TD 6

Exercice 1 :

1. Nous choisissons l'hypothèse H_0 suivante :"les individus occupant des postes de responsabilité lisent en moyenne $8,6$ minutes par jour", et l'hypothèse H_1 suivante: "les individus occupant des postes de responsabilité lisent en moyenne plus de $8,6$ minutes par jour"

2. L'erreur de première espèce consiste à rejeter à tort l'hypothèse nulle H_0. La conséquence de cette erreur est d'affirmer que les individus occupant des postes de responsabilité lisent en moyenne plus de $8,6$ minutes par jour, alors qu'en fait ils lisent en moyenne $8,6$ minutes par jour.

3. L'erreur de seconde espèce consiste à accepter à tort l'hypothèse nulle H_0. La conséquence de cette erreur est d'affirmer que les individus occupant des postes de responsabilité lisent en moyenne $8,6$ minutes par jour, alors qu'en fait ils lisent en moyenne plus de $8,6$ minutes par jour.

Exercice 2 :

1. Notons X_i, $i = 1, 2, ..., 36$ les salaires des professeurs en école de commerce de l'échantillon considéré et $\bar{X}_{36}$ la moyenne empirique de leurs salaires. On suppose que les X_i sont des variables indépendantes identiquement distribuées selon une loi dont on note μ l'espérance et σ^2 la variance. On considère que la taille de l'échantillon (36 individus) est suffisamment grande pour approcher la loi de $\frac{\bar{X}_{36}-\mu}{\frac{\sigma}{\sqrt{36}}} = \frac{\bar{X}_{36}-\mu}{\frac{\sigma}{6}}$ par $N(0,1)$ (rappelons que cette approximation est fondée sur le théorème central limite). On ne connaît pas σ mais on dispose d'un estimateur convergent de ce paramètre qu'on note $\hat{\sigma}_{36}$ et qui, pour l'échantillon considéré, prend la valeur 5000. On estime que $n = 36$ est suffisamment grand pour approcher la loi de $\frac{\bar{X}_{36}-\mu}{\frac{\hat{\sigma}_{36}}{6}}$ par celle de $\frac{\bar{X}_{36}-\mu}{\frac{\sigma}{6}}$. On en conclut qu'on a approximativement :

$$\frac{\bar{X}_{36}-\mu}{\frac{\hat{\sigma}_{36}}{6}} \rightsquigarrow N(0,1)$$

ce qui permet d'affirmer que :

$$P\left(-1,96 \leq \frac{\bar{X}_{36}-\mu}{\frac{\hat{\sigma}_{36}}{6}} \leq 1,96\right) \simeq 0,95$$

or il est aisé de vérifier que

$$-1,96 \leq \frac{\bar{X}_{36}-\mu}{\frac{\hat{\sigma}_{36}}{6}} \leq 1,96 \Longleftrightarrow \bar{X}_{36} - 1,96\frac{\hat{\sigma}_{36}}{6} \leq \mu \leq \bar{X}_{36} + 1,96\frac{\hat{\sigma}_{36}}{6} \qquad (1)$$

donc :

$$P\left(\bar{X}_{36} - 1,96\frac{\hat{\sigma}_{36}}{6} \leq \mu \leq \bar{X}_{36} + 1,96\frac{\hat{\sigma}_{36}}{6}\right) \simeq 0.95$$

Pour les valeurs observées $\bar{X}_{36} = 72800$ et $\hat{\sigma}_{36} = 5000$, on obtient (approximativement) l'intervalle de confiance suivant : $[71166, 74333]$.

2. Montrons que le fait que $61650 \notin [71166, 74333]$ permet de rejeter l'hypothèse nulle $H_0 : \mu = 61650$ au seuil $\alpha = 0.05$. En fait, nous pouvons montrer (plus généralement) que pour toute valeur $\mu_0 \notin [71166, 74333]$ l'hypothèse $H_0 : \mu = \mu_0$ est rejetée au seuil $\alpha = 0.05$.

