STT-1920 Méthodes statistiques

Solutions des exercices du chapitre 4

Numéro 1. Pour la loi de Fisher avec 23 et 29 degrés de liberté (c'est-à-dire 23 degrés de liberté au numérateur et 29 degrés de liberté au dénominateur), trouvez les quantités suivantes :

(a) La moyenne.

(b) L'écart-type.

(c) Le 95^e centile.

(d) Le 5^e centile.

Solution.

(a)
$$\text{moyenne de la loi } F_{23,29} \;=\; \frac{29}{29-2} \;=\; \frac{29}{27} \;\approx\; 1.074.$$

(b)
$$\text{écart-type de la loi } F_{23,29} \;=\; \sqrt{\frac{2\times 29^2(23+29-2)}{23(29-2)^2(29-4)}} \;=\; \sqrt{0.20063} \;\approx\; 0.448.$$

(c)
$$\text{le } 95^e \text{ centile de la loi } F_{23,29} \;=\; F_{23,29,0.05} \;=\; 1.910.$$

(d)
$$\text{le } 5^e \text{ centile de la loi } F_{23,29} \;=\; F_{23,29,0.95} \;=\; \frac{1}{F_{29,23,0.05}} \;=\; \frac{1}{1.967} \;=\; 0.5084.$$

Numéro 2. Deux machines sont utilisées pour remplir des sacs de carottes de 2 kg. La distribution des poids des sacs remplis par la machine A est la loi $N(2.080, (0.050)^2)$. La distribution des poids des sacs remplis par la machine B est la loi $N(2.050, (0.050)^2)$.

(a) Quel pourcentage des sacs remplis par la machine A pèsent moins de 2 kg ?

(b) Quel pourcentage des sacs remplis par la machine B pèsent moins de 2 kg ?

Je choisis au hasard 24 sacs remplis par la machine A et 30 sacs remplis par la machine B. Je calcule $\overline{x}_{\text{A}}$, s_{A}, $\overline{x}_{\text{B}}$, s_{B}.

(c) Je m'attends à ce que $\overline{x}_{\text{A}} - \overline{x}_{\text{B}}$ soit environ _____, plus ou moins environ _____.

(d) Je m'attends à ce que $s^2_{\text{A}}/s^2_{\text{B}}$ soit environ _____, plus ou moins environ _____.

Petites questions portant sur la matière du chapitre deux. Je pose $N =$ *le nombre de sacs pesant moins de 2 kg parmi les 24 sacs remplis par la machine A.*

(e) Quelle est la distribution de la variable aléatoire N ?

(f) Quelle est l'espérance de la variable aléatoire N ?

(g) Quel est l'écart-type de la variable aléatoire N ?

(h) Comment calcule-t-on $\mathbb{P}[N \geq 4]$?

SOLUTION.

(a)

$$\mathbb{P}[X \leq 2] \;=\; \mathbb{P}[Z \leq (2.00 - 2.08)/0.05] \;=\; \mathbb{P}[Z \leq -1.60] \;=\; 0.0548.$$

(b)

$$\mathbb{P}[Y \leq 2] \;=\; \mathbb{P}[Z \leq (2.00 - 2.05)/0.05] \;=\; \mathbb{P}[Z \leq -1.00] \;=\; 0.1587.$$

(c) On utilise le fait que

$$\overline{X}_{\text{A}} - \overline{X}_{\text{B}} \sim N\left(\mu_{\text{A}} - \mu_{\text{B}}, \sigma^2\left(\frac{1}{n_{\text{A}}} + \frac{1}{n_{\text{B}}}\right)\right).$$

Ici ça donne

$$\overline{X}_{\text{A}} - \overline{X}_{\text{B}} \sim N\left(2.08 - 2.05, (0.05)^2\left(\frac{1}{24} + \frac{1}{30}\right)\right).$$

c'est-à-dire

$$\overline{X}_{\text{A}} - \overline{X}_{\text{B}} \sim N\left(0.03, 0.0001875\right).$$

On s'attend donc à ce que $\overline{X}_{\text{A}} - \overline{X}_{\text{B}}$ soit environ 0.030, plus ou moins environ $\sqrt{0.0001875} \approx 0.014$.

(d) On utilise le fait que

$$S_{\text{A}}^2 / S_{\text{B}}^2 \sim F_{n_{\text{A}}-1, n_{\text{B}}-1}.$$

Ici ça donne

$$S_{\text{A}}^2 / S_{\text{B}}^2 \sim F_{23,29}.$$

On a obtenu la moyenne et l'écart-type de la loi $F_{23,29}$ au numéro 1. On s'attend donc à ce que $S_{\text{A}}^2/S_{\text{B}}^2$ soit environ 1.074, plus ou moins environ 0.448.

(e) On obtient $N \sim \text{binomiale}(n, p)$, avec $n = 24$ et $p = 0.0548$.

(f) $\mathbb{E}[N] = np = 24 \times (0.0548) = 1.3152$.

(g) $\sigma_{\text{N}} = \sqrt{np(1-p)} = \sqrt{24 \times 0.0548 \times 0.9452} = 1.1150$.

(h)

$$\begin{aligned}
\mathbb{P}[N \geq 4] &= 1 - (\mathbb{P}[N = 0] + \mathbb{P}[N = 1] + \mathbb{P}[N = 2] + \mathbb{P}[N = 3]) \\
&= 1 - \left\{ \binom{24}{0}(0.0548)^0(0.9452)^{24} + \binom{24}{1}(0.0548)^1(0.9452)^{23} \right. \\
&\qquad \left. + \binom{24}{2}(0.0548)^2(0.9452)^{22} + \binom{24}{3}(0.0548)^3(0.9452)^{21} \right) \\
&= 1 - (0.2586 + 0.3598 + 0.2399 + 0.1020) = 0.0398.
\end{aligned}$$

NUMÉRO 3. À partir de la population A, on obtient un échantillon aléatoire de taille 16. La moyenne de ces 16 observations est 36.7 et l'écart-type est 20.60. À partir de la population B, on obtient un échantillon aléatoire de taille 21. La moyenne de ces 21 observations est 45.9 et l'écart-type est 8.20. On suppose que la loi $N(\mu_A, \sigma_A^2)$ est un bon modèle pour la population A et que la loi $N(\mu_B, \sigma_B^2)$ est un bon modèle pour la population B. Obtenez un intervalle de confiance de niveau 90% pour le rapport des écart-types théoriques σ_A/σ_B.

SOLUTION. On utilise l'intervalle

$$\left(\frac{1}{\sqrt{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}} \, \frac{s_1}{s_2} \, , \, \frac{1}{\sqrt{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}} \, \frac{s_1}{s_2} \right)$$

Ici on a $s_1 = 20.60$ et $s_2 = 8.20$. À l'aide de la table de la loi de Fisher, on obtient

$$\begin{aligned} F_{n_1-1,n_2-1,\frac{\alpha}{2}} &= F_{15,20,0.05} = 2.203 \\ F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} &= F_{15,20,0.95} = \frac{1}{F_{20,15,0.05}} = \frac{1}{2.328} = 0.4296. \end{aligned}$$

On insère tout ça dans l'intervalle ci-dessus et on obtient l'intervalle $(1.69, 3.83)$.

