.0684$$

Il y a environ 7% de chance qu'en moyenne le prix observé sur les 20 stations service soit supérieur à 1.45 euros.

C : « le prix le plus faible observé sur les 20 stations est supérieur à 1.25 euros » ;

La statistique à considérer est la première statistique d'ordre $X_{(1)}$.

$$\mathbb{P}\left(X_{(1)} \geqslant 1.25\right) = \prod_{i=1}^{20} \mathbb{P}(X_i \geqslant 1.25) = \prod_{i=1}^{20} \left(F(1)\right) = (0.8413)^{20} \simeq 0.0316$$

Il y a environ 3% de chance que le prix le plus faible observé soit supérieur à 1.25 euros.

D : « la variance empirique corrigée du prix moyen sur les 20 stations étudiées est inférieure à 0.036 ».

La statistique à considérer est la variance empirique corrigée S^2_{cor}. On a $\dfrac{19 S^2_{cor}}{0.15^2} \sim \chi^2(19)$. On obtient

$$\mathbb{P}(S^2_{cor} \leqslant 0.036) = \mathbb{P}\left(\frac{19 S^2_{cor}}{0.15^2} \leqslant \frac{19 \times 0.036}{0.15^2}\right) = \mathbb{P}\left(\frac{19 S^2_{cor}}{0.15^2} \leqslant 30.4\right) \simeq 1 - 0.05 \simeq 0.95$$

Donc il y a environ 95% de chance que l'observation de la variance empirique corrigée soit inférieure à 0.036 (cela correspond à un écart type corrigé inférieur à 0.19).

EXERCICE 4.
Un astronome souhaite mesurer la distance, en années-lumière, entre son observatoire et une étoile lointaine. Bien qu'il connaisse une technique de mesure, il sait aussi que chaque résultat ne constitue qu'une distance approchée, en raison des influences atmosphériques et d'autre causes d'erreur inévitables.
Par conséquent, notre astronome prévoit de prendre plusieurs mesures et d'accepter leur moyenne comme estimation de la distance réelle. Il a des raisons de penser que les différentes valeurs mesurées sont des variables aléatoires indépendantes et identiquement distribuées d'espérance commune d (la vraie valeur) et de variance commune 4 (l'unité étant toujours l'année lumière).
Combien de mesures doit-il réaliser pour être sûr à 95% que l'erreur soit inférieure à une demi-année-lumière ?

Les différentes mesures correspondent à un n-échantillon $X_1, X_2, \dots, X_n$ tel que la loi mère a pour espérance d et écart type 2. On cherche le nombre de mesures n à effectuer tel que $\mathbb{P}\left(\left|\overline{X}_n - d\right| < 0.5\right) \geqslant 0.95$.

$$\mathbb{P}\left(\left|\overline{X}_n - d\right| < 0.5\right) \geqslant 0.95 \Leftrightarrow \mathbb{P}\left(-0.5 < \overline{X}_n - d < 0.5\right) \geqslant 0.95 \Leftrightarrow \mathbb{P}\left(\frac{-0.5}{2/\sqrt{n}} < \overline{X}_n - d < \frac{0.5}{2/\sqrt{n}}\right) \geqslant 0.95$$
$$\Leftrightarrow 2F\left(\frac{\sqrt{n}}{4}\right) - 1 \geqslant 0.95 \Leftrightarrow F\left(\frac{\sqrt{n}}{4}\right) \geqslant 0.975 \Leftrightarrow \frac{\sqrt{n}}{4} \geqslant 1.96$$
$$\Leftrightarrow n \geqslant (1.96 \times 4)^2 = 61.4656$$

Il faudra donc que l'astronome effectue au moins 62 mesures pour s'assurer d'avoir 95% de chance que l'erreur soit inférieure à une demi-année-lumière.

EXERCICE 5.
Une compagnie d'assurance assure n personnes contre un même risque, de probabilité p. Si ce risque se réalise, la compagnie doit payer à l'assuré la somme M. La cotisation de chaque assuré est $M(p+a)$. On suppose que les sinistres sont indépendants.

