STT-1920
Méthodes statistiques

Chapitre 4

Estimation et tests d'hypothèses : Problèmes à deux échantillons

Claude Bélisle
Département de mathématiques et de statistique
Université Laval
Mars 2014

Chapitre 4

Estimation et tests d'hypothèses : Problèmes à deux échantillons

4.1 Introduction

Récapitulons ce que nous avons fait uu chapitre 3. Nous avons considéré les problèmes à un échantillon. Le scénario était toujours le même :

- On considère une certaine population.
- On s'intéresse à une certaine variable statistique, disons la variable X.
- On s'intéresse à un certain paramètre de la distribution de la variable X, disons le paramètre θ.
- À partir de cette population, on obtient un échantillon aléatoire de taille n, disons $X_1, X_2, ..., X_n$.

Les trois exemples de paramètres θ que nous avons étudiés en détails sont les suivants : une moyenne μ, une variance σ^2, une proportion p. Nous avons étudié les deux principaux problèmes d'inférence statistique : l'estimation d'un paramètre et les tests d'hypothèses sur un paramètre.

Pour les problèmes d'estimation, nous avons étudié les concepts de statistique, d'estimateur, d'estimation, de biais, d'erreur quadratique moyenne et d'erreur type. Nous avons également vu

- comment estimer le paramètre ;
- comment calculer interpréter l'erreur type associée à une estimation du paramètre ;
- comment calculer et interpréter un intervalle de confiance pour le paramètre.

Pour les problèmes de tests d'hypothèses, nous avons étudié les concepts d'hypothèse nulle, d'hypothèse alternative, de règle de décision, d'erreur de première espèce, d'erreur de deuxième espèce, de seuil et de *p-value*. Nous avons également vu

- comment choisir une bonne règle de décision ;
- comment calculer et interpréter un *p-value*.

Nous allons maintenant considérer les problèmes à deux échantillons. Commençons par un exemple simple. On veut comparer deux types d'engrais pour plants de tomates, disons l'engrais A et l'engrais B. La variable qui nous intéresse est la taille (hauteur) du plant de tomates trois semaines après germination. Nous allons faire l'expérience suivante. Nous disposons de 40 plants de tomates qui viennent tout juste de germer. Nous allons utiliser l'engrais A sur 20 plants et l'engrais B sur les 20 autres. Au bout de trois semaines, nous allons mesurer les tailles de nos 40 plants de tomates. Écrivons μ_A pour la moyenne théorique des tailles des plants de tomates soumis à l'engrais A et μ_B pour la moyenne théorique des tailles de ceux soumis à l'engrais B. Nous allons considérer les deux problèmes suivants.

Problème d'estimation. Comment estime-t-on la différence $\mu_A - \mu_B$? C'est facile : il suffit de prendre comme estimation la différence des moyennes échantillonnales, c'est-à-dire $\overline{x}_A - \overline{x}_B$. Mais, alors, quelle est l'erreur type associée à cette estimation et comment calcule-t-on un intervalle de confiance pour $\mu_A - \mu_B$?

Problème de test d'hypothèse. Imaginez qu'on veuille tester $H_0 : \mu_A = \mu_B$ contre l'alternative $H_1 : \mu_A > \mu_B$ (ou peut-être $H_1 : \mu_A < \mu_B$, ou peut-être même $H_1 : \mu_A \neq \mu_B$). Quelle sera notre règle de décision et comment calculera-t-on notre *p-value* ?

Dans les pages qui suivent, nous allons examiner plusieurs scénarios de problèmes à deux échantillons. Le problème qui nous intéresse le plus est le problème de la comparaison de deux moyennes. Mais avant de s'attaquer aux moyennes, examinons brièvement le problème de la comparaison de deux variances.

4.2 Comparaison de deux variances

4.2.1 Introduction

On suppose ici que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma_1^2)$.
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma_2^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les paramètres μ_1, μ_2, σ_1^2 et σ_2^2 sont inconnus.

Notre objectif est de comparer les variances théoriques σ_1^2 et σ_2^2. Plus précisément, nous aimerions tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'une ou l'autre des hypothèses alternatives usuelles (unilatérale à droite $H_1 : \sigma_1^2 > \sigma_2^2$, unilatérale à gauche $H_1 : \sigma_1^2 < \sigma_2^2$ ou bilatérale $H_1 : \sigma_1^2 \neq \sigma_2^2$). Dans certains cas, nous aimerions plutôt estimer le rapport σ_1^2/σ_2^2, calculer l'erreur type associée à notre estimation et calculer un intervalle de confiance pour σ_1^2/σ_2^2. Avant d'aller plus loin, il faut se familiariser avec une loi de probabilité importante. Il s'agit de la *loi de Fisher*, aussi appelée la *loi de Fisher et Snedecor*.

4.2.2 La loi de Fisher

Supposons que

(i) $U \sim \chi^2_k$ (autrement dit, U suit la loi du khi-deux avec k degrés de liberté).

(ii) $V \sim \chi^2_\ell$ (autrement dit, V suit la loi du khi-deux avec ℓ degrés de liberté).

(iii) Les variables aléatoires U et V sont indépendantes.

Alors la distribution de la variable aléatoire $R = \frac{U/k}{V/\ell}$ s'appelle la loi F de Fisher avec k et ℓ degrés de liberté. Cette loi sera dénotée $F_{k,\ell}$. Le paramètre k est appelé le nombre de degrés de liberté du numérateur alors que le paramètre ℓ est appelé le nombre de degrés de liberté du dénominateur La densité de cette loi de probabilité de la loi $F_{k,\ell}$ est donnée par l'équation suivante :

$$f(r) = \begin{cases} \dfrac{\Gamma\left((k+\ell)/2\right)(k/\ell)^{k/2}}{\Gamma\left(k/2\right)\Gamma\left(\ell/2\right)} \dfrac{r^{(k/2)-1}}{\left(1+kr/\ell)\right)^{(k+\ell)/2}} & \text{si } r \geq 0, \\ 0 & \text{si } r < 0. \end{cases}$$

Heureusement, nous n'aurons jamais à utiliser cette formule très complexe. Pour nous, l'important est de connaître les principales propriétés de cette loi, de savoir reconnaître les situations où cette loi s'applique et d'être capable de trouver certains quantiles de cette loi à partir d'une table, d'une calculatrice ou d'un logiciel de statistique comme R.