Sous l'hypothèse $H_0 : \mu = \mu_0$, la statistique $\frac{\bar{X}_{36} - \mu_0}{\frac{\hat{\sigma}_{36}}{6}}$ suit (approximativement) la loi $N(0, 1)$. On a donc, sous H_0,

$$P\left(-1,96 \leq \frac{\bar{X}_{36} - \mu_0}{\frac{\hat{\sigma}_{36}}{6}} \leq 1,96\right) \simeq 0,95$$

La procédure suivante permet donc d'obtenir un test de H_0 contre H_1 au seuil $\alpha = 0.05$: on accepte H_0 si et seulement si

$$-1,96 \leq \frac{\bar{X}_{36} - \mu_0}{\frac{\hat{\sigma}_{36}}{6}} \leq 1,96$$

ce qui est équivalent, d'après (1) à

$$\bar{X}_{36} - 1,96\frac{\hat{\sigma}_{36}}{6} \leq \mu_0 \leq \bar{X}_{36} + 1,96\frac{\hat{\sigma}_{36}}{6} \tag{2}$$

Or on sait d'après la question précédente que pour les valeurs observées $\bar{X}_{36} = 72800$ et $\hat{\sigma}_{36} = 5000$,on a $\bar{X}_{36} - 1,96\frac{\hat{\sigma}_{36}}{6} = 71166$ et $\bar{X}_{36} + 1,96\frac{\hat{\sigma}_{36}}{6} = 74333$, donc *avec les valeurs observées de* $\bar{X}_{36}$ *et* $\hat{\sigma}_{36}$, l'hypothèse $H_0 : \mu = \mu_0$ est acceptée si et seulement si

$$\mu_0 \in [71166, 74333]$$

Ce n'est clairement pas le cas de $\mu_0 = 61650$. L'hypothèse $H_0 : \mu = 61650$ est donc rejetée.

Exercice 3 :

Notons p la proportion de personnes dans la population totale ayant l'intention de voter pour le candidat en question.

1. L'hypothèse nulle H_0 est celle qu'on privilégie au sens où on veut que le risque de la rejeter à tort soit très faible (par exemple 5%). D'après l'énoncé, on veut que le risque d'accepter l'hypothèse $p > 0.3$ alors qu'elle est fausse soit de 5%. Autrement dit, on veut que le risque de rejeter l'hypothèse $p \leq 0.3$ alors qu'elle est vraie soit de 5%. On choisit donc comme hypothèse nulle l'hypothèse $H_0 : p \leq 0.3$

2. L'hypothèse alternative est H_1 est : $p > 0.3$

3. Le risque de première espèce est le risque de rejeter à tort l'hypothèse H_0, c'est-à-dire d'affirmer que plus de 30% de la population a l'intention de voter pour le candidat alors qu'en réalité la proportion de personnes ayant l'intention de voter pour ce candidat est inférieure ou égale à 30%.

4. Notons $\hat{p}_n$ la proportion de personnes qui ont l'intention de voter pour un candidat dans un échantillon de taille n tiré aléatoirement dans la population. On sait que $n\hat{p}_n$ suit la loi $B(n,p)$. L'approximation normale de la loi binômiale par la loi normale (qu'on justifie dans ce cas par le fait que pour la valeur observée de $\hat{p}_n$, on a : $n\hat{p}_n\left(1-\hat{p}_n\right) = 500\times 0.32\times 0.68 = 108,8 >> 15$) permet d'affirmer qu'on a approximativement $n\hat{p}_n \rightsquigarrow N(np, np(1-p))$ d'où $\hat{p}_n \rightsquigarrow N(p, \frac{p(1-p)}{n})$ ce qui implique que

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

Le fait que la taille de l'échantillon ($n = 500$) soit grande permet de faire une autre approximation : celle qui consiste à considérer que la loi de $\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ et celle de $\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$ sont très proches, de sorte qu'on a approximativement :

$$\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \rightsquigarrow N(0,1)$$

Rappelons que si on note F la fonction de répartition de la loi normale alors $F(1.645) = 0.95$. On peut donc adopter le test suivant dont le risque de première espèce est limité à 5% : on accepte l'hypothèse $H_0 : p \leq 0.3$ si

$$\frac{\hat{p}_n - 0.3}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \leq 1.645$$

et on la rejette sinon. Autrement dit, la région critique associée à ce test est :

$$W_C = \left\{ (x_1, x_2, ..., x_n) \ / \ \frac{\hat{p}_n - 0.3}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} > 1.645 \right\}$$

Puisque la valeur observée de $\hat{p}_n$ est de 0.32 et $n = 500$ alors la statistique $\frac{\hat{p}_n - 0.3}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$ prend la valeur 0.959 qui est inférieure à 1.645. On accepte donc l'hypothèse $H_0 : p \leq 0.3$.

