Statistique - TD1 - Loi normale et TCL

Exercice 1. *la loi normale et ses propriétés*

X et Y sont deux variables gaussiennes indépendantes. X a pour moyenne 5 et pour variance 4. Y a pour moyenne 6 et pour variance 9.

1. Avec quelle probabilité X est-elle (i) inférieure à 6 ? (ii) supérieure à 3 ? (iii) comprise entre 3 et 7 ?

   (i) $\mathbb{P}(X < 6) = \mathbb{P}(\mathcal{N}(5;2) < 6) \overset{\text{R6}}{=} \phi((6-5)/2) = \phi(1/2) \approx 0,69$ (R : `pnorm(1/2)`)

   (ii) $\mathbb{P}(X > 3) = \mathbb{P}(\mathcal{N}(5;2) > 3) \overset{\text{R6}}{=} 1 - \phi((3-5)/2) = 1 - \phi(-1) \overset{\text{R5}}{=} \phi(1) \approx 0,84$ (R : `pnorm(1)`)

   (iii) $\mathbb{P}(3 \leq X \leq 7) = \mathbb{P}(3 \leq \mathcal{N}(5;2) \leq 7) \overset{\text{R6}}{=} \phi((7-5)/2) - \phi((3-5)/2) = \phi(1) - \phi(-1) \overset{\text{R5}}{=} 2 \times \phi(1) - 1 \approx 0,68.$

2. Avec quelle probabilité la variable $2Y + 3$ est-elle supérieure à 21 ?

   $2Y + 3 \overset{\text{R1}}{\sim} \mathcal{N}(2 \times 6 + 3; 2 \times 3)$; $\mathbb{P}(2Y + 3 > 21) = \mathbb{P}(\mathcal{N}(15;6) > 21) \overset{\text{R6}}{=} 1 - \phi((21-15)/6) = 1 - \phi(1) \approx 0,16.$

3. Avec quelle probabilité la variable $X + Y$ est-elle comprise entre 10 et 12 ?

   X et Y sont indépendantes, ainsi : $X + Y \overset{\text{R3}}{\sim} \mathcal{N}(5+6; \sqrt{4+9})$; $\mathbb{P}(10 \leq X + Y \leq 12) = \mathbb{P}(10 \leq \mathcal{N}(11; \sqrt{13}) \leq 12) \overset{\text{R6}}{=} \phi((12-11)/\sqrt{13}) - \phi((10-11)/\sqrt{13}) \overset{\text{R5}}{=} 2 \times \phi(1/\sqrt{13}) - 1 \approx 0,22.$ (R : `2*pnorm(1/sqrt(13))-1`)

Exercice 2. *propriétés de la loi normale*

Dans un pays où la plupart des actifs sont salariés, le salaire mensuel des hommes, exprimé en milliers d'euros, est distribué selon une loi normale de moyenne 2 de variance 1, celui des femmes est distribué selon une loi normale de moyenne $1,8$ et de variance $0,9$.

1. Avec quelle probabilité un couple gagne-t-il au moins 4300 euros par mois ?

   X : le salaire de l'homme en milliers d'euros, dans un couple choisi au hasard ; Y : le salaire de la femme.

   $X \sim \mathcal{N}(2; \sqrt{1})$, $Y \sim \mathcal{N}(1,8; \sqrt{0,9})$ et en supposant X et Y indépendantes : $X + Y \sim \mathcal{N}(3,8; \sqrt{1+0,9})$.

   $\mathbb{P}(X + Y \geq 4,3) = 1 - \phi((4,3-3,8)/\sqrt{1,9}) \approx 0,36$. 36% des couples environ gagne au moins 4300 euros par mois.

2. Avec quelle probabilité le salaire d'une femme est-il supérieur à celui de son mari ?

   $X - Y = 1 \times X + (-1) \times Y \overset{\text{R3}}{\sim} \mathcal{N}(1 \times 2 + (-1) \times 1,8; \sqrt{1^2 \times 1 + (-1)^2 \times 0,9})$.

   $\mathbb{P}(X < Y) = \mathbb{P}(X - Y < 0) = \mathbb{P}(\mathcal{N}(0,2; \sqrt{1,9}) < 0) \overset{\text{R6}}{=} \phi((0-0,2)/\sqrt{1,9}) \overset{\text{R5}}{=} 1 - \phi(0,2/\sqrt{1,9}) \approx 0,44.$

   Dans 44% des couples, le salaire du mari est inférieur à celui de son épouse.

Supposons que le franc soit remis en circulation et qu'un euro vaille cinq francs.

3. Cette opération change-t-elle la probabilité pour une femme de gagner plus que son mari ?

   $\tilde{X}$: le salaire du mari en milliers de francs. $\tilde{X} = 5X \overset{\text{R1}}{\sim} \mathcal{N}(10;5)$

   $\tilde{Y}$: le salaire de son épouse en milliers de francs. $\tilde{Y} = 5Y \overset{\text{R1}}{\sim} \mathcal{N}(9; 5 \times \sqrt{0,9})$.