Principales propriétés de la loi F de Fisher :

(i) La densité de la loi $F_{k,\ell}$ est en forme de cloche asymétrique étirée vers la droite.

(ii) Si $R \sim F_{k,\ell}$ et si $\ell > 2$, alors $\mu_R = \frac{\ell}{\ell-2}$.

(iii) Si $R \sim F_{k,\ell}$ et si $\ell > 4$, alors $\sigma^2_R = \frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}$

La propriété (ii) nous permet de voir que la moyenne de la loi $F_{k,\ell}$ est toujours un peu plus grande que 1 et elle vaut à peu près 1 lorsque ℓ est grand. Examinons maintenant la propriété (iii) dans le cas où k et ℓ sont tous les deux grands. Dans ce cas on obtient

$$\sigma^2_R = \frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)} \approx \frac{2\ell^2(k+\ell)}{k\,\ell^3} = \frac{2(k+\ell)}{k\,\ell} = \frac{2}{\ell} + \frac{2}{k}.$$

En particulier, si k et ℓ sont grands et si $k \approx \ell$, alors on a $\mu_R \approx 1$ et $\sigma_R \approx 2/\sqrt{\ell}$.

Pour la loi de Fisher avec $k = \ell = 100$, les approximations ci-dessus nous donne $\mu_R \approx 1$ et $\sigma_R \approx 2/\sqrt{\ell} = 2/\sqrt{100} = 2/10$ alors que les valeurs exactes sont

$$\mu = \frac{\ell}{\ell-2} = \frac{100}{98} = 1.0204,$$

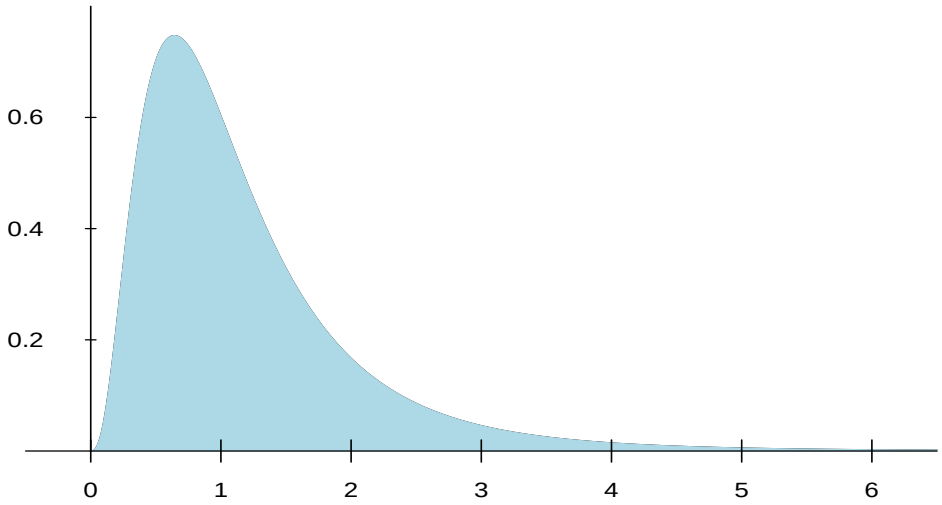
$$\sigma = \sqrt{\frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}} = \sqrt{\frac{2\times 100^2(100+100-2)}{100(100-2)^2(100-4)}} = 0.2072.$$

Exemple 1. Considérons la loi de Fisher avec 8 et 12 degrés de liberté. La moyenne et l'écart-type de cette distribution sont, respectivement,

$$\mu = \frac{\ell}{\ell-2} = \frac{12}{10} = 1.20,$$

$$\sigma = \sqrt{\frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}} = \sqrt{\frac{2\times 12^2(8+12-2)}{8(12-2)^2(12-4)}} = 0.90.$$

Voici le graphe de la densité de la loi de Fisher de l'exemple 1 :



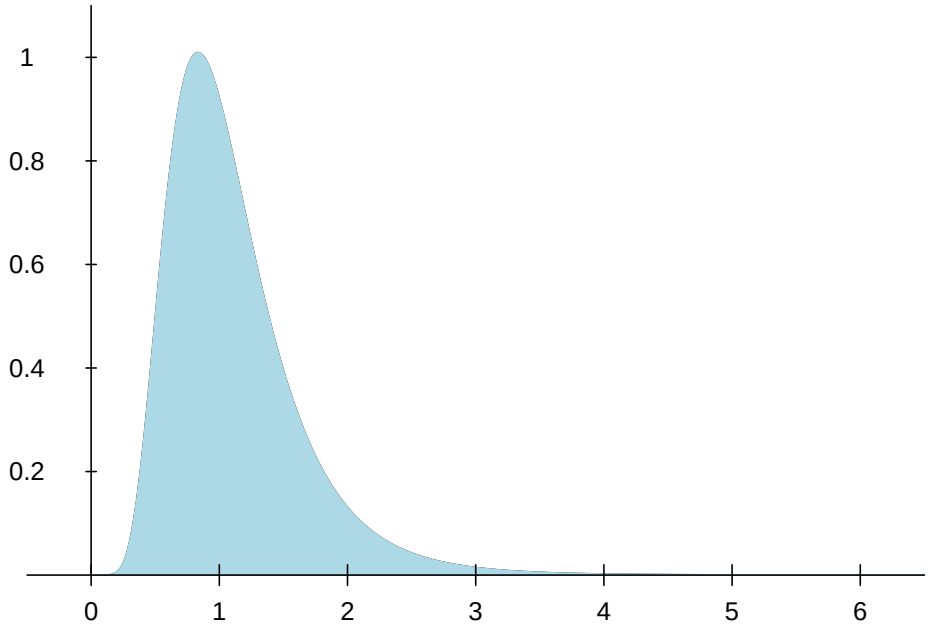
La loi de Fisher avec 8 et 12 degrés de liberté.