Conclusion : le fait qu'on ait obtenu 32% d'intentions de vote dans l'échantillon interrogé n'est donc pas significatif, au sens où ceci ne permet pas de rejeter l'hypothèse que le candidat obtiendra au plus 30% des voix.

Exercice 4 :

Notons p la proportion de personnes dans la population totale qui réussiraient le test en question si on le leur faisait passer dans des conditions de stress. On désire tester l'hypothèse nulle $H_0 : p = 0.25$ (le stress ne diminue pas les performances) contre l'hypothèse alternative $H_1 : p < 0.25$ (le stress diminue les prformances). Avec des notations et un raisonnement analogues à ceux de l'exercice précédent on a (approximativement) :

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

Sous l'hypothèse $H_0 : p = 0.25$ on a donc :

$$\frac{\hat{p}_n - 0.25}{\sqrt{\frac{0.25\times(1-0,25)}{n}}} \rightsquigarrow N(0,1)$$

et par conséquent, pour $n = 200$,

$$P_{H_0}\left(\frac{\hat{p}_{200} - 0.25}{\sqrt{\frac{0,25\times 0.75}{200}}} \geq -1.645\right) = 0.95$$

ce qui peut se réécrire sous la forme

$$P_{H_0}\left(\hat{p}_{200} \geq 0.25 - 1.645\sqrt{\frac{0,25 \times 0.75}{200}}\right) = 0.95$$

soit :

$$P_{H_0}\left(\hat{p}_{200} \geq 0.20\right) = 0.95$$

On peut donc adopter la règle de décision suivante qui est équivalente à un test de H_0 contre H_1 au seuil de 5% : on accepte l'hypothèse H_0 si

$$\hat{p}_{200} \geq 0.20$$

et on la rejette sinon. La région critique de ce test est donc :

$$W_c = \{(x_1, x_2, ..., x_{200}) \ / \ \hat{p}_{200} < 0.20\}$$

La valeur observée de $\hat{p}_{200}$ est $\frac{42}{200} = 0.21$ ce qui conduit à accepter l'hypothèse H_0.

Conclusion : on ne peut donc pas affirmer à une erreur de première espèce de 5% que le stress fait diminuer les performances au test des individus.

Exercice 5 :

1- L'analyse du collègue est faussée par le fait qu'il compare le taux de survie des créations pures : $\frac{960}{2000} = 0.48$ à celui de toutes les entreprises (créations pures et

reprises) : $\frac{1490}{3000} = 0.496$ au lieu de comparer le taux de survie des créations pures à celui des reprises.

2- Le modèle statistique adopté est le suivant : $X_1, X_2, ..., X_{n_1} \overset{i.i.d}{\rightsquigarrow} B(1, p_1)$ et $Y_1, Y_2, ..., Y_{n_2} \overset{i.i.d}{\rightsquigarrow} B(1, p_2)$ où $X_i = 1$ si l'entreprise i issue d'une création pure survit après 5 ans d'exercice et 0 sinon, et $Y_j = 1$ si l'entreprise j issue d'une reprise survit après 5 ans d'exercice et 0 sinon. Nous supposons en outre que $X_1, X_2, ..., X_{n_1}, Y_1, Y_2, ..., Y_{n_2}$ sont indépendantes. La taille de l'échantillon des créations pures est $n_1 = 2000$ et la taille de celui des reprises est $n_2 = 1000$.