   Si X et Y sont indépendantes, $\tilde{X}$ et $\tilde{Y}$ le sont aussi et : $\tilde{X} - \tilde{Y} = 1 \times \tilde{X} + (-1) \times \tilde{Y}$ est gaussienne (R3) de moyenne : $1 \times 10 + (-1) \times 9 = 1$, de variance : $1^2 \times 5^2 + (-1)^2 \times (5\sqrt{0,9})^2 = 47,5$.

   $\mathbb{P}(\tilde{X} < \tilde{Y}) = \mathbb{P}(\tilde{X} - \tilde{Y} < 0) = \mathbb{P}(\mathcal{N}(1; \sqrt{47,5}) < 0) \overset{\text{R6}}{=} \phi((0-1)/\sqrt{47,5}) \approx 0,44$. La conversion des salaires en francs ne change pas la proportion de couples dans lesquels une femme gagne plus que son mari.

   On parvient au même résulat, plus rapidement, par un argument algébrique : $\mathbb{P}(\tilde{X} < \tilde{Y}) = \mathbb{P}(5X < 5Y) = \mathbb{P}(X < Y)$.

4. Avec quelle probabilité le salaire d'un couple est-il supérieur à 25000 francs ?

   $\tilde{X} + \tilde{Y} \overset{\text{R3}}{\sim} \mathcal{N}(19; \sqrt{47,5})$; $\mathbb{P}(\tilde{X} + \tilde{Y} \geq 25) = \mathbb{P}(\mathcal{N}(19; \sqrt{47,5}) \geq 25) \overset{\text{R6}}{=} 1 - \phi((25-19)/\sqrt{47,5}) \approx 0,19.$

   19% des couples environ ont un salaire de 25000 francs au moins.

   NB. On peut aussi passer par : $\mathbb{P}(\tilde{X} + \tilde{Y} \geq 25) = \mathbb{P}(5X + 5Y \geq 25) = \mathbb{P}(X + Y \geq 5) = \mathbb{P}(\mathcal{N}(3,8; \sqrt{1,9}) \geq 5)$.

Exercice 3. *propriétés de la loi normale*

M. Li détient pour 100 \$ du titre FlyCorp. Lorsque les marchés sont stables, les rendements de cet actif sont distribués selon une loi normale centrée de variance 4×10^{-4}.

A. Lourme, Faculté d'économie, gestion & AES, Université de Bordeaux `http://alexandrelourme.free.fr`

1. Avec quelle probabilité M. Li gagne-t-il 1 \$ au moins ?

   X : rendement journalier de FlyCorp. $X \sim \mathcal{N}(0; 2 \times 10^{-2})$.

   Le gain aléatoire de M. Li un jour où il détient 100 \$ de FlyCorp est : $100X$.

   $\mathbb{P}(100X \geq 1) = \mathbb{P}(X \geq 1/100) = \mathbb{P}(\mathcal{N}(0; 2 \times 10^{-2}) \geq 1/100) = 1 - \phi((1/100 - 0)/(2 \times 10^{-2})) = 1 - \phi(1/2) \approx 0,31$.

   Un jour où il détient pour 100 \$ du titre FlyCorp, M. Li a 31% de chances environ de gagner au moins un dollar.

En période de crise, la volatilité des rendements de FlyCorp double.

2. Avec quelle probabilité M. Li gagne-t-il au moins 1 \$ dans une telle période ?

   Y : rendement journalier de FlyCorp en période de crise. La volatilité (écart-type des rendements) double en période de crise : $\sqrt{V(Y)} = 2 \times \sqrt{V(X)}$. Ainsi : $V(Y) = 4 \times V(X)$ et $Y \sim \mathcal{N}(0; \sqrt{16 \times 10^{-4}})$.

   Gain de M. Li, en période de crise, un jour où il détient 100 \$ de FlyCorp : $100Y$.

   $\mathbb{P}(100Y \geq 1) = \mathbb{P}(Y \geq 1/100) = \mathbb{P}(\mathcal{N}(0; \sqrt{16 \times 10^{-4}}) \geq 1/100) = 1 - \phi((1/100 - 0)/\sqrt{16 \times 10^{-4}}) = 1 - \phi(1/4) \approx 0,40$.

   Un jour de crise où il détient pour 100 \$ de FlyCorp, M. Li a 40% de chances environ de gagner au moins un dollar.

Lors de la prochaine crise financière, Flycorp augmentera artificiellement son rendement moyen de $0,01$.

3. Avec quelle probabilité M. Li gagnera-t-il 1 \$ au moins ?

   Z : rendement journalier de FlyCorp lors de la prochaine crise. $Z \sim \mathcal{N}(0,01; \sqrt{16 \times 10^{-4}})$.