Exemple 2. Considérons maintenant la loi de Fisher avec 24 et 20 degrés de liberté. La moyenne et l'écart-type de cette distribution sont, respectivement,

$$\mu = \frac{\ell}{\ell-2} = \frac{20}{18} \approx 1.11,$$

$$\sigma = \sqrt{\frac{2\ell^2(k+\ell-2)}{k(\ell-2)^2(\ell-4)}} = \sqrt{\frac{2\times 20^2(24+20-2)}{24(20-2)^2(20-4)}} \approx 0.52.$$

Voici le graphe de la densité de la loi de Fisher de l'exemple 2 :



La loi de Fisher avec 24 et 20 degrés de liberté.

Principale application de la loi F de Fisher : Nous savons que si les conditions énoncées au début de la présente section sont satisfaites, alors on a

(a) $\frac{(n_1-1)\ S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$,

(b) $\frac{(n_2-1)\ S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$,

(c) Ces deux variables aléatoires sont indépendantes.

Donc, avec $U = \frac{(n_1-1)\ S_1^2}{\sigma_1^2}, V = \frac{(n_2-1)\ S_2^2}{\sigma_2^2}, k = n_1 - 1$ et $\ell = n_2 - 2$, on obtient le résultat suivant :

Théorème : Sous les conditions énoncées au début de la présente section, on a

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

Ce résultat peut aussi s'écrire sous la forme suivante :

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}. \tag{4.1}$$

D'une part, ce théorème met en évidence la réalité suivante : lorsqu'il s'agit de comparer les variances théoriques σ_1^2 et σ_2^2, il est plus simple de considérer le rapport σ_1^2/σ_2^2 plutôt que la différence $\sigma_1^2 - \sigma_2^2$. D'autre part, ce théorème nous permet d'obtenir un intervalle de

confiance pour σ_1^2/σ_2^2 ou d'obtenir une bonne règle de décision pour tester $H_0 : \sigma_1^2 = \sigma_2^2$ (contre l'une ou l'autre des trois hypothèses alternatives usuelles). Mais d'abord, il faut savoir obtenir les quantiles de la loi de Fisher.

4.2.3 Les quantiles de la loi de Fisher

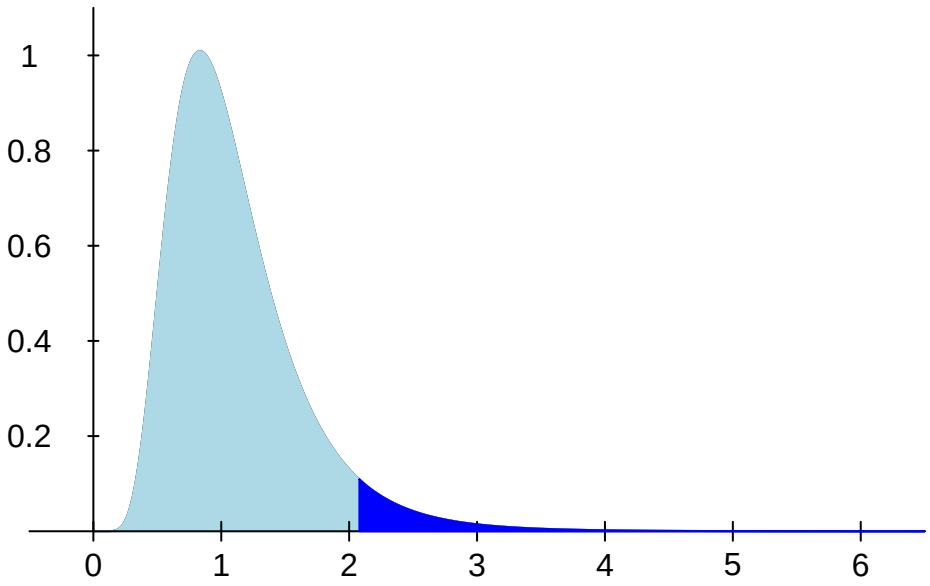
On écrit $F_{k,\ell,\gamma}$ pour dénoter le quantile d'ordre $1-\gamma$ de la loi de Fisher avec k et ℓ degrés de liberté. Donc si $R \sim F_{k,\ell}$, alors on a

$$\mathbb{P}[R \leq F_{k,\ell,\gamma}] = 1 - \gamma$$

ou, de façon équivalente,

$$\mathbb{P}[R > F_{k,\ell,\gamma}] = \gamma.$$

Le graphe suivant est celui de la loi de Fisher avec 24 degrés de liberté au numérateur et 20 degrés de liberté au dénominateur. Sur le graphe, on a indiqué le quantile d'ordre 95%, c'est-à-dire $F_{24,20,0.05} = 2.082$. Ce quantile a été obtenu avec la commande `qf(0.95,24,20)` dans le logiciel R. Sur le graphe, la surface à gauche de 2.082, ombragée pâle, est égale à 0.95. La surface à droite de 2.082, ombragée foncée, est égale à 0.05.



Le quantile d'ordre 95% de la loi $F_{24,20}$.

La table présentée à l'annexe A.4 donne le quantile d'ordre 95% de la loi de Fisher pour différentes combinaisons des deux nombres de degrés de liberté. Notez que cette table permet aussi d'obtenir les quantiles d'ordre 5% grâce à la propriété suivante de la loi de Fisher :

$$F_{k,\ell,\gamma} = \frac{1}{F_{\ell,k,1-\gamma}}. \tag{4.2}$$

Par exemple, pour trouver le quantile d'ordre 5% de la loi $F_{24,20}$, on fait

$$F_{24,20,0.95} = \frac{1}{F_{20,24,0.05}} = \frac{1}{2.027} = 0.4933.$$

Avec l'aide de la table de la loi de Fisher et en utilisant la propriété (4.2) ci-dessus, le lecteur devrait pouvoir vérifier les affirmations suivantes :

1. Si $R \sim F_{10,16}$, alors $\mathbb{P}[R > 2.494] = 0.05$.
2. Si $R \sim F_{10,16}$, alors $\mathbb{P}[R < 0.3536] = 0.05$.
3. Si $R \sim F_{10,16}$, alors $\mathbb{P}[0.3536 < R < 2.494] = 0.90$.