NUMÉRO 4. Je veux tester $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 > \sigma_2^2$. J'ai des échantillons indépendants de tailles n_1 et n_2. J'obtiens $s_1^2/s_2^2 = 1.887$. Est-ce que le *p-value* est plus petit dans le cas $n_1 = 25$ et $n_2 = 31$ ou dans le cas $n_1 = 38$ et $n_2 = 41$? Suggestion : de quoi a l'air la loi de Fisher avec $n_1 - 1$ et $n_2 - 1$ degrés de liberté ?

SOLUTION. Dans les deux cas, le *p-value* est la surface à droite de 1.887 sous la densité de la loi de Fisher F_{n_1-1,n_2-1}. Plus les degrés de liberté sont grand et plus la loi de Fisher est concentrée autour de la valeur 1. Donc, plus les degrés de liberté sont grand et plus la surface à droite de 1.887 est petite. Le *p-value* est donc plus petit avec $n_1 = 38$ et $n_2 = 41$ qu'avec $n_1 = 25$ et $n_2 = 31$. D'ailleurs, avec le logiciel R j'obtiens

$$\begin{aligned} \text{Surface à droite de 1.887 sous la } F_{37,40} &= 0.0255 \\ \text{Surface à droite de 1.887 sous la } F_{24,30} &= 0.0500 \end{aligned}$$

NUMÉRO 5. On a mesuré les poids de 16 kiwis provenant d'une ferme de Bay of Plenty en Nouvelle-Zélande. Voici ces 16 poids, en grammes :

65.06	71.44	67.93	69.02	67.28	62.34	66.23	64.16
68.56	70.45	64.91	69.90	65.52	66.75	68.54	67.90

On suppose que la loi normale avec moyenne μ_1 et variance σ^2 est un bon modèle pour décrire la distribution des poids des kiwis de Bay of Plenty.

On a mesuré les poids de 18 kiwis provenant d'une ferme de la péninsule de Banks en Nouvelle-Zélande. Voici ces 18 poids, en grammes :

66.00	71.79	65.19	67.25	65.12	61.17
69.72	64.04	67.93	63.95	63.85	68.82
67.54	63.22	61.82	66.81	65.40	69.02

On suppose que la loi normale avec moyenne μ_2 et variance σ^2 est un bon modèle pour décrire la distribution des poids des kiwis de la péninsule de Banks.

On veut tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'alternative $H_1 : \mu_1 > \mu_2$.

(a) Énoncez la règle de décision au seuil 1%.

(b) Calculez votre statistique de test et comparez la à la valeur critique appropriée au seuil 1%. Au seuil 1%, est-ce que vous acceptez ou est-ce que vous rejetez l'hypothèse nulle ?

(c) Quel est votre *p-value* ?

(d) Les hypothèses de normalité et d'égalité des variances théoriques semblent-elles raisonnables ? Justifiez votre réponse.

SOLUTION.

(a) On rejette H_0 si $T \geq t_{n_1+n_2-2,\alpha}$, avec

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{S_c\sqrt{\frac{1}{16} + \frac{1}{18}}}$$

$$t_{n_1+n_2-2,\alpha} = t_{32,0.01} = 2.449$$

(b) On obtient

$$\overline{x}_1 = 67.2494 \qquad s_1 = 2.4553 \qquad \overline{x}_2 = 66.0356 \qquad s_2 = 2.8255$$

L'écart-type combiné est

$$s_c = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{15 \times (2.4553)^2 + 17 \times (2.8255)^2}{32}} = 2.6584.$$

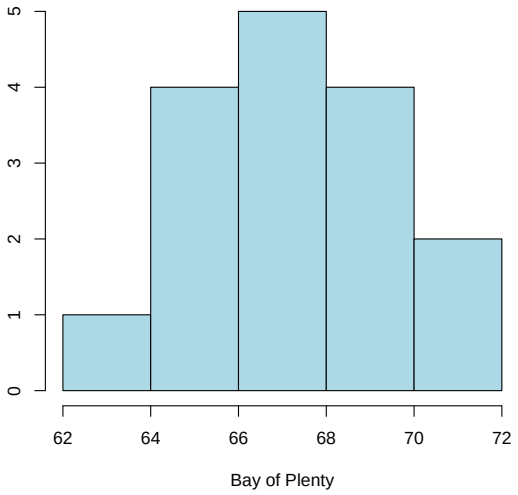
La valeur observée de notre statistique de test est donc

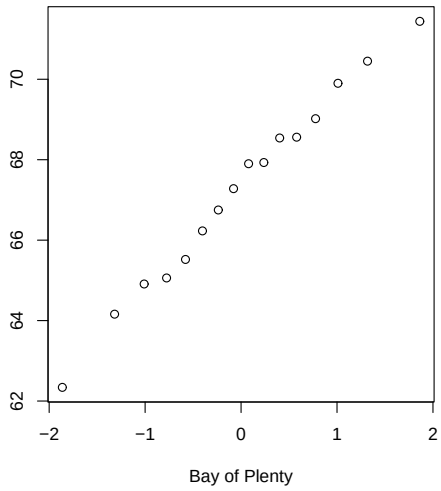
$$T_{obs} = \frac{67.2494 - 66.0356}{2.6584\sqrt{\frac{1}{16} + \frac{1}{18}}} = 1.329.$$

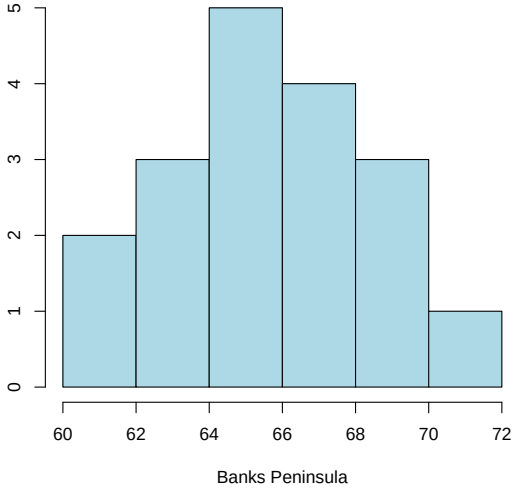
Conclusion : au seuil 1%, il n'y a pas leiu de rejeter H_0.

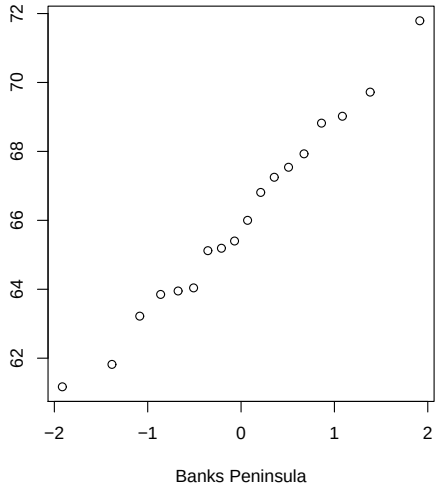
(c) Le *p-value* est la surface à droite de 1.329 sous la densité de la loi de Student avec 32 degrés de liberté. La table me permet de conclure que le *p-value* est légèrement inférieur à 0.10. Le logiciel R me donne un *p-value* de 0.0966.

(d) Voici les histogrammes et les graphes quantile-quantile gaussiens. Il est clair que la loi normale est un modèle raisonnable.