   Gain de M. Li, lors de la prochaine crise, un jour où il détient 100 \$ de FlyCorp : $100Z$

   $\mathbb{P}(100Z \geq 1) = \mathbb{P}(Z \geq 1/100) = \mathbb{P}(\mathcal{N}(0,01; \sqrt{16 \times 10^{-4}}) \geq 1/100) = 1 - \phi((1/100 - 0,01)/\sqrt{16 \times 10^{-4}}) = 1 - \phi(0) = 1/2$. Lors de la prochaine crise, un jour où il détient pour 100 \$ de FlyCorp, M. Li a 50% de chances environ de gagner au moins un dollar.

Exercice 4. *propriétés de la loi normale*

Lorsque M. Chan se rend au travail il prend le métro devant chez lui, descend dix stations plus loin et marche deux kilomètres environ. La durée en minutes du trajet entre deux stations de métro est distribuée selon $\mathcal{N}(2; \sqrt{0,1})$; la durée en minutes du trajet à pied est distribuée selon une loi normale de moyenne 30 et de variance 20.

On affecte le numéro : 0 à la station de métro située devant la maison de M. Chan, le numéro 1 à la station suivante, ..., le numéro 10 à la station où il descend. Pour $i \in \{1, \ldots, 10\}$, X_i désigne la durée du trajet en métro entre les stations $i-1$ et i : $X_i \sim \mathcal{N}(2; \sqrt{0,1})$; Y est la durée du trajet à pied : $Y \sim \mathcal{N}(30; \sqrt{20})$; $Z = Y + \sum_{i=1}^{10} X_i$ est la durée totale du trajet.

1. Avec quelle probabilité le trajet de M. Chan dure-t-il plus d'une heure en supposant :

   (a) que les durées entre deux stations de métro sont indépendantes

   On suppose $Y, X_1, \ldots, X_{10}$ indépendantes. Alors Z est gaussienne d'espérance : $30 + 10 \times 2 = 50$, de variance : $20 + 10 \times 0,1 = 21$. Ainsi : $\mathbb{P}(Z \geq 60) = \mathbb{P}(\mathcal{N}(50; \sqrt{21}) \geq 60) = 1 - \phi((60 - 50)/\sqrt{21}) \approx 0,015$. Il y a 1,5% de chances que le trajet de M. Chan dure une heure au moins.

   (b) que les durées entre deux stations de métro sont égales

   On suppose $X_1 = X_2 = \cdots = X_{10}$. Alors, $\sum_{i=1}^{10} X_i = 10 \times X_1 \overset{\text{R1}}{\sim} \mathcal{N}(10 \times 2; 10 \times \sqrt{0,1})$. En supposant la durée du trajet en métro : $\sum_{i=1}^{10} X_i$ indépendante de Y, Z est gaussienne d'espérance : $30 + 20 = 50$, de variance : $20 + (10\sqrt{0,1})^2 = 30$. Ainsi : $\mathbb{P}(Z \geq 60) = \mathbb{P}(\mathcal{N}(50; \sqrt{30}) \geq 60) = 1 - \phi((60 - 50)/\sqrt{30}) \approx 0,034$. Il y a 3,4% de chances que le trajet de M. Chan dure une heure au moins.

2. Quelle est l'hypothèse la plus vraisemblable parmi (a) et (b) ?

   Il est plus vraisemblable de supposer que les durées de trajet entre stations successives sont indépendantes qu'égales.

M. Chan aimerait arriver à l'heure dans 95% des cas au moins en partant cinquante-cinq minutes à l'avance.

3. Quelle devrait être pour cela la durée moyenne du trajet entre deux stations de métro ?

   On suppose que les variables aléatoires X_i sont gaussiennes d'espérance μ inconnue, de variance $0,1$. Si Y, X_1, ..., X_{10} sont indépendantes, la durée du trajet : $Z = Y + \sum_{i=1}^{10} X_i$ est distribuée selon une loi normale d'espérance : $30 + 10 \times \mu$, de variance : $20 + 10 \times 0,1 = 21$. On cherche μ de sorte que : $\mathbb{P}(Z \leq 55) \geq 0,95$; $\mathbb{P}(\mathcal{N}(30 + 10 \times \mu; \sqrt{21}) \leq 55) \geq 0,95$; $\phi((55 - (30 + 10\mu))/\sqrt{21}) \geq 0,95$; $\phi((25 - 10\mu)/\sqrt{21}) \geq 0,95$; $(25 - 10\mu)/\sqrt{21} \geq \phi^{-1}(0,95)$; $\mu \leq (25 - \sqrt{21} \times \phi^{-1}(0,95))/10 \approx 1,75$. La durée moyenne du trajet entre deux stations successives doit être inférieure ou égale à $1,75$ minutes pour que M. Chan parvienne au travail sans retard dans 95% des cas au moins.