4.2.4 Intervalle de confiance de niveau $1-\alpha$ pour le rapport σ_1^2/σ_2^2

Le résultat (4.1) nous permet d'écrire

$$\mathbb{P}\left[F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{n_1-1,n_2-1,\frac{\alpha}{2}}\right] = 1-\alpha.$$

Cette équation peut être réécrite de la façon suivante :

$$\mathbb{P}\left[\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\frac{S_1^2}{S_2^2}\right] = 1-\alpha.$$

On obtient donc l'intervalle de confiance de niveau $1-\alpha$ pour le rapport σ_1^2/σ_2^2 :

$$\left(\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\ \frac{s_1^2}{s_2^2}\ ,\ \frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\ \frac{s_1^2}{s_2^2}\right). \tag{4.3}$$

Puisque la densité de probabilité de la loi de Fisher est en forme de cloche (asymétrique) centrée aux alentours de 1, on voit que le quantile $F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ est un nombre plus grand que 1 alors que le quantile $F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ est un nombre plus petit que 1. La constante $1/F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ qui apparaît dans l'intervalle ci-dessus est donc un nombre compris entre 0 et 1 alors que la constante $1/F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ est un nombre plus grand que 1. On voit donc que l'intervalle de confiance pour σ_1^2/σ_2^2 contient l'estimation s_1^2/s_2^2 et est de la forme

$$\big(\text{une fraction de l'estimation } s_1^2/s_2^2\ ,\text{un multiple de l'estimation } s_1^2/s_2^2\ \big)$$

Exemple 3. On a obtenu des échantillons aléatoires indépendants à partir de deux populations. Voici un résumé de nos observations :

Population	1	2
Taille de l'échantillon	12	15
Moyenne échantillonnale	23.37	20.55
Écart-type échantillonnal	4.83	2.21

Obtenez un intervalle de confiance de niveau 90% pour le rapport des écarts-types théoriques σ_1/σ_2. Sous quelles conditions votre intervalle est-il approprié ?

SOLUTION. La formule (4.3) nous donne l'intervalle pour le rapport σ_1^2/σ_2^2. L'intervalle pour le rapport σ_1/σ_2 est donc

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

Avec l'aide du logiciel R on obtient

$$\begin{aligned} F_{n_1-1,n_2-1,\frac{\alpha}{2}} &= F_{11,14,0.05} = \texttt{qf(0.95,11,14)} = 2.5655, \\ F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} &= F_{11,14,0.95} = \texttt{qf(0.05,11,14)} = 0.3651. \end{aligned}$$

L'intervalle désiré est donc

$$\left(\frac{1}{\sqrt{2.5655}} \frac{4.83}{2.21} \, , \, \frac{1}{\sqrt{0.3651}} \frac{4.83}{2.21} \right) = (1.36, 3.62).$$

Cet intervalle est approprié à condition qu'on ait des échantillons aléatoires indépendants provenant de distributions normales.

Ci-dessus, on a utilisé R pour obtenir les valeurs $F_{11,14,0.05}$ et $F_{11,14,0.95}$. On peut aussi obtenir ces valeurs à partir de la table de la loi de Fisher qui apparaît à l'annexe A.4. Dans la table, On peut lire directement la valeur $F_{11,14,0.05} = 2.565$. Pour obtenir la valeur $F_{11,14,0.95}$, on utilise la propriété (4.2) et on obtient

$$F_{11,14,0.95} = \frac{1}{F_{14,11,0.05}} = \frac{1}{2.739} \approx 0.365.$$

4.2.5 Tests d'hypothèses sur deux variances

Supposons qu'on veuille tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 > \sigma_2^2$. Ces hypothèses peuvent être réécrites sous la forme $H_0 : \sigma_1^2/\sigma_2^2 = 1$ et $H_1 : \sigma_1^2/\sigma_2^2 > 1$. La règle suivante est donc une bonne règle de décision :

$$\text{On rejette } H_0 \text{ si } \frac{S_1^2}{S_2^2} \text{ est trop grand.}$$

Or si l'hypothèse H_0 est vraie, le résultat (4.1) nous assure que la statistique S_1^2/S_2^2 suit la loi F_{n_1-1,n_2-1}. Donc, au seuil α, la règle de décision précédente prend la forme suivante :

$$\text{On rejette } H_0 \text{ si } \frac{S_1^2}{S_2^2} \geq F_{n_1-1,n_2-1,\alpha}.$$

Le cas $H_1 : \sigma_1^2 < \sigma_2^2$ et le cas $H_1 : \sigma_1^2 \neq \sigma_2^2$ peuvent être traités de façon similaire. Voici les règles de décision :

Alternative	On rejette H_0 si
$H_1 : \sigma_1^2 > \sigma_2^2$	$S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\alpha}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$S_1^2/S_2^2 \leq F_{n_1-1,n_2-1,1-\alpha}$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ ou $S_1^2/S_2^2 \leq F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$

EXEMPLE 4. Si, à l'exemple 3, on nous avait demandé de tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 > \sigma_2^2$ au seuil 1%, quelle aurait été notre décision ? Calculez le *p-value*.

SOLUTION. La règle de décision nous dit de rejeter H_0 si $S_1^2/S_2^2 \geq F_{n_1-1,n_2-1,\alpha}$. Ici on obtient $s_1^2/s_2^2 = (4.83)^2/(2.21)^2 = 4.78$ et le logiciel R nous donne

$$F_{n_1-1,n_2-1,\alpha} = F_{11,14,0.01} \ = \ 3.864.$$

Donc au seuil 1%, rejette H_0. Le *p-value* est

$$\textit{p-value} = \text{ la surface à droite de 4.78 sous la densité } F_{11,14} = 0.0038.$$

Cette probabilité a été obtenue avec la commande `1 - pf(4.78, 11, 14)` dans R.