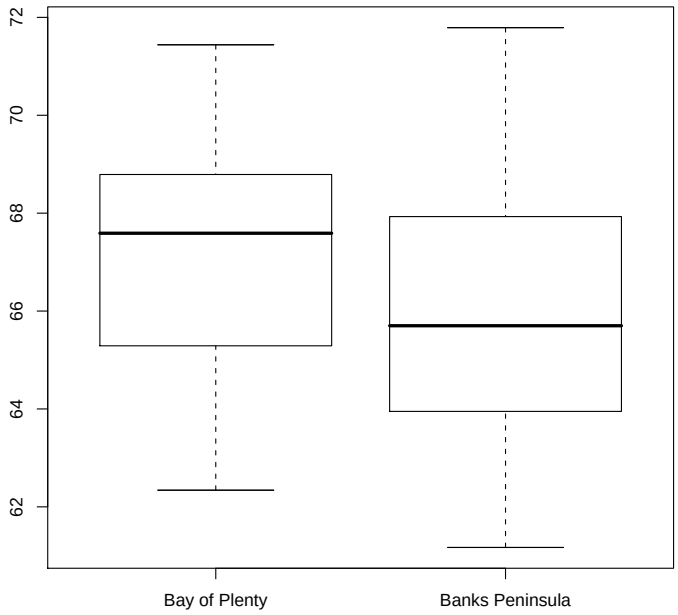








Est-il raisonnable de supposer que les deux populations ont (à peu près) la même variance théorique ? On pourrait examiner les deux histogrammes ci-dessus. Mais pour ça, il aurait fallu que ces deux histogrammes soient dessinés à la même échelle. Voici les boxplots juxtaposés. Il n'y a pas lieu de douter de l'égalité des variances théoriques.



On peut aussi faire un test formel : $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$. Par exemple, au seuil 10%, la règle de décision est

$$\text{on rejette } H_0 \text{ si } S_1^2/S_2^2 \geq F_{15,17,0.05} \text{ ou si } S_1^2/S_2^2 \leq F_{15,17,0.95}.$$

À l'aide de la table, on obtient

$$\begin{aligned} F_{15,17,0.05} &= 2.308 \\ F_{15,17,0.95} &= \frac{1}{F_{17,15,0.05}} = \frac{1}{2.368} = 0.422. \end{aligned}$$

La valeur observée de notre statistique de test est $s_1^2/s_2^2 = (2.4553)^2/(2.8255)^2 = 0.755$. Conclusion : au seuil 10%, on ne rejette pas H_0. Avec le logiciel R, j'obtiens

$$p\text{-}value = 2 \times \mathbb{P}_{H_0}[S_1^2/S_2^2 \leq 0.755] = 2 \times 0.295 = 0.590.$$

Numéro 6.
Voici les poids de 15 fraises provenant du champ A :

48.73	43.44	46.71	51.62	47.24	54.64	47.00	48.40
45.86	47.70	46.14	47.68	44.73	51.69	50.54	

Voici les poids de 15 fraises provenant du champ B :

44.89	34.31	42.74	53.36	41.98	41.64	47.24	37.86
45.89	40.88	40.85	38.60	44.38	44.52	38.26	

(a) Calculez une estimation pour $\mu_{\text{A}} - \mu_{\text{B}}$.

(b) Calculez l'erreur type associée à l'estimation obtenue en (a).

(c) Calculez un intervalle de confiance de niveau 95% pour $\mu_{\text{A}} - \mu_{\text{B}}$.

(d) La méthode utilisée à la partie (c) est valide sous certaines conditions. Énoncez ces conditions.

(e) Vérifiez si les conditions énoncées en (d) sont satisfaites.

(f) Si on avait utilisé 200 fraises du champ A et 200 fraises du champ B (au lieu de 15 et 15), quelle aurait été la longueur de l'intervalle obtenu en (c) ?

SOLUTION. Voici un petit résumé des données :

Champ :	A	B
Taille de l'échantillon :	15	15
Moyenne échantillonnale :	48.142	42.496
Écart-type échantillonnal :	2.935	4.577

L'écart-type échantillonnal combiné est

$$s_c = \sqrt{\frac{(15-1)\times(2.935)^2 + (15-1)\times(4.577)^2}{15+15-2}} = 3.8447.$$

(a) Estimation pour $\mu_{\text{A}} - \mu_{\text{B}}$: $\overline{x}_{\text{A}} - \overline{x}_{\text{B}} = 48.142 - 42.496 = 5.646$.

(b) L'erreur type : $s_c\sqrt{(1/n_1) + (1/n_2)} = 3.8447\sqrt{(1/15) + (1/15)} = 1.404$.

(c) L'intervalle de confiance de niveau 95% pour $\mu_{\text{A}} - \mu_{\text{B}}$:

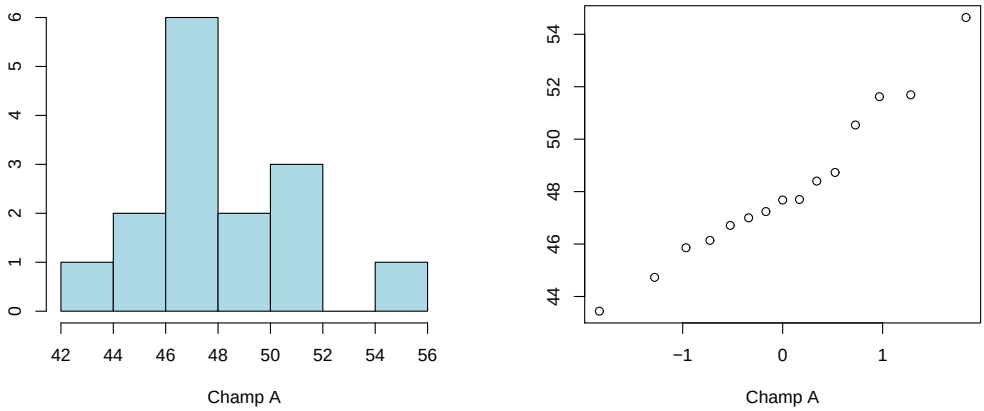
$$(\overline{x}_{\text{A}} - \overline{x}_{\text{B}}) \pm t_{n_1+n_2-2,\frac{\alpha}{2}}\ s_c\sqrt{(1/n_1) + (1/n_2)}$$

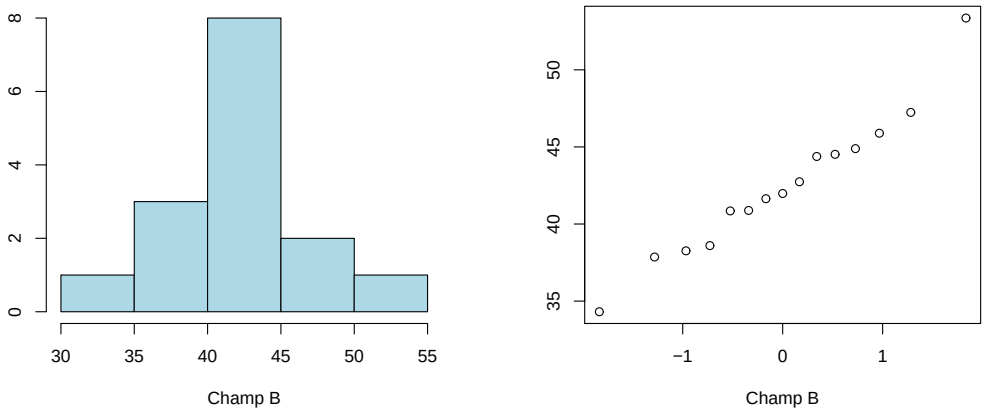
Avec la table j'obtiens $t_{n_1+n_2-2,\frac{\alpha}{2}} = t_{28,0.025} = 2.048$. J'insère tout ça dans la formule ci-dessus et j'obtiens l'intervalle $(2.77, 8.52)$.