4. Quelle devrait être la variance de la durée du trajet à pied ?

   On suppose que Y est gaussienne d'espérance 30, de variance inconnue : σ^2. Si Y, X_1, ..., X_{10} sont indépendantes, la durée du trajet : $Z = Y + \sum_{i=1}^{10} X_i$ est distribuée selon une loi normale d'espérance : $30 + 10 \times 2 = 50$, de variance $\sigma^2 + 10 \times 0,1 = \sigma^2 + 1$. On cherche σ^2 de sorte que : $\mathbb{P}(Z \leq 55) \geq 0,95$; $\mathbb{P}(\mathscr{N}(50; \sqrt{\sigma^2+1}) \leq 55) \geq 0,95$; $\phi((55-50)/\sqrt{\sigma^2+1}) \geq 0,95$; $\phi(5/\sqrt{\sigma^2+1}) \geq 0,95$; $5/\sqrt{\sigma^2+1} \geq \phi^{-1}(0,95)$; $\sigma^2 \leq (5/\phi^{-1}(0,95))^2 - 1 \approx 8,24$. La variance de la durée du trajet à pied doit être inférieure ou égale à $8,24$ pour que M. Chan parvienne au travail sans retard dans 95% des cas au moins.

Exercice 5. *loi normale et simulation*

$Q_\alpha(\mathscr{N}(\mu;\sigma))$: quantile d'ordre α de la loi $\mathscr{N}(\mu,\sigma)$

1. Générez sous R un échantillon de taille 100 issu de $\mathscr{N}(5, \sqrt{3})$

   ```
   sample = rnorm(100,5,sqrt(3))
   ```

2. Grâce aux données simulées, estimez le quantile d'ordre $0,9$ de $\mathscr{N}(5, \sqrt{3})$

   `quantile(sample,0.9)` # estimation de $Q_{0,9}(\mathscr{N}(5; \sqrt{3}))$ sur un échantillon de taille 100

3. Déterminez le quantile d'ordre $0,9$ de $\mathscr{N}(5, \sqrt{3})$ grâce à la fonction `qnorm` de R

   `qnorm(0.9,5,sqrt(3))` # valeur informatique de $Q_{0,9}(\mathscr{N}(5; \sqrt{3}))$

4. Proposez une formule donnant le quantile d'ordre α $(0 < \alpha < 1)$ de $\mathscr{N}(\mu, \sigma)$

   $\mathbb{P}(\mathscr{N}(\mu,\sigma) \leq q) = \alpha$; $\phi((q-\mu)/\sigma) = \alpha$; $(q-\mu)/\sigma = \phi^{-1}(\alpha)$; $q = \mu + \sigma \times \phi^{-1}(\alpha) = \mu + \sigma \times z_\alpha$.

   Ainsi : $Q_\alpha(\mathscr{N}(\mu,\sigma)) = \mu + \sigma \times z_\alpha$.

   En particulier : $Q_{0,9}(\mathscr{N}(5; \sqrt{3})) = 5 + \sqrt{3} \times z_{0,9}$ (valeur exacte de $Q_{0,9}(\mathscr{N}(5; \sqrt{3}))$).

Exercice 6. *théorème Central Limit et loi de Bernoulli*

70% des étudiants vivent en ville.

n étudiants sont choisis au hasard ; la variable $\mathrm{x_i}$ indique si l'étudiant i vit en ville ($\mathrm{x_i} = 1$) ou à la campagne ($\mathrm{x_i} = 0$). Ainsi, $\mathrm{x_1}, \ldots, \mathrm{x_n}$ est un échantillon aléatoire de la loi de Bernoulli de paramètre $p = 0,7$; les variables $\mathrm{x_1}, \ldots, \mathrm{x_n}$ ont la même espérance : $\mu = 0,7$ et la même variance : $\sigma^2 = 0,7 \times 0,3 = 0,21$.

$\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x_i}/\mathrm{n}$ est la proportion d'étudiants citadins dans l'échantillon ; quand n est grand, $\bar{\mathrm{x}}$ est approximativement distribuée selon $\mathscr{N}(0,7; \sqrt{0,21/n})$, d'après TCL.

1. Avec quelle probabilité un échantillon de cent étudiants contient-il au moins 65% de citadins ?

   $n = 100$ et $\bar{\mathrm{x}} \underset{\text{ap.}}{\sim} \mathscr{N}(0,7; \sqrt{0,21/100})$.

   $\mathbb{P}(\bar{\mathrm{x}} \geq 0,65) = \mathbb{P}(\mathscr{N}(0,7; \sqrt{0,21/100}) \geq 0,65) = 1 - \phi((0,65-0,7)/\sqrt{0,21/100}) = 1 - \phi(-0,5/\sqrt{0,21}) = \phi(0,5/\sqrt{0,21}) \approx 0,86$. Il y a 86% de chances qu'un échantillon de cent étudiants comporte au moins 65% de citadins.