4.3 Comparaison de deux moyennes : le cas où $\sigma_1^2 = \sigma_2^2$

On suppose ici que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma^2)$.
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les variances théoriques sont égales ; leur valeur commune est dénotée σ^2.
- Les paramètres μ_1, μ_2 et σ^2 sont inconnus.

Il s'agit du même scénario qu'à la section précédente, avec une différence : nous supposons maintenant que les variances théoriques σ_1^2 et σ_2^2 sont égales. Nous écrivons σ^2 pour dénoter cette variance théorique commune aux deux populations. Notre objectif est de comparer les moyennes théoriques μ_1 et μ_2. Plus précisément, nous aimerions tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'une ou l'autre des hypothèses alternatives usuelles (unilatérale à droite $H_1 : \mu_1 > \mu_2$, unilatérale à gauche $H_1 : \mu_1 < \mu_2$ ou bilatérale $H_1 : \mu_1 \neq \mu_2$). Dans certains cas, nous aimerions plutôt estimer la différence $\mu_1 - \mu_2$, calculer l'erreur type associée à notre estimation et calculer un intervalle de confiance pour $\mu_1 - \mu_2$. Mais avant de s'attaquer à ces problèmes, il faut d'abord considérer le problème de l'estimation de la variance théorique commune à nos deux populations.

Inférence pour la variance théorique σ^2 :

À partir de nos deux échantillons, on calcule nos moyennes échantillonnales et nos variances échantillonnales de la façon usuelle :

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \qquad S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1,i}-\overline{X}_1)^2,$$

$$\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}, \qquad S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2,i}-\overline{X}_2)^2.$$

La moyenne échantillonnale $\overline{X}_1$ sert à estimer la moyenne théorique μ_1 et la moyenne échantillonnale $\overline{X}_2$ sert à estimer la moyenne théorique μ_2. L'estimation de la variance théorique σ^2 est un peu plus délicate. Nous disposons de deux estimateurs : la variance échantillonnale S_1^2 et la variance échantillonnale S_2^2. Il est naturel de combiner ces deux estimateurs. Une façon simple de combiner ces deux estimateurs serait de prendre leur moyenne arithmétique $(S_1^2 + S_2^2)/2$. Cette approche est valide dans le cas où les tailles de nos échantillons sont identiques, c'est-à-dire dans le cas où $n_1 = n_2$. Si les tailles de nos échantillons ne sont pas égales, alors on s'attend à ce que la variance échantillonnale du plus grand échantillon donne une meilleure estimation de σ^2 que la variance échantillonnale du plus petit échantillon. Plutôt que de prendre $(S_1^2 + S_2^2)/2$, on devrait alors prendre une moyenne pondérée de S_1^2 et S_2^2, avec des poids qui tiennent compte des tailles de nos échantillons. En statistique mathématique, on montre que la solution optimale est de prendre l'estimateur

$$S_c^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{n_1-1}{n_1+n_2-2}\ S_1^2 + \frac{n_2-1}{n_1+n_2-2}\ S_2^2. \tag{4.4}$$

La statistique S_c^2 est appelée la *variance échantillonnale combinée*. Dans les ouvrages de langue anglaise, on l'appelle la *pooled sample variance* et on la dénote S_p^2. Si $n_1 = n_2$, l'équation (4.4) se réduit à

$$S_c^2 = \frac{1}{2}\ S_1^2 + \frac{1}{2}\ S_2^2 = \frac{S_1^2 + S_2^2}{2}.$$

Si $n_1 = 6$ et $n_2 = 11$, l'équation (4.4) se réduit à

$$S_c^2 = \frac{5\ S_1^2 + 10\ S_2^2}{15} = \frac{1}{3}\ S_1^2 + \frac{2}{3}\ S_2^2.$$

Pour le problème à un échantillon, nous avions le résultat suivant :

$$\frac{(n-1)\ S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{4.5}$$

C'est à partir de ce résultat qu'on avait obtenu (a) une formule pour l'erreur type associée à la variance, (b) l'intervalle de confiance pour la variance théorique et (c) les règles de décision pour les tests d'hypothèses sur la variance théorique. Pour la variance échantillonnale combinée, le résultat analogue au résultat (4.5) est le suivant :

$$\frac{(n_1+n_2-2)\ S_c^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}. \tag{4.6}$$

Petit truc pour se souvenir du nombre de degrés de liberté associé à S_c^2 : le nombre de degrés de liberté associé à S_1^2 est $n_1 - 1$ et celui associé à S_2^2 est $n_2 - 1$; le nombre de degrés de liberté associé à S_c^2 est donc $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. En procédant comme dans le cas du problème à un échantillon, on arrive aux résultats suivants :

1. La variance échantillonnale combinée S_c^2 est un estimateur sans biais pour la variance théorique σ^2. Autrement dit, $\mathbb{E}[S_c^2] = \sigma^2$.
2. L'erreur type associée à la variance échantillonnale combinée est donnée par
$$\sqrt{2}\ s_c^2/\sqrt{n_1 + n_2 - 2}.$$
3. L'intervalle de confiance de niveau $1 - \alpha$ pour σ^2 est donné par
$$\left(\frac{(n_1 + n_2 - 2)\ S_c^2}{\chi^2_{n_1+n_2-2,\frac{\alpha}{2}}},\ \frac{(n_1 + n_2 - 2)\ S_c^2}{\chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}}} \right).$$
4. Pour tester $H_0 : \sigma^2 = \sigma_o^2$, on utilise la règle de décision suivante :
   - Avec $H_1 : \sigma^2 > \sigma_o^2$, on rejette H_0 si $U \geq \chi^2_{n_1+n_2-2,\alpha}$.
   - Avec $H_1 : \sigma^2 < \sigma_o^2$, on rejette H_0 si $U \leq \chi^2_{n_1+n_2-2,1-\alpha}$.
   - Avec $H_1 : \sigma^2 \neq \sigma_o^2$, on rejette H_0 si $U \geq \chi^2_{n_1+n_2-2,\frac{\alpha}{2}}$ ou si $U \leq \chi^2_{n_1+n_2-2,1-\frac{\alpha}{2}}$.