(d) Les conditions :

1. Les poids des 15 fraises du champ A peuvent être vues comme étant un échantillon aléatoire de taille 15 issu d'une distribution $N(\mu_{\text{A}}, \sigma^2)$.
2. Les poids des 15 fraises du champ B peuvent être vues comme étant un échantillon aléatoire de taille 15 issu d'une distribution $N(\mu_{\text{B}}, \sigma^2)$.
3. Ces deux échantillons aléatoires sont indépendants l'un de l'autre.
4. Les variances théoriques sont égales.

(e) Vérification des conditions énoncées en (d). Il n'y a aucune raison de douter que nos échantillons soient indépendants l'un de l'autre. Pour les autres conditions, on peut procéder comme au numéro 5. Voici les histogrammes et les graphes quantile-quantile gaussiens. Il est clair que la loi normale est un modèle raisonnable.



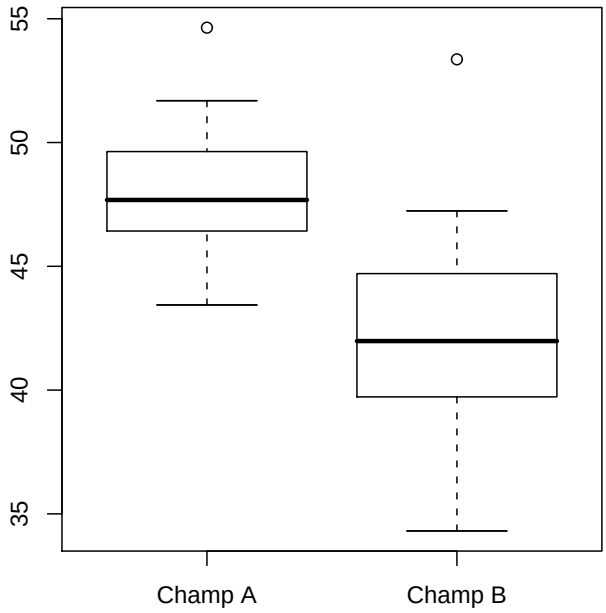


Est-il raisonnable de supposer que les deux populations ont (à peu près) la même variance théorique ? On pourrait examiner les deux histogrammes ci-dessus. Mais pour ça, il aurait fallu que ces deux histogrammes soient dessinés à la même échelle. Le graphe ci-dessous nous montre les boxplots juxtaposés. Il semble y avoir un peu plus de variation dans les poids du champ B. Faisons un test formel : $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$. Au seuil 10%, la règle de décision est

$$\text{on rejette } H_0 \text{ si } S_1^2/S_2^2 \geq F_{14,14,0.05} \text{ ou si } S_1^2/S_2^2 \leq F_{14,14,0.95}.$$

À l'aide de la table, on obtient

$$\begin{aligned} F_{14,14,0.05} &= 2.484 \\ F_{14,14,0.95} &= \frac{1}{F_{14,14,0.05}} = \frac{1}{2.484} = 0.403. \end{aligned}$$



La valeur observée de notre statistique de test est $s_1^2/s_2^2 = (2.935)^2/(4.577)^2 = 0.411$. Conclusion : au seuil 10%, on ne rejette pas H_0. Avec le logiciel R, j'obtiens

$$p\text{-}value = 2 \times \mathbb{P}_{H_0}[S_1^2/S_2^2 \leq 0.411] = 2 \times 0.0539 = 0.1078.$$

(f) Avec 200 fraises du champ A et 200 fraises du champ B, la longueur de l'intervalle obtenu en (c) aurait été (en supposant les mêmes écart-types échantillonnaux) :

$$2 \times t_{398,0.025} \times 3.8447 \times \sqrt{(1/200) + (1/200)} = 1.507.$$

NUMÉRO 7. Au numéro précédent, on suppose qu'on a des distributions normales de même variance, disons σ^2. Calculez un intervalle de confiance de niveau 95% pour cette variance σ^2.

SOLUTION. J'utilise l'intervalle

$$\left(\frac{(n_1+n_2-2)\ S_c^2}{\chi^2_{n_1+n_2-2,\frac{\alpha}{2}}},\ \frac{(n_1+n_2-2)\ S_c^2}{\chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}}} \right).$$

La table de la loi du khi-deux me donne

$$\chi^2_{n_1+n_2-2,\frac{\alpha}{2}} = \chi^2_{28,0.025} = 44.46$$

$$\chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}} = \chi^2_{28,0.975} = 15.31.$$

J'insère dans l'intervalle ci-dessus et j'obtiens l'intervalle $(9.31, 27.03)$.

NUMÉRO 8. J'utilise l'intervalle de confiance

$$\left((\overline{X}_1 - \overline{X}_2) - t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + t_{n_1+n_2-2,\alpha/2}\ S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right).$$

pour une différence de moyennes théoriques, $\mu_1 - \mu_2$. Cette méthode est appropriée lorsque certaines conditions sont satisfaites. Dans chacun des cas suivants, au moins une condition n'est sans doute pas satisfaite. Laquelle ?

(a) μ_1 est le poids moyen des garçons québécois de 6 ans et μ_2 est le poids moyen des garçons québécois à la naissance. J'obtiens un échantillon aléatoire de 20 garçons de 6 ans et un échantillon aléatoire de 20 garçons nouveaux-nés. $X_{1,i}$ est le poids du i^e garçon de 6 ans. $X_{2,i}$ est le poids du i^e garçon nouveau-né.

(b) μ_1 est le salaire moyen des employés de la compagnie Microsoft à Seattle et μ_2 est le salaire moyen des employés de McDonald. J'obtiens un échantillon aléatoire de 20 employés de Microsoft et un échantillon aléatoire de 20 employés de McDonald. $X_{1,i}$ est le salaire du i^e employé de Microsoft. $X_{2,i}$ est le salaire du i^e employé de McDonald.

(c) On veut comparer deux crèmes pour la peau sèche. μ_1 est la réponse moyenne à la crème A. μ_2 est la réponse moyenne à la crème B. On travaille avec 20 patients qui ont la peau sèche. Pour chaque patient, on applique la crème A sur un main et la crème B sur l'autre main. $X_{1,i}$ est la réponse à la crème A pour le i^e patient. $X_{2,i}$ est la réponse à la crème B pour le i^e patient.

SOLUTION.

(a) L'hypothèse d'égalité des variances n'est pas raisonnable. Il y a beaucoup plus de variabilité dans les poids des garçons de 6 ans que dans le poids des garçons nouveaux-nés.

(b) L'hypothèse de normalité n'est pas raisonnable. Habituellement, les distributions de salaire sont asymétriques, étirées vers la droite.