2. Quelle est la taille d'un échantillon contenant au moins 65% de citadins avec neuf chances sur dix ?

   On cherche n de sorte que : $\mathbb{P}(\bar{\mathrm{x}} \geq 0,65) = 0,9$; $\mathbb{P}(\mathscr{N}(0,7; \sqrt{0,21/n}) \geq 0,65) = 0,9$; $1 - \phi((0,65-0,7)/\sqrt{0,21/n}) = 0,9$; $\phi(0,05\sqrt{n/0,21}) = 0,9$; $0,05\sqrt{n/0,21} = \phi^{-1}(0,9)$; $n = 0,21 \times (\phi^{-1}(0,9)/0,05)^2 \approx 138$.

   Il faut interroger 138 étudiants pour que la proportion observée de citadins ait 90% de chances de dépasser $0,65$.

Exercice 7. *théorème Central Limit en l'absence de modèle*

Les français gagnent 2000 € par mois en moyenne et la variance de leurs salaires vaut 1000.

X : le salaire annuel d'un français, exprimé en euros. X est une variable aléatoire d'espérance $\mu = 2000$, de variance $\sigma^2 = 1000$, dont on ignore la loi de probabilité $\mathscr{L}(X)$.

Les salaires $\mathrm{x_1}, \ldots, \mathrm{x_n}$ de n français choisis au hasard constituent un échantillon aléatoire de $\mathscr{L}(X)$. Ainsi, la moyenne des salaires : $\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x_i}/\mathrm{n}$ est approximativement distribuée quand la taille n de l'échantillon est grande, selon $\mathscr{N}(2000; \sqrt{1000/n})$.

1. Avec quelle probabilité le salaire moyen parmi trente français est-il supérieur à 2010 € ?

   $n = 30$ et $\bar{\mathrm{x}} \underset{\text{ap.}}{\sim} \mathscr{N}(2000; \sqrt{1000/30})$. $\mathbb{P}(\bar{\mathrm{x}} > 2010) = \mathbb{P}(\mathscr{N}(2000; \sqrt{1000/30}) > 2010) = 1 - \phi((2010-2000)/\sqrt{1000/30}) = 1 - \phi(\sqrt{3}) \approx 0,04$. Il y a 4% de chances que le salaire moyen parmi trente français soit supérieur à 2010 €.

2. Quelle serait la taille d'un échantillon de français dont le salaire moyen serait compris entre 1990 et 2010 € avec 99% de chances ?

   On cherche n de sorte que $\mathbb{P}(1990 \leq \bar{\mathrm{x}} \leq 2010) = 0,99$; $\mathbb{P}(1990 \leq \mathscr{N}(2000; \sqrt{1000/n}) \leq 2010) = 0,99$; $\phi((2010-2000)/\sqrt{1000/n}) - \phi((1990-2000)/\sqrt{1000/n}) = 0,99$; $2 \times \phi(\sqrt{n/10}) - 1 = 0,99$; $\phi(\sqrt{n/10}) = 0,995$; $n = 10 \times (\phi^{-1}(0,995))^2 \approx 67$.

   Il faut interroger 67 français pour que la moyenne de leurs salaires ait 99% de chances d'être comprise entre 1990 et 2010 €.

Exercice 8. *théorème Central Limit et loi binomiale*

Le nombre mensuel d'absences d'un salarié est distribué selon une loi binomiale de paramètre $n = 20$ et $p = 0,1$.

n salariés sont choisis au hasard ; $\mathrm{x_i}$ compte le nombre d'absences mensuelles du salarié i. Ainsi, $\mathrm{x_1}, \ldots, \mathrm{x_n}$ est un échantillon aléatoire de la loi binomiale $b(20; 0,1)$; $\mathrm{x_1}, \ldots, \mathrm{x_n}$ ont la même espérance : $\mu = 20 \times 0,1 = 2$ et la même variance : $\sigma^2 = 20 \times 0,1 \times 0,9 = 1,8$.

$\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x_i}/\mathrm{n}$ est le nombre moyen d'absences calculé sur n salariés ; quand n est grand, $\bar{\mathrm{x}}$ est approximativement distribuée selon $\mathscr{N}(2; \sqrt{1,8/n})$, d'après TCL

1. Avec quelle probabilité le nombre moyen d'absences observé en un mois parmi cinquante salariés est-il inférieur à $2,1$?

   $n = 50$ et $\bar{\mathrm{x}} \underset{\text{ap.}}{\sim} \mathscr{N}(2; \sqrt{1,8/50})$. $\mathbb{P}(\bar{\mathrm{x}} \leq 2,1) = \mathbb{P}(\mathscr{N}(2; \sqrt{1,8/50}) \leq 2,1) = \phi((2,1-2)/\sqrt{1,8/50}) \approx 0,7$.