Nous voulons tester l'hypothèse nulle $H_0 : p_1 = p_2$ contre l'hypothèse alternative $H_1 : p_1 < p_2$ au seuil de 5%. Notons $F_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et $F_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_j$ les taux de survie empiriques au sein de l'échantillon des créations pures et de l'échantillon des reprises respectivement, et notons $F = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_j\right)$ le taux de survie empirique au sein de l'échantillon global. Le test de comparaison de fréquences est fondé sur la statistique :

$$Z = \frac{F_2 - F_1}{\sqrt{F(1-F)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Sous l'hypothèse $H_0 : p_1 = p_2$ cette statistique suit approximativement la loi $N(0, 1)$ (la démonstration de ce résultat est donnée en annexe) aussitôt que la taille de chacun des deux échantillons est suffisamment grande pour que l'approximation normale de la loi binômiale soit valable (il est aisé de vérifier que c'est le cas ici). La procédure de test que nous adoptons est la suivante : on rejette H_0 si et seulement si $Z > A$ où le nombre A est déterminé par le seuil du test (l'intuition derrière ce raisonnement est qu'on considère que $p_2 > p_1$ si Z est suffisamment grand). Pour un seuil de 5%, on doit avoir $P_{H_0}(Z > A) = 0.05$. Or sous H_0, $Z \rightsquigarrow N(0, 1)$ et par conséquent $P_{H_0}(Z \leq 1.645) = 0.95$ et donc $P_{H_0}(Z > 1.645) = 0.05$ ce qui implique que $A = 1.645$. Le test de H_0 contre H_1 au seuil de 5% décrit précédemment correspond donc à la règle de décision suivante : on accepte H_0 si $Z \leq 1.645$ et on rejette H_0 si $Z > 1.645$. La région critique de ce test est donc

$$W_c = \left\{ (x_1, x_2, ..., x_{2000}, y_1, y_2, ..., y_{1000}) \ / \ \frac{F_2 - F_1}{\sqrt{F(1-F)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > 1.645 \right\}$$

Au vu de nos observations, la valeur prise par la statistique Z est

$$\frac{\frac{530}{1000} - \frac{960}{2000}}{\sqrt{\frac{1490}{3000}(1 - \frac{1490}{3000})\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = 2.582 > 1.645$$

Nous sommes dans la région critique et par conséquent nous rejetons H_0.

Conclusion : au seuil de 5%, nous constatons que la proportion d'entreprises qui survivent après 5 ans d'exercice est plus grande parmi les reprises que parmi les créations pures.

Annexe : Démontrons le résultat $Z \underset{H_0:p_1=p_2}{\rightsquigarrow} N(0,1)$. Il est aisé de vérifier que l'approximation normale de la loi binômiale est raisonnable pour les deux échantillons ($2000 \times 0.48 \times 0.52 > 15$ et $1000 \times 0.53 \times 0.47 > 15$), ce qui permet d'aboutir à l'approximation suivante : $F_1 \rightsquigarrow N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$ et $F_2 \rightsquigarrow N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$ donc, vu que les X_i et les Y_j sont indépendants alors F_1 et F_2 sont indépendants et suivent chacun une loi normale, ce qui implique que $F_2 - F_1$ suit une loi normale et plus précisément que $F_2 - F_1 \rightsquigarrow N\left(p_2 - p_1, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$. Or sous l'hypothèse $H_0 : p_1 = p_2$, on a $p_1 = p_2 = p$ où p est le taux de survie dans la population globale des entreprises (ceci découle immédiatement du fait que p est une moyenne pondérée de p_1 et p_2) donc sous l'hypothèse $H_0 : p_1 = p_2$, on a $F_2 - F_1 \rightsquigarrow N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ c'est-à-dire $\frac{F_2 - F_1}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \rightsquigarrow N(0,1)$. Comme F est un estimateur convergent de p et que la taille de l'échantillon global $n_1 + n_2$ est grande, on peut "remplacer" p par son estimateur F et affirmer qu'on a approximativement :

$$Z = \frac{F_2 - F_1}{\sqrt{F(1-F)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0:p_1=p_2}{\rightsquigarrow} N(0,1)$$

Exercice 6 :