1. Quelle est l'espérance du bénéfice de la compagnie ? Sa variance ?

   Soit X la variable aléatoire représentant le nombre d'assurés qui rencontre ce risque. On a $X \sim \text{Bin}(n; p)$.
   On note Y la variable représentant le bénéfice de la compagnie d'assurance. On a $Y = nM(p+a) - MX$.
   On a $\mathbb{E}(Y) = nM(p+a) - M\mathbb{E}(X) = nM(p+a) - Mnp = nMa$.
   De même, on a $\text{V}(Y) = M^2\text{V}(X) = M^2np(1-p)$.

2. La compagnie d'assurance fixe la valeur de a de telle façon qu'elle ne perde de l'argent que dans 0.1 % des cas.
   Déterminer, en fonction de n et p, la valeur de a pour que la compagnie ait une probabilité 0.001 de perdre de l'argent.

   On suppose que n est grand. Dans ce cas, la loi de X peut être approchée par une loi normale et donc celle de Y aussi. Cela donne $Y \underset{\text{approx}}{\sim} \mathcal{N}\left(nMa; M\sqrt{p(1-p)}\right)$. Commençons par exprimer la probabilité citée en fonction de n et p.

   $$\mathbb{P}(Y < 0) = \mathbb{P}\left(\frac{Y - nMa}{M\sqrt{np(1-p)}} < -\frac{na}{\sqrt{np(1-p)}}\right) \simeq F\left(-\frac{na}{\sqrt{np(1-p)}}\right) \simeq 1 - F\left(\frac{\sqrt{n}a}{\sqrt{p(1-p)}}\right)$$

   On a donc

   $$1 - F\left(\frac{\sqrt{n}a}{\sqrt{p(1-p)}}\right) = 0.001 \Leftrightarrow F\left(\frac{\sqrt{n}a}{\sqrt{p(1-p)}}\right) = 0.999 \Leftrightarrow \frac{\sqrt{n}a}{\sqrt{p(1-p)}} = 3.1$$
   $$\Leftrightarrow a = 3.1\sqrt{\frac{p(1-p)}{n}}$$

3. Les assurés ont-ils intérêt à être nombreux ?

   Plus les assurés sont nombreux (n croît), plus la valeurs a diminue. Les assurés ont donc intérêt à être nombreux si la compagnie ajuste effectivement ses tarifs.

Exercice 6.
Soit (X_1, X_2) un couple de variables aléatoires de densité $f(x,y) = e^{-x}e^{-y}\,\mathbb{1}_{(\mathbb{R}_+^*)^2}(x,y)$.
On va s'intéresser à la loi des statistiques d'ordre : $X_{(1)} = \min(X_1, X_2)$ et $X_{(2)} = \max(X_1, X_2)$.

1. Donner les lois marginales : densités et fonctions de répartion marginales.

   On a $X_1 \sim \mathcal{E}(1)$ et $X_2 \sim \mathcal{E}(1)$.
   On n'a pas vu les lois exponentielles en L1.
   On a donc $f_{X_1}(x) = f_{X_2}(x) = e^{-x}\mathbb{1}_{\mathbb{R}_+^*}(x)$ et $F_{X_1}(x) = F_{X_2}(x) = (1-e^{-x})\,\mathbb{1}_{\mathbb{R}_+^*}(x)$.

2. Les variables aléatoires X_1 et X_2 sont-elles indépendantes entre elles ?

   Oui.

3. Donner les fonctions de répartition des deux statistiques d'ordre.

   $X_{(1)}(\Omega) = \mathbb{R}_+^*$. Soit $t \in \mathbb{R}_+^*$. On a $F_{X_{(1)}}(t) = \mathbb{P}(X_1 \leqslant t \text{ ou } X_2 \leqslant t) = 1 - (1-F(t))^2 = 1 - e^{-2z}$. Donc $X_{(1)} \sim \mathcal{E}(2)$.
   De même on obtient $F_{X_{(2)}}(t) = (1 - 2e^{-t} + e^{-2t})\mathbb{1}_{\mathbb{R}_+^*}(x)$.