(c) On n'as pas deux échantillons aléatoires indépendants l'un de l'autre. Si on avait utilisé la crème A sur 20 patients et la crème B sur 20 autres patients, alors on aurait eu des échantillons indépendants l'un de l'autre. Mais ici on a des données en paires : on observe $(X_{1,1}, X_{2,1})$ sur le patient numéro 1, on observe $(X_{1,2}, X_{2,2})$ sur le patient numéro 2, on observe $(X_{1,3}, X_{2,3})$ sur le patient numéro 3,... et enfin on observe $(X_{1,20}, X_{2,20})$ sur le patient numéro 20. Au lieu d'avoir deux échantillons indépendants, on a un échantillon bivarié.

NUMÉRO 9. Pour chacune des lois suivantes, trouver la moyenne, l'écart-type, le 5^e centile et le 95^e centile.

(a) La loi $N(0, 1)$.

(b) La loi $N(40, 16)$.

(c) La loi du khi-deux avec 24 degrés de liberté.

(d) La loi de Student avec 24 degrés de liberté.

(e) La loi de Fisher avec 8 et 11 degrés de liberté.

SOLUTION.

(a) La loi $N(0,1)$.
   (i) Moyenne = 0
   (ii) Écart-type = 1
   (iii) 5^e centile = -1.645
   (iv) 95^e centile = 1.645

(b) La loi $N(40,16)$.
   (i) Moyenne = 40
   (ii) Écart-type = 4
   (iii) 5^e centile = 33.42
   (iv) 95^e centile = 46.58

(c) La loi du khi-deux avec 24 degrés de liberté.
   (i) Moyenne = 24
   (ii) Écart-type = 6.93
   (iii) 5^e centile = 13.848
   (iv) 95^e centile = 36.415

(d) La loi de Student avec 24 degrés de liberté.
   (i) Moyenne = 0
   (ii) Écart-type = 1.04
   (iii) 5^e centile = -1.711
   (iv) 95^e centile = 1.711

(e) La loi de Fisher avec 8 et 11 degrés de liberté.
   (i) Moyenne = 1.22
   (ii) Écart-type = 0.952
   (iii) 5^e centile = 0.302
   (iv) 95^e centile = 2.948

NUMÉRO 10. On veut comparer l'efficacité de 2 types de cire (ou *fart*) pour le ski de fond sur de la neige granuleuse, sous une température de -3 C à -2 C. Vingt-huit skieurs ont participé à notre expérience. Nos skieurs étaient tous à peu près du même niveau, tous à peu près du même poids, et ils utilisaient tous le même type de skis. Chaque skieur a skié la même boucle de 20 km de niveau intermédiaire. Pour les 12 skieurs qui ont utilisé la cire A, le temps moyen pour parcourir la boucle a été de 85.50 minutes et l'écart-type a été de 4.10 minutes. On suppose que ces 12 observations constituent un échantillon aléatoire issu d'une loi normale de moyenne μ_A et d'écart-type σ. Pour les 16 skieurs qui ont utilisé la cire B, le temps moyen pour parcourir la boucle a été de 82.25 minutes et l'écart-type a été de 4.80 minutes. On suppose que ces 16 observations constituent un échantillon aléatoire issu d'une loi normale de moyenne μ_B et d'écart-type σ, le même σ pour les 2 types de cire.

(a) Calculez un intervalle de confiance de niveau 90% pour l'écart-type σ.

(b) Calculez un intervalle de confiance de niveau 80% pour la différence $\mu_A - \mu_B$.

SOLUTION.

Pour la cire A : $n_A = 12, \quad \overline{x}_A = 85.50 \text{ min} \quad s_A = 4.10 \text{ min}$

Pour la cire B : $n_B = 16, \quad \overline{x}_B = 82.25 \text{ min} \quad s_B = 4.80 \text{ min}$

L'écart-type combiné :

$$s_c = \sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}} = \sqrt{\frac{(12-1)(4.10)^2 + (16-1)(4.80)^2}{12+16-2}} = 4.5171$$

Intervalle de confiance de niveau 0.90 pour σ :

$$\begin{aligned}
&\left(\sqrt{\frac{(n_A+n_B-2)s_c^2}{\chi^2_{\alpha/2,n_A+n_B-2}}} \, , \, \sqrt{\frac{(n_A+n_B-2)s_c^2}{\chi^2_{1-\alpha/2,n_A+n_B-2}}} \right) \\
&\qquad = \left(\sqrt{\frac{(26)(4.5171)^2}{\chi^2_{0.05,26}}} \, , \, \sqrt{\frac{(26)(4.5171)^2}{\chi^2_{0.95,26}}} \right) \\
&\qquad = \left(\sqrt{\frac{(26)(4.5171)^2}{38.885}} \, , \, \sqrt{\frac{(26)(4.5171)^2}{15.379}} \right) \\
&\qquad = (\sqrt{13.643}, \sqrt{34.496}) = (3.69, 5.87)
\end{aligned}$$

Intervalle de confiance de niveau 0.80 pour $\mu_A - \mu_B$:

$$\begin{aligned}
&(\overline{x}_A - \overline{x}_B) \pm t_{\alpha/2,n_A+n_B-2} \; s_c \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \\
&\qquad = \quad (85.50 - 82.25) \pm t_{0.10,26}(4.5171)\sqrt{\frac{1}{12} + \frac{1}{16}} \\
&\qquad = \quad (85.50 - 82.25) \pm (1.3150)(4.5171)\sqrt{\frac{1}{12} + \frac{1}{16}} \\
&\qquad = \quad 3.25 \pm 2.27 = (0.98, 5.52)
\end{aligned}$$

NUMÉRO 11. Le VO-2 MAX d'un athlète est une mesure de sa capacité aérobique. Pour les épreuves de longue distance dans les sports d'endurance comme la course à pied, le ski de fond, le cyclisme et la natation, le VO-2 MAX permet de prédire la performance de l'athlète. Par exemple, en course à pied, les coureurs ayant un VO-2 MAX de 55.0 ml/kg par minute peuvent s'attendre à courir le 10 000 mètres en 38 minutes et 6 secondes alors que ceux qui ont un VO-2 MAX de 60.0 ml/kg par minute peuvent s'attendre à courir cette même distance en 35 minutes et 22 secondes. Bien que le VO-2 MAX d'un individu soit en grande partie une affaire d'hérédité, il est possible de l'augmenter par l'entraînement. Seize nageuses du Club de Natation Rouge er Or de l'Université Laval ont participé, en début de saison,

à une étude pour comparer la nouvelle méthode d'entraînement *Immersion Totale* (voir www.totalimmersion.net) et la méthode d'entraînement Kinsella, une méthode développée dans les années 70 par John Kinsella, cet américain qui gagna la traversée du Lac St-Jean à 6 reprises consécutives, de 1974 à 1979 (voir www.traversee.qc.ca). Les 16 nageuses ont d'abord été regroupées en 8 paires de nageuses de niveaux comparables. Pour chacune des 8 paires, une nageuse a suivi la méthode d'entraînement Kinsella et l'autre a suivi la méthode *Immersion Totale*. Après 6 mois d'entraînement, on a mesuré, pour chaque nageuse, le gain en VO-2 MAX. Voici les résultats (en ml/kg par min) :

Numéro de la paire de nageuses :	1	2	3	4	5	6	7	8
Gain VO-2 MAX pour la nageuse ayant suivi la méthode *Immersion Totale*	2.17	1.06	1.84	2.44	3.61	2.73	1.94	2.29
Gain VO-2 MAX pour la nageuse ayant suivi la méthode Kinsella	1.35	1.16	0.32	1.81	2.28	1.01	0.80	1.71

En supposant que ces nageuses sont représentatives de l'ensemble des nageuses de niveau universitaire canadien, et en supposant que l'hypothèse de normalité est valide, calculez un intervalle de confiance de niveau 90% pour $\mu_{IT} - \mu_{JK}$. Ici, μ_{IT} représente l'espérance du gain VO-2 MAX pour les nageuses qui suivent le programme d'entraînement *Immersion Totale* et μ_{JK} représente l'espérance du gain de VO-2 MAX pour les nageuses qui suivent le programme d'entraînement de John Kinsella.