1. Nous adoptons le modèle statistique suivant :

$$X_1, X_2, ..., X_{10} \overset{i.i.d.}{\rightsquigarrow} N(m_1, \sigma_1^2)$$

$$Y_1, Y_2, ..., Y_{10} \overset{i.i.d.}{\rightsquigarrow} N(m_2, \sigma_2^2)$$

$$X_1, X_2, ..., X_{10}, Y_1, Y_2, ..., Y_{10} \text{ sont indépendantes}$$

Nous souhaitons tester l'hypothèse $m_1 = m_2$ contre l'hypothèse $m_1 \neq m_2$. Vu que les échantillons sont de petite taille, ceci n'est possible (avec les outils dont nous disposons) que si $\sigma_1^2 = \sigma_2^2$. Nous devons donc commencer par tester l'hypothèse $\sigma_1^2 = \sigma_2^2$ avant de tester l'hypothèse $m_1 = m_2$.

2. Dans un premier temps, nous testons l'hypothèse $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$ au seuil de 10%. Le test que nous adoptons est fondé sur la statistique $\frac{S_1^2}{S_2^2}$ où $S_1^2 = \frac{1}{9}\sum_{i=1}^{10}\left(X_i - \bar{X}_{10}\right)^2$ et $S_2^2 = \frac{1}{9}\sum_{i=1}^{10}\left(Y_i - \bar{Y}_{10}\right)^2$ sont respectivement les variances

empiriques modifiées du premier et du second échantillon. Sous l'hypothèse H_0, cette statistique suit la loi de Fisher(9,9) :

$$\frac{S_1^2}{S_2^2} \underset{H_0}{\rightsquigarrow} Fisher(9,9)$$

La zone d'acceptation du test que nous choisissons est de la forme $\left\{A \leq \frac{S_1^2}{S_2^2} \leq B\right\}$. Vu que le seuil du test est de 10%, on doit donc avoir $P_{H_0}(A \leq \frac{S_1^2}{S_2^2} \leq B) = 0.9$. Autrement dit, si on note F une variable aléatoire suivant la loi de Fisher(9,9), il faut qu'on ait $P(A \leq F \leq B) = 0.9$. Il suffit par exemple de choisir A et B tels que $P(A < F) = 0.05$ et $P(F < B) = 0.05$. La valeur de B se lit directement sur la table statistique de la loi de Fisher : $B = 3.18$. Quant à la valeur de A, on la trouve en remarquant que si F suit la loi de Fisher(9,9) alors $\frac{1}{F}$ suit également la loi de Fisher(9,9) donc $P(A < F) = P(A < \frac{1}{F})$ d'où $0.05 = P(A < \frac{1}{F}) = P(F < \frac{1}{A})$ et par conséquent $\frac{1}{A} = B = 3.18$ ce qui implique que $A = 0.31$.

La règle de décision définissant notre test est donc la suivante : on accepte H_0 : $\sigma_1^2 = \sigma_2^2$ si $0.31 \leq \frac{S_1^2}{S_2^2} \leq 3.18$ et on rejette H_0 sinon. Autrement dit, la région critique associée à ce test est :

$$W_c = \left\{ (x_1, x_2, ..., x_{10}, y_1, y_2, ..., y_{10}) \ / \ \frac{S_1^2}{S_2^2} < 0.31 \text{ ou } \frac{S_1^2}{S_2^2} > 3.18 \right\}$$

La valeur de la statistique $\frac{S_1^2}{S_2^2}$ correspondant aux observations est $\frac{0.11176}{0.06299} = 1.7742$. Par conséquent nous ne sommes pas dans la région critique $\Longrightarrow$ nous acceptons donc l'hypothèse H_0.

Conclusion : au seuil de 10%, les observations n'ont pas mis en évidence de différence significative entre les variances des deux échantillons.