   Parmi 50 salariés, le nombre moyen d'absences en un mois vaut au plus $2,1$ avec 70% de chances.

2. Quelle serait la taille d'un échantillon de salariés dans lequel le nombre moyen d'absences mensuelles serait supérieur à $1,9$ avec 95% de chances ?

   On cherche n de sorte que : $\mathbb{P}(\bar{\mathrm{x}} > 1,9) = 0,95$; $\mathbb{P}(\mathscr{N}(2; \sqrt{1,8/n}) > 1,9) = 0,95$; $1 - \phi((1,9-2)/\sqrt{1,8/n}) = 0,95$; $\phi(0,1\sqrt{n/1,8}) = 0,95$; $0,1\sqrt{n/1,8} = \phi^{-1}(0,95)$; $n = 1,8 \times (\phi^{-1}(0,95)/0,1)^2 \approx 487$.

   Il faut interroger 487 salariés pour que le nombre moyen d'absences mensuelles soit supérieur à $1,9$ avec 95% de chances.

Exercice 9. *théorème Central Limit et loi de Poisson*

Le nombre de pannes hebdomadaires du tram bordelais est distribué selon une loi de Poisson de paramètre $\lambda = 10$. Avec quelle probabilité y aura-t-il au plus onze pannes par semaine en moyenne en 2019 ?

n semaines d'observation sont choisies ; la variable $\mathrm{x_i}$ compte le nombre de pannes du tram dans la semaine i. Ainsi, $\mathrm{x_1}, \ldots, \mathrm{x_n}$ est un échantillon aléatoire de la loi de Poisson : $\mathscr{P}(10)$; les variables $\mathrm{x_1}, \ldots, \mathrm{x_n}$ ont la même espérance : $\mu = 10$ et la même variance : $\sigma^2 = 10$.

$\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x_i}/\mathrm{n}$ est le nombre moyen de pannes hebdomadaires, calculé sur n semaines ; quand n est grand, $\bar{\mathrm{x}}$ est approximativement distribuée selon $\mathscr{N}(10; \sqrt{10/n})$, d'après TCL.

Comme l'année 2022 compte cinquante-deux semaines ($n = 52$), cette année là le nombre moyen de pannes par semaine : $\bar{\mathrm{x}}$ est distribué selon $\mathscr{N}(10; \sqrt{10/52})$. $\mathbb{P}(\bar{\mathrm{x}} \leq 11) = \mathbb{P}(\mathscr{N}(10; \sqrt{10/52}) \leq 11) = \phi((11-10)/\sqrt{10/52}) = \phi(\sqrt{5,2}) \approx 0,99$.

Il y a 99% de chances que le tram connaisse au plus onze pannes par semaine en moyenne au cours de l'année 2022.

Exercice 10. *théorème Central Limit et loi exponentielle*

1. Avec quelle probabilité la moyenne d'un 100-échantillon issu d'une loi exponentielle de paramètre $\lambda = 0,01$ est-elle comprise entre 95 et 105 ?

   $\mathrm{x_1}, \ldots, \mathrm{x_n}$ est un n-échantillon de la loi $\mathscr{E}(0,01)$; les variables $\mathrm{x_1}, \ldots, \mathrm{x_n}$ ont la même espérance : $\mu = 100$ et la même variance : $\sigma^2 = 10000$. Quand n est grand, $\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x_i}/\mathrm{n}$ est approximativement distribuée selon $\mathscr{N}(100; \sqrt{10000/n})$.

   Si $n = 100$, alors $\bar{\mathrm{x}} \underset{\text{ap.}}{\sim} \mathscr{N}(100; \sqrt{100})$.

   $\mathbb{P}(95 \leq \bar{\mathrm{x}} \leq 105) = \mathbb{P}(95 \leq \mathscr{N}(100; \sqrt{100}) \leq 105) = \phi((105-100)/\sqrt{100}) - \phi((95-100)/\sqrt{100}) = \phi(0,5) - \phi(-0,5) = 2\phi(0,5) - 1 \approx 2 \times 0,69 - 1 = 0,38$.

   Il y a 38% de chances que la moyenne d'un 100-échantillon d'une loi exponnetielle de paramètre $0,01$ soit comprise entre 95 et 105.

2. Ecrivez sous R un script qui :

   (a) génère $N = 1000$ échantillons de taille $n = 100$ issus chacun d'une loi exponentielle de paramètre $\lambda = 0,01$

   (b) calcule la moyenne de chaque échantillon

   (c) détermine la proportion d'échantillons dont la moyenne est comprise entre 95 et 105.

```
N = 1000 ; barx <- NULL
for (k in 1:N){sample <- rexp(100,0.01) ; barx[k]=mean(sample)}
prop = sum((barx >= 95)&(barx <= 105))/N
```

3. Ecrivez l'énoncé d'un problème auquel 1. et 2. apporteraient une solution.

   $\mathscr{E}(0,01)$ peut modéliser la durée de vie d'une tortue dans une espèce où l'espérance de vie est proche de 100. Ainsi, 1. permettrait d'affirmer : il y a 38% de chances que la durée de vie moyenne de cent tortues choisies au hasard soit comprise entre 95 et 105 ans.