   Ici la statistique de test est $U = (n_1 + n_2 - 2)S_c^2/\sigma_o^2$.

Inférence pour la différence $\mu_1 - \mu_2$:

L'estimateur naturel de la différence $\mu_1 - \mu_2$ est la différence des moyennes échantillonnales $\overline{X}_1 - \overline{X}_2$. Jusqu'ici, pas de surprise. Mais qu'en est-il de l'erreur type ? Et comment calcule-t-on un intervalle de confiance ? Tout repose sur le résultat suivant :

Théorème : Sous les conditions énoncées au début de la présente section, on a

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1) \tag{4.7}$$

et

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}. \tag{4.8}$$

L'équation (4.7) est l'analogue de l'équation (2.23) du chapitre 2. L'équation (4.8) est l'analogue de l'équation (3.10) du chapitre 3. Voici les principales conséquences de ce théorème :

1. La différence des moyennes échantillonnales $\overline{X}_1 - \overline{X}_2$ est un estimateur sans biais pour la différence des moyennes théoriques $\mu_1 - \mu_2$. Autrement dit, on a
$$\mathbb{E}[\overline{X}_1 - \overline{X}_2] = \mu_1 - \mu_2.$$

2. L'erreur type associée à la différence des moyennes échantillonnales est donnée par

$$\text{erreur type } = s_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

3. L'intervalle de confiance de niveau $1-\alpha$ pour $\mu_1-\mu_2$ est donné par

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

4. Pour tester $H_0:\mu_1=\mu_2$, on utilise la règle de décision suivante :
   - Avec $H_1:\mu_1>\mu_2$, on rejette H_0 si $T\geq t_{n_1+n_2-2,\alpha}$.
   - Avec $H_1:\mu_1<\mu_2$, on rejette H_0 si $T\leq -t_{n_1+n_2-2,\alpha}$.
   - Avec $H_1:\mu_1\neq\mu_2$, on rejette H_0 si $|T|\geq t_{n_1+n_2-2,\frac{\alpha}{2}}$.

   Ici la statistique de test est $T=\frac{\overline{X}_1-\overline{X}_2}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$.

Exemple 5. On veut comparer l'effet de l'engrais A et l'effet de l'engrais B sur la croissance des carottes. On divise notre jardin de 500 carottes en deux sections contenant chacune 250 carottes. Dans une section on utilise l'engrais A, dans l'autre section on utilise l'engrais B. La variable d'intérêt est le poids de la carotte (en grammes) au moment de la récolte. On fait les hypothèses suivantes :

- La loi normale avec moyenne μ_A est un bon modèle pour décrire la distribution des poids des carottes soumises à l'engrais A.
- La loi normale avec moyenne μ_B est un bon modèle pour décrire la distribution des poids des carottes soumises à l'engrais B.
- Ces deux lois normales ont la même variance, disons σ^2, mais cette variance est inconnue.
- Les poids des 250 carottes soumises à l'engrais A peuvent être vus comme étant un échantillon aléatoire de taille $n_A=250$ issu de la loi $N(\mu_A,\sigma^2)$.
- Les poids des 250 carottes soumises à l'engrais B peuvent être vus comme étant un échantillon aléatoire de taille $n_B=250$ issu de la loi $N(\mu_B,\sigma^2)$.
- Ces deux échantillons sont indépendants l'un de l'autre.

Voici un résumé des données recueillies à la fin de l'expérience, au moment de la récolte. Une partie du jardin a été détruite lorsque le responsable de l'application des engrais a fait une fausse manoeuvre avec son tracteur, détruisant ainsi 23 carottes.

Engrais	A	B
Taille de l'échantillon	250	227
Moyenne échantillonnale	92.7	80.2
Écart-type échantillonnal	4.93	4.55

(a) Calculez une estimation pour $\mu_A - \mu_B$.

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

(c) Calculez un intervalle de confiance de niveau 95% pour $\mu_A - \mu_B$.

(d) Discutez la validité des méthodes utilisées en (a), (b) et (c).

SOLUTION.

(a) Notre estimation pour $\mu_A - \mu_B$ est $\overline{x}_A - \overline{x}_B = 92.7 - 80.2 = 12.5$.

(b) D'abord on calcule s_c et on obtient

$$s_c = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} = \sqrt{\frac{(249 \times (4.93)^2) + (226 \times (4.55)^2)}{475}} = 4.753.$$

L'erreur type associée à l'estimation obtenue en (a) est donc

$$\text{erreur type } = s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4.753 \times \sqrt{\frac{1}{250} + \frac{1}{227}} = 0.436$$

(c) On utilise l'intervalle donné à la page précédente. Il nous faut la quantité

$$t_{n_A+n_B-2,\alpha/2} = t_{475,0.025}.$$

Avec 475 degrés de liberté, la loi de Student est essentiellement identique à la loi $N(0,1)$. Donc on peut prendre $t_{475,0.025} \approx z_{0.025} = 1.96$. L'intervalle désiré est donc l'intervalle

$$\text{estimation } \pm \text{ (1.96 fois l'erreur type).}$$

On utilise l'estimation obtenue en (a) et l'erreur type obtenue en (b) et on obtient l'intervalle $(11.65, 13.35)$.

(d) Si on avait accès aux données (les poids des 250 carottes soumises à l'engrais A et ceux des 227 carottes soumises à l'engrais B), on pourrait examiner nos deux histogrammes pour voir si l'hypothèse de normalité est raisonnable. Mais les tailles de nos échantillons étant grandes, l'hypothèse de normalité n'est pas essentielle (en vertu du théorème limite central). Les histogrammes nous permettraient aussi de voir si l'hypothèse d'égalité des variances est raisonnable. Mais on peut utiliser un critère plus objectif (en supposant que les deux histogrammes présentent des formes de lois normales) : le test d'égalité des variances de la section 4.2. On veut tester $H_0 : \sigma_A^2 = \sigma_B^2$ contre $H_1 : \sigma_A^2 \neq \sigma_B^2$. La valeur observée de notre statistique de test est $s_A^2/s_B^2 = 1.174$. Le *p-value* est donc

$$\begin{aligned} \textit{p-value} &= 2 \times \mathbb{P}_{H_0}[S_A^2/S_B^2 \geq 1.174] \\ &= 2 \times \text{ surface à droite de 1.174 sous la densité } F_{249,226} \\ &= 2 \times 0.1095 \approx 0.22. \end{aligned}$$

Ce *p-value* est très grand. Il n'y a pas lieu de douter de l'hypothèse d'égalité des variances théoriques.