Nous pouvons donc travailler maintenant avec le modèle statistique suivant :

$$X_1, X_2, ..., X_{10} \overset{i.i.d.}{\rightsquigarrow} N(m_1, \sigma^2)$$

$$Y_1, Y_2, ..., Y_{10} \overset{i.i.d.}{\rightsquigarrow} N(m_2, \sigma^2)$$

$$X_1, X_2, ..., X_{10}, Y_1, Y_2, ..., Y_{10} \text{ sont indépendantes}$$

et tester l'hypothèse $H_0 : m_1 = m_2$ contre l'hypothèse $H_1 : m_1 \neq m_2$. Pour cela, nous utilisons la statistique de test suivante

$$Z = \frac{\bar{X}_{10} - \bar{Y}_{10}}{\sqrt{S^2(\frac{1}{10} + \frac{1}{10})}}$$

où

$$S^2 = \frac{1}{9+9}\left(\sum_{i=1}^{10}\left(X_i - \bar{X}_{10}\right)^2 + \sum_{i=1}^{10}\left(Y_i - \bar{Y}_{10}\right)^2\right) = \frac{9S_1^2 + 9S_2^2}{18} = \frac{S_1^2 + S_2^2}{2}$$

La statistique Z suit, sous l'hypothèse $H_0 : m_1 = m_2$ la loi de Student à $9 + 9 = 18$ degrés de libertés :

$$Z = \frac{\bar{X}_{10} - \bar{Y}_{10}}{\sqrt{S^2(\frac{1}{10} + \frac{1}{10})}} \underset{H_0}{\rightsquigarrow} T_{18}$$

Pour tester H_0 contre H_1 nous adoptons la règle de décision suivante : nous rejetons H_0 si et seulement si $|Z| > A$ où A est déterminé par le seuil du test. Au seuil de 10% , A doit vérifier $P_{H_0}(|Z| > A) = 0.10$ c'est-à-dire $P_{H_0}(|Z| \leq A) = 0.90$ ou encore $P_{H_0}(Z \leq A) = 0.95$ (à cause du caractère symétrique de la loi de Student). En utilisant la table statistique de la loi de Student, on trouve $A = 1.734$. La région critique du test est donc définie par la condition :

$$\left| \frac{\bar{X}_{10} - \bar{Y}_{10}}{\sqrt{S^2(\frac{1}{10} + \frac{1}{10})}} \right| > 1.734$$

La valeur de la statistique $\frac{\bar{X}_{10} - \bar{Y}_{10}}{\sqrt{S^2(\frac{1}{10} + \frac{1}{10})}}$ correspondant à nos observations est $\frac{3.665 - 3.509}{\sqrt{0.2 \times (0.1176 + 0.06299)}} = 1.1801$.Nous ne sommes pas dans la région critique du test et par conséquent nous acceptons H_0.

Conclusion : au seuil de 10%, les observations faites ne montrent pas de différence significative entre les espérances m_1 et m_2.

Exercice 7 : Intervalles de confiance et tests

L'objet de cet exercice est d'analyser l'échantillon constitué par vos prédécesseurs l'année dernière, en vue de répondre aux deux questions suivantes :

- Quelle est la probabilité que vous trouviez un emploi dans l'année qui suivra la fin de vos études ?
- Quel salaire pouvez-vous espérer obtenir après une licence ? après un master ?

Pour ce faire :

1. Construire des intervalles de confiance à 95% pour le logarithme du salaire moyen et pour la probabilité d'obtenir un emploi dans l'année qui suit la fin des études.

   Rappel du TD précédent : sur les données du groupe 13 "nettoyé", on avait, pour le log du salaire : $\overline{lw} = 10.02$ et $\widehat{\sigma}^2 = (0.326)^2$.
   Pour la probabilité de trouver un emploi en un an au plus, on avait $\widehat{p} = 0.8$.

   Par conséquent, $V(\overline{lw}) = (0.326)^2/35 = (0.326/5.916)^2 = 0.055^2$. Et $V(\widehat{p}) = \widehat{p}(1-\widehat{p})/35 = 0.16/35 = 0.067^2$.

   En supposant que le log du salaire suit une distribution normale et que, de même, on peut approximer la loi de la fréquence empirique par une loi normale, on a
   $I(mlw)_{95\%} = [10.02 - 1.96 \times 0.055; 10.02 + 1.96 \times 0.055] = [9.91; 10.13]$

   $I(p)_{95\%} = [0.8 - 1.96 \times 0.067; 0.8 + 1.96 \times 0.067] = [0.67; 0.93]$

   On voit que le log du salaire est correctement estimé. En prenant l'exponentielle des bornes, on obtient un intervalle [20175, 25029]. La précision relative est moins bonne pour la probabilité estimée de trouver un emploi.