R1 *la loi normale est stable par transformation affine*

$X \sim \mathscr{N}(\mu, \sigma)$ et $\alpha,\ \beta \in \mathbb{R}$; alors $\alpha X + \beta \sim \mathscr{N}(\alpha\mu + \beta, |\alpha| \times \sigma)$.

Exemple. Si $X \sim \mathscr{N}(4, 3)$ alors $-2X + 1 \sim \mathscr{N}(-2 \times 4 + 1 = -7, |-2| \times 3 = 6)$.

Application. Les salaires belges (en k€) sont distribués selon une loi normale de moyenne 4 et de variance 3 ; si on augmente les salaires de 1000 euros après les avoir doublés, ils restent distribués selon une loi normale mais de moyenne est 9 et de variance 12.

R2 *la loi normale est stable par centrage et réduction*

$X \sim \mathscr{N}(\mu, \sigma)$; alors $(X - \mu)/\sigma \sim \mathscr{N}(0, 1)$.

Exemple. Si $X \sim \mathscr{N}(5, 3)$ alors $(X - 5)/3 \sim \mathscr{N}(0, 1)$.

Application. La taille (en mm) d'une coccinelle est distribuée selon une loi normale de moyenne 5 et de variance 9 ; la proportion de coccinelles mesurant au plus 8 mm est égale à la probabilité pour une variable normale centrée réduite d'être inférieure à 1.

R3 *la combinaison linéaire de variables gaussiennes indépendantes est une variable gaussienne*[a]

$X_1 \sim \mathscr{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathscr{N}(\mu_2, \sigma_2)$ sont indépendantes et $\alpha_1,\ \alpha_2 \in \mathbb{R}$; alors $\alpha_1 X_1 + \alpha_2 X_2$ est distribuée selon la loi normale de moyenne $\alpha_1\mu_1 + \alpha_2\mu_2$ et de variance $\alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2$.

Exemple. Si $X_1 \sim \mathscr{N}(7, 3)$ et $X_2 \sim \mathscr{N}(5, 2)$ sont indépendantes alors $2X_1 + 3X_2$ est distribuée selon la loi normale de moyenne : $2 \times 7 + 3 \times 5 = 29$ et de variance : $2^2 \times 9 + 3^2 \times 4 = 72$.

Application. On confectionne des filets contenant deux pommes de même masse et trois poires de même masse. La masse (en g) d'une pomme est distribuée selon $\mathscr{N}(7, 3)$ et celle d'une poire selon $\mathscr{N}(5, 2)$. La masse des filets est alors distribuée selon $\mathscr{N}(29, \sqrt{72})$ en admettant que les masses d'une pomme et d'une poire choisies au hasard sont indépendantes.

Attention. Si $X_1 \sim \mathscr{N}(7, 3)$ et $X_2 \sim \mathscr{N}(5, 2)$ sont indépendantes alors $2X_1 - 3X_2$ est distribuée selon la loi normale de moyenne : $2 \times 7 - 3 \times 5 = -1$ et de variance : $2^2 \times 9 + (-3)^2 \times 4 = 72$.

R4 *la somme de variables gaussiennes indépendantes et identiquement distribuées (iid) est gaussienne*

$X_1, \ldots, X_n \underset{iid}{\sim} \mathscr{N}(\mu, \sigma)$; alors $\sum_{i=1}^n X_i \sim \mathscr{N}(n\mu, \sqrt{n} \times \sigma)$.

Exemple. Si $X_1, \ldots, X_9 \underset{iid}{\sim} \mathscr{N}(5, 2)$ alors $\sum_{i=1}^9 X_i \sim \mathscr{N}(9 \times 5 = 45, 3 \times 2 = 6)$.

Application. On confectionne des barquettes contenant neuf abricots. Or la masse (en g) d'un abricot est distribuée selon $\mathscr{N}(5, 2)$. La masse d'une barquette est alors distribuée selon $\mathscr{N}(45, 6)$ en supposant indépendantes les masses de neuf abricots choisis au hasard.

R5 *symétrie de la fonction de répartition de $\mathscr{N}(0, 1)$.*

ϕ désigne la fonction de répartition de la loi $\mathscr{N}(0, 1)$: $\forall \alpha \in \mathbb{R}, \phi(\alpha) = \mathbb{P}(\mathscr{N}(0, 1) \leq \alpha)$. Alors pour tout α :

i. $\phi(-\alpha) = 1 - \phi(\alpha)$,

ii. $\phi(\alpha) - \phi(-\alpha) = 2\phi(\alpha) - 1$.