4.4 Comparaison de deux moyennes : le cas où $\sigma_1^2 \neq \sigma_2^2$

Nous reprenons ici le problème traité à la section précédente mais nous ne supposons plus que les variances théoriques sont égales. Le scénario est donc le suivant :

On suppose que

- $X_{1,1}, X_{1,2}, X_{1,3}, ..., X_{1,n_1}$ est un échantillon aléatoire de taille n_1 issu d'une population avec distribution $N(\mu_1, \sigma_1^2)$
- $X_{2,1}, X_{2,2}, X_{2,3}, ..., X_{2,n_2}$ est un échantillon aléatoire de taille n_2 issu d'une population avec distribution $N(\mu_2, \sigma_2^2)$
- Ces deux échantillons sont indépendants l'un de l'autre.
- Les paramètres μ_1, μ_2, σ_1^2 et σ_2^2 sont inconnus.

À la section précédente, nos procédures statistiques (intervalle de confiance et règles de décision pour tests d'hypothèses) étaient basées sur le résultat (4.8). Dans le présent contexte, on peut utiliser le résultat suivant, démontré par le mathématicien britannique Bernard L. Welch en 1947 :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_{k_*} \tag{4.9}$$

avec

$$k_* = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\sigma_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{\sigma_2^2}{n_2}\right)^2}. \tag{4.10}$$

Détails techniques :

(a) La notation $\approx t_{k_*}$ est utilisé ci-dessus pour signifier que la statistique suit *a peu près* la loi de Student avec k_* degrés de liberté.

(b) Jusqu'à maintenant, toutes les fois que nous avons rencontré une loi de Student avec k degrés de liberté, le nombre k était un entier positif. Or ici le nombre k_* donné par l'équation (4.10) n'est pas nécessairement un entier. Cela ne pose pas de problème. La loi de Student avec k_* degrés de liberté est bien définie dès que $k_* > 0$, entier ou pas. Sa densité est donnée par l'équation suivante :

$$f(t) = \frac{\Gamma\left((k_* + 1)/2\right)}{\Gamma\left(k_*/2\right)\sqrt{k_*\pi}} \frac{1}{\left(1 + (t^2/k_*)\right)^{(k_*+1)/2}}.$$

(c) Bien que les tables de la loi de Student se limitent au cas où le nombre de degrés de liberté est un entier positif, avec le logiciel R on peut spécifier un nombre de degré de liberté quelconque, entier ou pas.

(d) Lorsqu'on utilise de résultat (4.9) pour faire un test d'hypothèse ou pour calculer un intervalle de confiance, on doit obtenir certains quantiles de la loi de Student t_{k_*}. Or k_* est inconnu puisqu'il dépend des variances théoriques et que celles-ci sont inconnues. Pour contourner ce problème, on remplace σ_1^2 et σ_2^2 par s_1^2 et s_2^2 dans

l'équation (4.10). Autrement dit, on utilise k degrés de liberté, avec k donné par l'équation suivante :

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}. \tag{4.11}$$

Le test de Welch. Sous les conditions énoncées au début de la présente section, voici les règles de décision pour tester $H_0 : \mu_1 = \mu_2$ contre les hypothèse alternatives usuelles :

- Avec $H_1 : \mu_1 > \mu_2$, on rejette H_0 si $T' \geq t_{k,\alpha}$.
- Avec $H_1 : \mu_1 < \mu_2$, on rejette H_0 si $T' \leq -t_{k,\alpha}$.
- Avec $H_1 : \mu_1 \neq \mu_2$, on rejette H_0 si $|T'| \geq t_{k,\frac{\alpha}{2}}$.

Ici la statistique de test est

$$T' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

et le nombre de degrés de liberté k est donné par l'équation (4.11).

Exemple 6 : On obtient un échantillon aléatoire de $n_1 = 6$ louveteaux du Yukon et un échantillon aléatoire de $n_2 = 9$ louveteaux de Sibérie. On s'intéresse au poids des louveteaux à la naissance. On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. On peut supposer que les deux distributions sont normales mais rien ne nous permet de croire *a priori* que les variances théoriques sont égales. Voici les données :

Yukon : 4.76, 3.58, 5.04, 4.84, 4.29, 4.37.
Sibérie : 3.32, 7.53, 5.83, 8.22, 4.70, 5.20, 6.36, 9.73, 3.38.

Si on fait le test de la section 4.2 pour $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, on obtient un *p-value* égal à 0.0063. Il ne serait donc pas du tout raisonnable de supposer que $\sigma_1^2 = \sigma_2^2$ et d'appliquer la procédure présentée à la section 4.3. L'hypothèse de normalité, quant à elle, est raisonnable. C'est une hypothèse difficile à vérifier à partir d'échantillons de tailles 6 et 9. Mais les histogrammes obtenus avec le logiciel R n'ont rien d'alarmant et on sait que la loi normale est presque toujours un bon modèle pour décrire les distributions de poids dans les populations animales. On peut donc utiliser le test de Welch. Avec les données ci-dessus, on obtient

$$\overline{x}_1 = 4.48 \qquad s_1 = 0.5253 \qquad \overline{x}_2 = 6.03 \qquad s_2 = 2.1711.$$

Pour déterminer le nombre de degrés de liberté approprié, on utilise l'équation (4.11) et on obtient

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \;=\; 9.35.$$

La valeur observée de notre statistique de test est

$$T'_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \;=\; -2.0535.$$

Notre *p-value* est donc

$$\begin{aligned} \textit{p-value} &= \mathbb{P}_{H_0}[|T'| \geq 2.0535] \\ &\approx \text{ deux fois la surface à droite de 2.0535 sous la densité } t_{9.35} \\ &= 2 \times 0.03453 \;=\; 0.069. \end{aligned}$$

Conclusion : au seuil 5%, il n'y a pas lieu de rejeter H_0.