Exercice 7.
On suppose que le temps avant qu'une personne panique lorsqu'elle se trouve dans un ascenseur bloqué suit une loi exponentielle de paramètre 0.05. Trois personnes dont on suppose la panique indépendante sont dans un ascenseur. L'ascenseur se bloque.

1. Quelle est la loi du temps avant qu'au moins une personne panique ?

   $F(x) = \left(1 - e^{-0.05x}\right)\mathbb{1}_{\mathbb{R}_+^*}$
   $F_{X_{(1)}}(t) = \left(1 - e^{-0.15t}\right)\mathbb{1}_{\mathbb{R}_+^*}$

2. Quelle est la probabilité que personne ne panique pendant les 10 premières minutes ?

   $\mathbb{P}\left(X_{(1)} \geqslant 10\right) = 1 - F_{X_{(1)}}(10) = e^{-1.5} \simeq 0.2231$

3. Quelle doit être la durée de l'intervention avant débloquage pour que plus de la moitié du temps personne ne panique ?

   On cherche le plus grand réel t tel que

$$P\left(X_{(1)} \geqslant t\right) \geqslant 0.5 \Leftrightarrow e^{-0.15t} \geqslant 0.5 \Leftrightarrow -0.15t \geqslant \ln(0.5) \Leftrightarrow t \leqslant -\frac{\ln(0.5)}{0.15}\ (\simeq 4.62)$$

   Il faut donc que l'intervention dure au plus 4 minutes et 37 secondes.

Exercice 8.
On considère un n échantillon $X_1, X_2, \ldots, X_n$ de loi mère à densité. Quelle est la probabilité que le maximum empirique dépasse la médiane de la loi mère ? le troisième quartile ? le neuvième décile ?

Comme la loi mère est à densité, on a que la médiane est égale à $F^{-1}(0.5)$. On cherche donc
$\mathbb{P}\left(X_{(n)} \geqslant F^{-1}(0.5)\right) = 1 - F_{X_{(n)}}\left(F^{-1}(0.5)\right) = 1 - \left(F\left(F^{-1}(0.5)\right)\right)^n = 1 - \frac{1}{2^n}$.
$\mathbb{P}\left(X_{(n)} \geqslant F^{-1}(0.75)\right) = 1 - \left(\frac{3}{4}\right)^n$
$\mathbb{P}\left(X_{(n)} \geqslant F^{-1}(0.9)\right) = 1 - \left(\frac{9}{10}\right)^n$

EXERCICE 9.
On considère un n échantillon $X_1, X_2, \dots, X_n$ de loi mère $\mathcal{U}([0;1])$.

1. Déterminer la fonction de répartition de $X_{(1)}$. En déduire sa densité.

   Soit $t \in [0;1]$. On a $F_{X_{(1)}}(t) = 1 - (1 - F(x))^n = 1 - (1-x)^n$.
   En dérivant on obtient $f_{X_{(1)}}(x) = n(1-x)^{n-1}\, \mathbb{1}_{]0;1[}(x)$.

2. Calculer son espérance.

$$\mathbb{E}\left(X_{(1)}\right) = \int_0^1 nx(1-x)^{n-1}\,dx = n\int_0^1 \left((1-x)^{n-1} - (1-x)^n\right)dx = n\left[-\frac{(1-x)^n}{n} + \frac{(1-x)^{n+1}}{n+1}\right]_0^1$$
$$= n\left(-0 + 0 + \frac{1}{n} - \frac{1}{n+1}\right) = 1 - \frac{n}{n+1} = \frac{n+1-n}{n+1} = \frac{1}{n+1}$$

3. Déterminer la fonction de répartition de $X_{(n)}$. En déduire sa densité, puis son espérance.

   Il s'agit maintenant d'étudier la loi de la dernière statistique d'ordre $X_{(n)}$.
   Soit $t \in [0;1]$. On a $F_{X_{(n)}}(t) = (F(x))^n = x^n$. En dérivant on obtient $f_{X_{(n)}}(x) = nx^{n-1}\, \mathbb{1}_{]0;1[}(x)$.

$$\mathbb{E}\left(X_{(n)}\right) = \int_0^1 nxx^{n-1}\,dx = n\int_0^1 x^n dx = n\left[\frac{x^{n+1}}{n+1}\right]_0^1 = \frac{n}{n+1} = 1 - \frac{1}{n+1}$$

EXERCICE 10.
Soient X_1 et X_2 deux variables aléatoires indépendantes de loi $\mathcal{U}(\llbracket 1;10 \rrbracket)$.