Solution :

Attention ! Les données sont appariées !
Nous ne sommes pas en présence du scénario

- $X_{1,1}, X_{1,2}, ..., X_{1,n_1}$, un échantillon aléatoire issu de la première population ;
- $X_{2,1}, X_{2,2}, ..., X_{2,n_2}$, un échantillon aléatoire issu de la deuxième population ;
- les deux échantillons sont indépendants l'un de l'autre.

Nous somme plutôt en présence du scénario

- $(X_{1,1}, X_{2,1}), (X_{1,2}, X_{2,2}), (X_{1,3}, X_{2,3}), ..., (X_{1,n}, X_{2,n})$ est un échantillon aléatoire bivarié.

Il faut donc travailler avec les différences :

$$\begin{aligned}
d_1 &= x_{1,1} - x_{2,1} = 2.17 - 1.35 = 0.82 \\
d_2 &= x_{1,2} - x_{2,2} = 1.06 - 1.16 = -0.10 \\
d_3 &= x_{1,3} - x_{2,3} = 1.84 - 0.32 = 1.52 \\
d_4 &= x_{1,4} - x_{2,4} = 2.44 - 1.81 = 0.63 \\
d_5 &= x_{1,5} - x_{2,5} = 3.61 - 2.28 = 1.33 \\
d_6 &= x_{1,6} - x_{2,6} = 2.73 - 1.01 = 1.72 \\
d_7 &= x_{1,7} - x_{2,7} = 1.94 - 0.80 = 1.14 \\
d_8 &= x_{1,8} - x_{2,8} = 2.29 - 1.71 = 0.58
\end{aligned}$$

La moyenne échantillonnale est $\overline{d} = \overline{x}_D = 0.955$ et l'écart-type échantillonnal est $s_D = 0.5924$. L'intervalle de confiance de niveau 90% pour $\mu_D = \mu_{IT} - \mu_{JK}$ est donc

$$\begin{aligned}
&\left(\overline{d} - t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}\,,\, \overline{d} + t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}\right) \\
&\quad= \left(0.955 - t_{0.05,7}\frac{0.5924}{\sqrt{8}}\,,\, 0.955 + t_{0.05,7}\frac{0.5924}{\sqrt{8}}\right) \\
&\quad= \left(0.955 - (1.8946)\frac{0.5924}{\sqrt{8}}\,,\, 0.955 + (1.8946)\frac{0.5924}{\sqrt{8}}\right) \\
&\quad= (0.955 - 0.397, 0.955 + 0.397) \\
&\quad= (0.558, 1.352)
\end{aligned}$$

Numéro 12. À l'Université de Montréal, deux types d'étudiants prennent le cours IFT-12550 *Introduction au language C++* : les étudiants inscrits au baccalauréat en informatique et les étudiants inscrits au programme de génie informatique. On veut comparer ces 2 groupes. On suppose que la loi normale avec moyenne μ_{BI} et variance σ^2_{BI} est un bon modèle pour les notes à l'examen final de IFT-12550 pour les étudiants inscrits au baccalauréat en informatique et que la loi normale avec moyenne μ_{GI} et variance σ^2_{GI} est un bon modèle pour les notes à l'examen final de IFT-12550 pour les étudiants inscrits au programme de génie informatique. De plus, on suppose que $\sigma^2_{BI} = \sigma^2_{GI}$ et on écrit tout simplement σ^2 pour dénoter cette variance théorique commune.

(a) On veut tester $H_0 : \mu_{GI} = \mu_{BI}$ contre $H_1 : \mu_{GI} \neq \mu_{BI}$. On a obtenu les notes pour un échantillon aléatoire de 12 étudiants inscrits au baccalauréat en informatique. La moyenne de ces 12 notes est 58.4 et l'écart-type est 6.30. On a également obtenu les notes pour un échantillon aléatoire de 18 étudiants inscrits en génie informatique. La moyenne de ces 18 notes est 66.2 et l'écart-type est 5.80. Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Quel est le *p-value* ?

(b) Parmi un échantillon de 80 étudiants inscrits au baccalauréat en informatique, il y avait 36 femmes et 44 hommes. Parmi un échantillon de 100 étudiants

inscrits en génie informatique, il y avait 28 femmes et 72 hommes. Testez $H_0 : p_{BI} = p_{GI}$ contre $H_1 : p_{BI} \neq p_{GI}$. Ici p_{GI} représente la proportion de femmes en génie informatique à l'Université de Montréal et où p_{BI} représente la proportion de femmes au baccalauréat en informatique à l'Université de Montréal. Au seuil 5%, est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Quel est le *p-value* ?

SOLUTION :

(a) On utilise la règle de décision

$$\text{On rejette } H_0 \text{ si } |T| \geq t_{\alpha/2, n_{GI}+n_{BI}-2}$$

avec

$$T = \frac{\overline{X}_{GI} - \overline{X}_{BI}}{S_c \sqrt{\frac{1}{n_{GI}} + \frac{1}{n_{BI}}}}$$

Ici on a

- $n_{BI} = 12 \quad \overline{x}_{BI} = 58.4 \quad s_{BI} = 6.30$
- $n_{GI} = 18 \quad \overline{x}_{GI} = 66.2 \quad s_{GI} = 5.80$
- $\alpha = 0.05$, donc $t_{\alpha/2, n_{GI}+n_{BI}-2} = t_{0.025,28} = 2.0484$

et

$$s_c = \sqrt{\frac{(n_{GI}-1)s_{GI}^2 + (n_{BI}-1)s_{BI}^2}{n_{GI}+n_{BI}-2}} = 6.0014$$

La valeur observée de notre statistique de test est donc

$$T_{obs} = \frac{66.2 - 58.4}{6.0014\sqrt{\frac{1}{18} + \frac{1}{12}}} = 3.487.$$

D'après la table, le *p-value* est entre 0.001 et 0.002. D'après R, le *p-value* est 0.0016. Au seuil, 5% on rejette H_0. Au seuil 1%, on rejette H_0. On rejette H_0 vivement !