2. Comparez ces intervalles de confiance avec les intervalles de confiance à 90% et à 99%. Que constatez-vous. Ces différences étaient-elles prévisibles ?

   $I(mlw)_{90\%} = [10.02 - 1.64 \times 0.055; 10.02 + 1.64 \times 0.055] = [9.93; 10.11]$

   $I(p)_{90\%} = [0.8 - 1.64 \times 0.067; 0.8 + 1.64 \times 0.067] = [0.69; 0.91]$

   $I(mlw)_{99\%} = [10.02 - 2.57 \times 0.055; 10.02 + 2.57 \times 0.055] = [9.88; 10.16]$

   $I(p)_{99\%} = [0.8 - 2.57 \times 0.067; 0.8 + 2.57 \times 0.067] = [0.63; 0.97]$

   Les bornes sont d'autant plus éloignées que l'on considère un seuil de confiance élevé : c'est logique !

3. Construisez un intervalle de confiance à 95% pour le salaire moyen en supposant que les salaires sont distribués selon une loi normale. Comparez cet intervalle avec l'intervalle obtenu en prenant l'exponentielle des bornes de l'intervalle construit dans la première question. Que constatez-vous ? Cette différence était-elle prévisible ?

   On obtient, sur l'échantillon, $\overline{w} = 23698$ avec un écart-type pour w de 7760. Par conséquent, l'écart-type de $\overline{w}$ est égal à $7760/\sqrt{35} = 1311$.

   $I(mw)_{95\%} = [23698 - 1.96 \times 1311; 23698 + 1.96 \times 1311] = [21127; 26268]$

   L'intervalle obtenu ici est un tout petit peu plus large que celui obtenu en modélisant le log du salaire. Surtout, il est décalé sur la droite, pour des valeurs plus élevées du salaire. Ceci s'explique par le caractère asymétrique de la loi log normale, plus concentrée sur des faibles valeurs du salaire.

4. Testez les hypothèses suivantes :
   - le logarithme du salaire moyen est égal à log(2000 euros) par mois
   - le logarithme du salaire moyen est supérieur ou égal à log(2000 euros) par mois
   - le logarithme du salaire moyen est inférieur ou égal à log(2000 euros) par mois

   Qu'en concluez-vous ?

   N.B. Le seuil n'étant pas précisé, on prend généralement un seuil de 5%.

   La région critique associée au premier test est :
   $\omega = \{(w_1, w_2, ..., w_N) / \left|\frac{l\overline{w} - 10.09}{0.055}\right| \geqslant 1.96\}$
   $= \{(w_1, w_2, ..., w_N) / \left|\frac{10.02 - 10.09}{0.055}\right| \geqslant 1.96\}$
   $= \{(w_1, w_2, ..., w_N) / \left|\frac{-0.070}{0.055}\right| \geqslant 1.96\}$
   $= \{(w_1, w_2, ..., w_N) / |-1.27| \geqslant 1.96\}$

   Par conséquent, on accepte H0.On ne peut pas, statistiquement, considérer comme fausse l'affirmation selon laquelle le (log du) salaire moyen est égal à (log) 24000 Euros.

   La région critique associée au deuxième test est :
   $\omega = \{(w_1, w_2, ..., w_N) / l\overline{w} \leq 10.09 - 1.64 \times 0.055\}$
   $= \{(w_1, w_2, ..., w_N) / 10.02 \leq 10.00\}$

   Par conséquent, on accepte H0 : On ne peut pas, statistiquement, considérer comme fausse l'affirmation selon laquelle le (log du) salaire moyen est supérieur ou égal à (log) 24000 Euros.