1. Quelle est la loi de $X_{(1)}$? Celle de $X_{(2)}$?

   On a $X_{(1)}(\Omega) = X_{(2)}(\Omega) = \llbracket 1;10 \rrbracket$. Soit $k \in \llbracket 1;10 \rrbracket$. On a

$$\mathbb{P}\left(X_{(1)} = k\right) = \mathbb{P}\left([X_1 = k] \cap [X_2 = k]\right) + \mathbb{P}\left([X_1 > k] \cap [X_2 = k]\right) + \mathbb{P}\left([X_1 = k] \cap [X_2 > k]\right)$$
$$= \frac{1}{100} + 2 \times \frac{1}{10} \times \frac{10-k}{10} = \frac{21-2k}{100}$$

   En procédant de la même façon, on obtient $\mathbb{P}\left(X_{(2)} = k\right) = \dfrac{2k-1}{100}$.

2. (*) Calculer son espérance.
   Indication : $\displaystyle\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

$$\mathbb{E}\left(X_{(1)}\right) = \sum_{k=1}^{10} k \times \frac{21-2k}{100} = \frac{21}{100}\sum_{k=1}^{10} k - \frac{2}{100}\sum_{k=1}^{10} k^2$$
$$= \frac{21}{100} \times 55 - \frac{1}{50} \times 385 = \frac{231}{20} - \frac{154}{20} = \frac{77}{20} = 3.85$$

EXERCICE 11.
Alexia et Anne ne sont pas des étudiantes très sérieuses. Chaque jour, de façon indépendante, elles arrivent avec plus d'un quart d'heure de retard : avec probabilité 0.02 pour Alexia et 0.01 pour Anne. On note X (respectivement Y) le numéro du premier jour où Alexia (respectivement Anne) arrive avec plus d'un quart d'heure de retard.
La politique de l'université est très sévère dès le premier gros retard (supérieur à 15 minutes) l'étudiant est interdit de cours.

1. On note Z la variable aléatoire correspondant au numéro de jour du premier gros retard d'Alexia ou de Anne.
   Déterminer la loi de Z.

   Soit $n \in \mathbb{N}^*$. On a $\mathbb{P}(X = n) = 0.98^{n-1} \times 0.02$ et $\mathbb{P}(X = n) = 0.99^{n-1} \times 0.01$.
   Ces lois sont appelées lois géométriques.
   On a $Z = \min(X, Y)$.
   $\mathbb{P}(Z > n) = \mathbb{P}(Y > n, X > n) = \mathbb{P}(X > n) \times \mathbb{P}(Y > n) = 0.98^n \times 0.99^n = 0.9702^n$.
   On constate que Z suit une loi géométrique de paramètre 0.0298.
   $Z(\Omega) = \mathbb{N}^*$ et, pour tout $n \in \mathbb{N}^*$, on a $\mathbb{P}(Z = n) = 0.9702^{n-1} \times 0.0298$.

2. On considère qu'un semestre comporte 40 jours de cours.
   (a) Quelle est la probabilité qu'au moins une des deux étudiantes soit exclues ?

   $\mathbb{P}(Z \geqslant 40) = 1 - \mathbb{P}(Z > 40) = 1 - 0.9702^{40} \simeq 0.702$

   (b) Quelle est la probabilité pour chaque étudiante d'être exclue ?

   $\mathbb{P}(X \geqslant 40) = 1 - \mathbb{P}(X > 40) = 1 - 0.98^{40} \simeq 0.554$
   $\mathbb{P}(Y \geqslant 40) = 1 - \mathbb{P}(Y > 40) = 1 - 0.99^{40} \simeq 0.331$