(b) On utilise la règle de décision

$$\text{on rejette } H_0 \text{ si } |Z| \geq z_{\alpha/2}$$

avec

$$Z = \frac{\hat{p}_{BI} - \hat{p}_{GI}}{\sqrt{\hat{p}_o(1-\hat{p}_o)\left(\frac{1}{n_{BI}} + \frac{1}{n_{GI}}\right)}}$$

avec

- $n_{BI} = 80 \quad \hat{p}_{BI} = 36/80 = 0.45$
- $n_{GI} = 100 \quad \hat{p}_{GI} = 28/100 = 0.28$
- $\alpha = 0.05$, donc $z_{\alpha/2} = z_{0.025} = 1.96$

et avec

$$\hat{p}_o = \frac{n_{BI}\hat{p}_{BI} + n_{GI}\hat{p}_{GI}}{n_{BI} + n_{GI}} = \frac{36 + 28}{80 + 100} = \frac{64}{180} = 0.3555$$

La valeur observée de notre statistique de test est donc

$$Z_{obs} = \frac{0.45 - 0.28}{\sqrt{0.3555(1 - 0.3555)\left(\frac{1}{80} + \frac{1}{100}\right)}} = 2.37.$$

D'après la table, le *p-value* est 0.0178. Au seuil 5%, on rejette H_0.

NUMÉRO 13. Un échantillon aléatoire de 150 étudiantes de l'Université Laval révèle que 24 d'entre elles fréquentent le PEPS régulièrement. Pour les étudiants, un échantillon aléatoire de taille 196 révèle que 49 d'entre eux fréquentent le PEPS régulièrement. Par *fréquentation régulière* on veut dire au moins 3 visites au PEPS par semaine. Calculez un intervalle de confiance de niveau 90% pour la différence $p_H - p_F$, où p_H et p_F représentent les proportions d'étudiants (H = homme) et d'étudiantes (F = femme) de l'Université Laval qui fréquentent le PEPS régulièrement.

SOLUTION : On utilise l'intervalle de confiance

$$(\hat{p}_H - \hat{p}_F) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_H(1 - \hat{p}_H)}{n_H} + \frac{\hat{p}_F(1 - \hat{p}_F)}{n_F}}$$

avec

- $n_H = 196$ et $\hat{p}_H = 49/196 = 0.25$
- $n_F = 150$ et $\hat{p}_F = 24/150 = 0.16$
- $\alpha = 0.10$

et on obtient l'intervalle $(0.019, 0.161)$.

NUMÉRO 14. Considérons l'intervalle de confiance pour une différence de proportions avec des échantillons de même taille n :

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{n}}\ ,\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{n}}\right]$$

Quelle valeur de n nous assure que la longueur de cet intervalle sera au plus 0.04 ?

SOLUTION : On veut

$$z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{n}} \leq 0.02.$$

On a

$$z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{n}} \leq z_{\alpha/2}\sqrt{\frac{(1/4) + (1/4)}{n}} = \frac{z_{\alpha/2}}{\sqrt{2n}}.$$

Il suffit donc de prendre n tel que

$$\frac{z_{\alpha/2}}{\sqrt{2n}} \leq 0.02$$

c'est-à-dire

$$n \geq \frac{1}{2} \left(\frac{z_{\alpha/2}}{0.02}\right)^2.$$

Par exemple, avec $\alpha = 0.05$, ça donne

$$n \geq \frac{1}{2} \left(\frac{1.96}{0.02}\right)^2 = 4\,802.$$

NUMÉRO 15. On obtient d'abord un échantillon de taille 10 à partir de la population 1. La moyenne échantillonnale est 37.65 et l'écart-type échantillonnal est 12.33. On obtient ensuite un échantillon de taille 18 à partir de la population 2. La moyenne échantillonnale est 26.51 et l'écart-type échantillonnal est 6.28. On suppose que les histogrammes sont en forme de cloche symétrique. Au seuil 5%, testez $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$. Quel est votre *p-value*? Justifiez le choix de votre règle de décision.

SOLUTION. Test de Student ou test de Welch ? S'il est raisonnable de supposer que les variances théoriques sont égales, alors on peut utiliser le test de Student. Si je fais le test pour $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$, j'obtiens un *p-value* de 0.016. C'est pas mal limite ! J'opte donc pour le test de Welch.

La valeur observée de notre statistique de test est

$$T'_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{37.65 - 26.51}{\sqrt{\frac{(12.33)^2}{10} + \frac{(6.28)^2}{18}}} = 2.671.$$

Pour déterminer le nombre de degrés de liberté approprié, on calcule

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{(12.33)^2}{10} + \frac{(6.28)^2}{18}\right)^2}{\frac{1}{9}\left(\frac{(12.33)^2}{10}\right)^2 + \frac{1}{17}\left(\frac{(6.28)^2}{18}\right)^2} = 11.65.$$

On prend donc $k = 12$. Notre *p-value* est donc

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[|T'| \geq 2.671] \\ &= \text{ deux fois la surface à droite de 2.671 sous la densité } t_{12} \\ &= 0.02. \end{aligned}$$

Au seuil 5%, on rejette H_0.

NUMÉRO 16. Avec les données du numéro précédent, obtenez un intervalle de confiance de niveau 90% pour la différence des moyennes, $\mu_1 - \mu_2$. Justifiez votre démarche.

SOLUTION. On utilise

$$(\overline{x}_1 - \overline{x}_2) \pm t_{k,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

avec la valeur $k = 12$ obtenue au numéro 15. J'obtiens l'intervalle $(3.71, 18.57)$.

NUMÉRO 17. Sous quelles conditions le test de la somme des rangs de Wilcoxon-Mann-Whitney est-il approprié ?

SOLUTION.

- On a un échantillon aléatoire de taille n_1 issu de la population 1.
- On a un échantillon aléatoire de taille n_2 issu de la population 2.
- Ces deux échantillons aléatoires sont indépendants l'un de l'autres.
- Les distributions à partir desquelles nos échantillons proviennent (c'est-à-dire la distribution de la variable d'intérêt dans la population 1 et la distribution de la variable d'intérêt dans la population 2) ont la même forme. Autrement dit, ces deux distributions sont, à une translation près, la même distribution.

REMARQUE : Si les distributions à partir desquelles nos échantillons proviennent sont en fait des distributions normales avec la même variance, alors la quatrième condition de la liste ci-dessus est satisfaite et, en principe, le test de la somme des rangs de Wilcoxon-Mann-Whitney est approprié. Toutefois dans ce cas le test de Student est préférable.

NUMÉRO 18. Considérons le test de la somme des rangs de Wilcoxon-Mann-Whitney dans le cas où $n_1 = 2$ et $n_2 = 4$. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Oui, oui, je sais, le cas $n_1 = 2$ et $n_2 = 4$ n'est pas très utile en pratique car avec de telles tailles d'échantillon on ne peut pas conclure grand chose. Mais la vraie vie, c'est pour demain ! Aujourd'hui on essaie de comprendre ce qui se passe !

(a) Déterminez l'ensemble des valeurs possibles de la statistique de Wilcoxon.

(b) Calculez $\mathbb{E}_{H_0}[W]$ et $\mathbb{V}\mathrm{ar}_{H_0}[W]$.

(c) [Optionnel, mais vous devriez lire la solution] En procédant comme à la section 4.7.5 des notes de cours, obtenez la distribution exacte de la statistique W de Wilcoxon et dessinez son graphe.

SOLUTION.