   La région critique associée au troisième test est
   $\omega = \{(w_1, w_2, ..., w_N) / l\overline{w} \geqslant 10.09 + 1.64 \times 0.055\}$

$= \{(w_1, w_2, ..., w_N)/10.02 \geqslant 10.18\}$

Par conséquent, on accepte H0 : On ne peut pas, statistiquement, considérer comme fausse l'affirmation selon laquelle le (log du) salaire moyen est inférieur ou égal à (log) 24000 Euros.

!!! Ceci montre que l'imprécision relative d'une estimation peut conduire à accepter des hypothèses contradictoires. Il faut donc toujours se méfier des résultats de tests statistiques réalisés à partir d'estimation qui ne sont pas très précises.

5. Testez l'hypothèse que la probabilité d'obtenir un emploi dans l'année qui suit la fin des études est supérieure à 90% ; qu'elle est inférieure à 50%

   La région critique associée au premier test est :
   $\omega = \{(w_1, w_2, ..., w_N)/\widehat{p} \leq 0.90 - 1.64 \times 0.067\}$
   $= \{(w_1, w_2, ..., w_N)/0.8 \leq 0.79\}$

   On accepte H0 (de justesse!).
   La région critique associée au second test est :
   $\omega = \{(w_1, w_2, ..., w_N)/\widehat{p} \geqslant 0.50 + 1.64 \times 0.067\}$
   $= \{(w_1, w_2, ..., w_N)/0.8 \geqslant 0.61\}$

   On refuse H0..

6. Testez l'hypothèse que la probabilité d'obtenir un emploi dans l'année qui suit la fin des études est identique pour les hommes et les femmes.

   C'est un test de comparaison de fréquences.

   La fréquence estimée pour les hommes est égale à 0.87 (avec Nh=23). La fréquence estimée pour les femmes est égale à 0.67 (avec Nf=12). On a déjà estimé la fréquence sous H0 : en pondérant les estimations pour les hommes et pour les femmes : $\widehat{p} = 0.8$.
   La variance estimée de la différence des fréquences estimées est, sous H0, égale à
   $V(\widehat{p_h} - \widehat{p_f}) = 0.8(1 - 0.8)[\frac{1}{Nh} + \frac{1}{Nf}] = 0.16([\frac{1}{23} + \frac{1}{12}] = 0.16 * 0.127 = 0.02.$

   La région critique associée au test est :
   $\omega = \{(w_1, w_2, ..., w_N)/\left|\frac{(0.87-0.67)-0.00}{0.14}\right| \geqslant 1.96\}$
   $= \{(w_1, w_2, ..., w_N)/\left|\frac{0.20}{0.14}\right| \geqslant 1.96\}$

   $= \{(w_1, w_2, ..., w_N)/|1.43| \geqslant 1.96\}$

   Par conséquent, on accepte H0 : il n'y a pas de différence significative entre hommes et femmes quant à la probabilité de trouver un emploi en moins d'un an.

7. Testez l'hypothèse que le salaire moyen à l'embauche est indépendant de l'Université

d'origine.

Les données disponibles ne permettent pas de faire ce test en pratique pour le groupe 13. Mais le principe est le même que le test ci-dessus : on compare les log(salaires) des 2 groupes. Soit $\overline{lnw}_{p1}$ la moyenne du log du salaire des anciens étudiants de Paris 1 (en nombre $N1$) et $\overline{lnw}_{autres}$ la moyenne du log du salaire des autres anciens étudiants (en nombre $N2$) Soient $\widehat{\sigma}_1$ et $\widehat{\sigma}_2$ les écarts-types des (log) salaires de ces deux groupes. La variance de la différence des log des salaires moyens peut être estimée par :
$V(\ \overline{lnw}_{p1} - \overline{lnw}_{autres}) = \frac{\widehat{\sigma}_1^2}{N1} + \frac{\widehat{\sigma}_2^2}{N2}$.

La région critique associée au test est :

$$\omega = \{(w_1, w_2, ..., w_N)/ \left| \frac{(\overline{lnw}_{p1} - \overline{lnw}_{autres}) - 0.00}{\sqrt{\frac{\widehat{\sigma}_1^2}{N1} + \frac{\widehat{\sigma}_2^2}{N2}}} \right| \geqslant 1.96\}$$