Exemple.

i. $\phi(-0,8) = 1 - \phi(0,8) \approx 0,21$

ii. $\phi(0.8) - \phi(-0.8) = 2\phi(0.8) - 1 \approx 0,58$:

Une variable normale centrée réduite est inférieure à $-0,8$ dans 21% des cas d'après i (on dit aussi que $-0,8$ est le quantile d'ordre $0,21$ de la loi $\mathscr{N}(0, 1)$) ; elle est comprise entre $-0,8$ et $0,8$ dans 58% cas d'après ii.

R6 $a < b$ sont deux réels et $X \sim \mathscr{N}(\mu, \sigma)$; alors :

i. $\mathbb{P}(X \leq b) = \mathbb{P}(X < b) = \phi((b - \mu)/\sigma)$

ii. $\mathbb{P}(X > a) = \mathbb{P}(X \geq a) = 1 - \phi((a - \mu)/\sigma)$

iii. $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b) = \phi((b - \mu)/\sigma) - \phi((a - \mu)/\sigma)$

[a] on dit qu'une variable est gaussienne lorsqu'elle est distribuée selon une loi normale

Exemple. Si $X \sim \mathscr{N}(5,2)$ alors $\mathbb{P}(4 \leq X \leq 7) = \phi\left((7-5)/2\right) - \phi\left((4-5)/2\right) = \phi(1) - \phi(-1/2) = \phi(1) - 1 + \phi(1/2) \approx 0,84 - 1 + 0,69 = 0,53$.

Application. Mon coucou annonce l'heure du thé à cinq heure en moyenne, avec une variance de 4 h^2 ; si son déclenchement suit une loi normale, j'ai 53% de chances de boire le thé entre 4 h et 7 h.

R7 *distribution d'une moyenne empirique (théorème Central Limit)*

$X_1, \ldots, X_n$ est un échantillon aléatoire d'une loi de probabilité appelée la loi mère, de moyenne μ et de variance finie σ^2 ; alors la moyenne d'échantillon $\bar{X} = \left(\sum_{i=1}^{n} X_i\right)/n$ est approximativement[b] distribuée selon une loi normale de moyenne μ et de variance σ^2/n.

Remarque 1. Aucune hypothèse n'est faite quant à la nature de la loi mère, c'est à dire la loi dont est issu l'échantillon. Ainsi, R7 est valable que l'échantillon soit issu de n'importe quelle loi discrète ou continue.

Remarque 2. Si la loi mère est $\mathscr{N}(\mu, \sigma)$ alors $\mathscr{N}(\mu, \sigma/\sqrt{n})$ est la loi exacte (et pas seulement une loi approchée) de $\bar{X}$; ce résultat peut être démontré en combinant R1 et R4.

Exemple. Si $X_1, \ldots, X_{36}$ est un échantillon aléatoire d'une loi de Poisson de paramètre 12, chaque variable X_i a pour moyenne $\mu = 12$, pour variance $\sigma^2 = 12$ et $\left(\sum_{i=1}^{36} X_i\right)/36 \underset{\text{approx}}{\sim} \mathscr{N}(12, \sqrt{12}/6)$.

Application. Le nombre d'appels reçus en une heure sur n'importe lequel des trente-six téléphones d'un centre de renseignements suit une loi de Poisson de paramètre 12. Ainsi, le nombre moyen d'appels par téléphone reçus en une heure est approximativement distribué selon $\mathscr{N}(12, 1/\sqrt{3})$.

R8 *distribution d'une fréquence empirique*

p est la proportion d'individus pourvus d'une qualité dans une population ; alors la proportion $\hat{p}$ de sujets ayant de cette qualité dans un échantillon aléatoire de taille n est approximativement[b] distribuée selon une loi normale de moyenne p et de variance $p(1-p)/n$.

Remarque. R8 est un cas particulier de R7 avec une loi de Bernoulli pour loi mère.

Exemple. Si 40% des individus d'une population possèdent une qualité, la proportion $\hat{p}$ de sujets pourvus de cette qualité dans un échantillon de taille 50 est supérieure à $0,5$ avec propbabilité : $\mathbb{P}(\hat{p} > 0,5) = 1 - \mathbb{P}(\hat{p} \leq 0,5) = 1 - \mathbb{P}\left(\mathscr{N}(0,4; \sqrt{0,4 \times 0,6/50} = \sqrt{0,0048}) \leq 0,5\right) = 1 - \phi\left((0,5 - 0,4)/\sqrt{0,0048}\right) \approx 0,07$.

Application. Si 40% de la population est favorable à M. Li, un institut de sondage a 7% de chances de donner M. Li gagnant sur la base de cinquante sondés.

---

[b] on considère généralement l'approximation comme acceptable lorsque $n \geq 30$