(a) L'ensemble des valeurs possibles du W de Wilcoxon : $\{3, 4, 5, 6, 7, 8, 9, 10, 11\}$

(b) La moyenne et la variance du W de Wilcoxon sous H_0 :

$$\begin{aligned} \mathbb{E}_{H_0}[W] &= \frac{n_1(n_1 + n_2 + 1)}{2} = 7 \\ \mathbb{V}\mathrm{ar}_{H_0}[W] &= \frac{n_1 n_2(n_1 + n_2 + 1)}{12} = \frac{14}{3}. \end{aligned}$$

(c)

Rangs échantillon 1	Probabilité	Valeur de W
$1-2$	$1/15$	3
$1-3$	$1/15$	4
$1-4$	$1/15$	5
$1-5$	$1/15$	6
$1-6$	$1/15$	7
$2-3$	$1/15$	5
$2-4$	$1/15$	6
$2-5$	$1/15$	7
$2-6$	$1/15$	8
$3-4$	$1/15$	7
$3-5$	$1/15$	8
$3-6$	$1/15$	9
$4-5$	$1/15$	9
$4-6$	$1/15$	10
$5-6$	$1/15$	11

Voici donc, sous forme de tableau, la distribution de W lorsque H_0 est vraie :

k	3	4	5	6	7	8	9	10	11
$\mathbb{P}_{H_0}[W=k]$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

NUMÉRO 19. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Nos tailles d'échantillons sont égales : $n_1 = n_2$. Dénotons par n cette taille commune aux deux échantillons. Notre règle de décision est de la forme

on rejette H_0 si W est trop grand.

Supposons que les n observations issues de la population 1 soient toutes plus grandes que chacune des n observations issues de la population 2. Est-ce qu'on rejette H_0 ? Calculez le *p-value* dans le cas $n = 1, 2, 3, ...$ À partir de quelle valeur de n obtient-on un *p-value* plus petit que 0.01 ?

SOLUTION. Dans le présent scénario, le *p-value* est simplement la probabilité que les n observations de l'échantillon 1 recoivent les n plus grands rangs. Le *p-value* est donc la probabilité de tirer les boules avec les numéros $n+1$, $n+2$, $n+3$,..., $2n$ lorsqu'on fait n tirages sans remise à partir d'un panier contenant $2n$ boules numérotées 1 à $2n$. De façon équivalente, c'est la probabilité de tirer les n boules rouges lorsqu'on fait n tirages sans remise à partir d'un panier contenant n boules noires et n boules rouges. On a donc

$$\textit{p-value} \ = \frac{1}{\binom{2n}{n}}.$$

Voici la valeur de ce *p-value* dans le cas $n = 1, 2, 3, 4, 5$ et 6. On voit qu'à partir de $n = 5$ le *p-value* est plus petit que 0.01.

n	$\frac{1}{\binom{2n}{n}}$
1	$\frac{1}{2} = 0.5000$
2	$\frac{1}{6} = 0.1667$
3	$\frac{1}{20} = 0.0500$
4	$\frac{1}{70} = 0.0143$
5	$\frac{1}{252} = 0.0040$
6	$\frac{1}{924} = 0.0011$

NUMÉRO 20. Le test de la somme des rangs de Wilcoxon-Mann-Whitney peut être très utile dans les scénarios où la variable d'intérêt est difficile à quantifier. Supposons qu'on veuille comparer les fleurs du champ A avec les fleurs du champ B. Ici la variable d'intérêt n'est ni le poids ni la hauteur mais bien *la beauté*. Et qui donc a dit qu'il n'y avait rien de poétique en statistique ! ? ! Nous obtenons 20 fleurs au hasard à partir du champ A et 20 fleurs au hasard à partir du champ B. Pour s'assurer qu'il n'y aura pas de biais, on a demandé à un aveugle de cueillir les 40 fleurs. En revenant au laboratoire, l'aveugle perd une fleur du champ A et deux fleurs du champ B. Bref, nos tailles d'échantillon sont $n_1 = 19$ et $n_2 = 18$. Prochaine étape : évaluer la beauté de chaque fleur ! Pas facile. Et pas nécessaire ! Il est difficile de quantifier la beauté, mais il est relativement facile de comparer deux fleurs et de déterminer laquelle des deux est la plus belle. Nous demandons à un comité d'experts de mettre nos 37 fleurs en ordre, de la moins belle à la plus belle. Ensuite, nous attribuons le rang 1 à la moins belle fleur, le rang 2 à la deuxième moins belle fleur, etc. Voici les résultats :

rang	1	2	3	4	5	6	7	8	9	10
champ de provenance	A	A	A	B	A	B	A	A	B	A
rang	11	12	13	14	15	16	17	18	19	20
champ de provenance	A	A	B	A	A	B	A	A	B	A
rang	21	22	23	24	25	26	27	28	29	30
champ de provenance	B	A	B	A	B	B	B	A	B	B
rang	31	32	33	34	35	36	37			
champ de provenance	A	B	B	A	B	B	B			

(a) Exprimez H_0 et H_1 en quelques mots.

(b) Quelle est la valeur observée de la statistique de Wilcoxon ?

(c) Si H_0 était vraie, à quoi devrait-on s'attendre ? Autrement dit, complétez la phrase suivante : *Si H_0 était vraie, je m'attendrais à ce que le W de Wilcoxon soit environ _____, plus ou moins environ _____.* Autrement dit, calculez

$$\mathbb{E}_{H_0}[W] \qquad \text{et} \qquad \sqrt{\mathbb{V}\text{ar}_{H_0}[W]}.$$

(d) Déterminez l'ensemble des valeurs possibles de la statistique W.

(e) D'après les résultats obtenus au points (c) et (d), est-ce que le W_{obs} obtenue en (b) vous semble cohérent ou incohérent avec H_0 ?

(f) À l'aide de l'approximation gaussienne de la distribution de W sous H_0, et en utilisant la correction pour la continuité, obtenez le *p-value* approprié.

SOLUTION.

(a) L'hypothèse nulle H_0 dit que les fleurs du champ A sont ni plus belles, ni moins belles, que les fleurs du champ B. L'hypothèse alternative H_1 dit qu'il y a un champ qui a tendance à produire des fleurs plus belles que l'autre champ (sans spécifier quel champ).

(b) $W_{obs} = 282$.

(c)

$$\begin{aligned} \mathbb{E}_{H_0}[W] &= \frac{n_1(n_1+n_2+1)}{2} = 361 \\ \sqrt{\mathbb{V}\text{ar}_{H_0}[W]} &= \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} = 19\sqrt{3} \approx 32.91. \end{aligned}$$

(d)

$$\begin{aligned} w_{min} &= \frac{n_1(n_1+1)}{2} = 190 \\ w_{max} &= \frac{n_1(n_1+1)}{2} + n_1 n_2 = 532. \end{aligned}$$

L'ensemble des valeurs possibles du W de Wilcoxon est donc l'ensemble de tous les entiers de 190 à 532, inclusivement.

(e) Le W_{obs} est plutôt incohérent avec H_0. En supposant que H_0 est vraie, notre valeur observée de 282 est 2.4 écarts-types en bas de l'espérance !

(f)

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[W \le 282] + \mathbb{P}_{H_0}[W \ge 643] \\ &= 2 \times \mathbb{P}_{H_0}[W \le 282] \\ &\approx 2 \times \mathbb{P}\left[Z \le \frac{282.5 - 361}{19\sqrt{3}}\right] = 2 \times \mathbb{P}[Z \le -2.3854] = 0.0171. \end{aligned}$$

Claude Bélisle
Le 27 octobre 2009