Dans cet exemple, on a obtenu la valeur 0.03453, c'est-à-dire la surface à droite de 2.0535 sous la densité $t_{9.35}$, avec l'aide du logiciel R. Si on a seulement accès à une table de la loi de Student, on peut obtenir une bonne approximation en interpolant entre les valeurs qu'on obtient avec 9 et 10 degrés de liberté. La table nous donne

$$\begin{aligned} \text{surface à droite de 2.0535 sous la densité } t_9 &\approx 0.03715 \\ \text{surface à droite de 2.0535 sous la densité } t_{10} &\approx 0.03675. \end{aligned}$$

Chacune de ces deux valeurs a été obtenue par interpolation linéaire. Enfin, on interpole entre ces deux valeurs et on obtient

$$\text{surface à droite de 2.0535 sous la densité } t_{9.35} \;\approx 0.03701.$$

La table nous donne donc

$$\textit{p-value} \approx 2 \times 0.03701 = 0.074.$$

C'est beaucoup de calcul et on comprend pourquoi plusieurs auteurs suggèrent de remplacer l'équation (4.11) par

$$k = \text{ l'entier le plus proche de } \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

4.5 Comparaison de deux moyennes : le cas où les données sont appariées

Voici un exemple illustratif. Supposons qu'une certaine espèce animale soit atteinte d'une maladie qui affecte les griffes des pattes avant. Deux traitements ont été proposés. On dispose de seulement $n = 15$ animaux malades sur lesquels on doit essayer nos 2 traitements afin de les comparer et de déterminer s'il y en a un qui est meilleur que l'autre.

Scénario 1 : Une approche possible serait d'utiliser le traitement A sur 8 animaux et le traitement B sur les 7 autres. On obtiendrait ainsi deux échantillons aléatoires indépendants de tailles respectives $n_1 = 8$ et $n_2 = 7$.

Scénario 2 : Une approche alternative consiste à utiliser les deux traitements sur chacun des $n = 15$ animaux : le traitement A sur une des deux pattes avant et le traitement B sur l'autre patte. Nous aurons ainsi deux échantillons aléatoires de taille $n = 15$. Mais attention ! Nos deux échantillons de tailles $n = 15$ ne seront pas des échantillons indépendants !

Notez qu'on suppose ici qu'il est possible d'appliquer un traitement sur une patte et l'autre traitement sur l'autre patte. Ceci est possible par exemple dans le cas où les traitements sont des crèmes qu'on applique directement sur les pattes. Si les traitements étaient des médicaments que les animaux doivent avaler, le scénario 2 ne serait pas applicable.

Le scénario 1 correspond au schéma suivant. On dispose de deux paniers, le panier A et le panier B. Le panier A correspond à la population des animaux malades et la variable d'intérêt est la réponse au traitement A alors que le panier B correspond à la population des animaux malades et la variable d'intérêt est la réponse au traitement B. Nous faisons 8 tirages à partir du panier A et 7 tirages à partir du panier B. Avec ce scénario, et en supposant que les conditions de normalité et d'égalité des variances théoriques sont vérifiées, on utilise le test présenté à la section 4.3.

Le scénario 2 correspond au schéma suivant. Il y a un seul panier. Ce panier représente la population de tous les animaux atteints de la maladie des griffes des pattes avant. Chaque boule du panier représente un animal malade. Sur chaque boule il y a deux nombres : le premier représente la réponse au traitement A et le deuxième représente la réponse au traitement B. Nous faisons $n = 15$ tirages à partir de ce panier. Au lieu d'avoir deux échantillons aléatoires indépendants, nous avons maintenant un seul échantillon aléatoire. Il s'agit d'un échantillon aléatoire *bivarié* de taille $n = 15$:

$$(X_{1,1}, X_{2,1}), (X_{1,2}, X_{2,2}), (X_{1,3}, X_{2,3}), ..., (X_{1,15}, X_{2,15}).$$

On dit alors que les données sont *appariées*. Elles sont en paires. Dans la paire $(X_{1,7}, X_{2,7})$, le $X_{1,7}$ représente la réponse de l'animal numéro 7 au traitement A alors que le $X_{2,7}$ représente la réponse de l'animal numéro 7 au traitement B. Pour analyser ces données appariées, nous allons considérer les différences

$$D_1 = X_{1,1} - X_{2,1}, \quad D_2 = X_{1,2} - X_{2,2}, \quad D_3 = X_{1,3} - X_{2,3}, ... \quad D_{15} = X_{1,15} - X_{2,15}$$

Nous allons supposer que les observations $D_1, D_2, D_3, ..., D_{15}$ constituent un échantillon aléatoire de taille $n = 15$ issu d'une population avec distribution $N(\mu_D, \sigma_D^2)$. L'hypothèse $H_0 : \mu_1 = \mu_2$ peut alors s'écrire sous la forme $H_0 : \mu_D = 0$. L'hypothèse $H_1 : \mu_1 \neq \mu_2$ prend la forme $H_1 : \mu_D \neq 0$. De même, l'hypothèse $H_1 : \mu_1 > \mu_2$ prend la forme $H_1 : \mu_D > 0$ et l'hypothèse $H_1 : \mu_1 < \mu_2$ prend la forme $H_1 : \mu_D < 0$. Nous sommes maintenant en présence d'un problème de test d'hypothèse sur la moyenne d'une seule distribution normale. Nous avons déjà vu comment traiter ce problème. Avec un échantillon bivarié de taille n, la règle de décision au seuil α prend donc la forme suivante :

- Avec $H_1 : \mu_D > 0$, on rejette H_0 si $T \geq t_{n-1,\alpha}$.
- Avec $H_1 : \mu_D < 0$, on rejette H_0 si $T \leq -t_{n-1,\alpha}$.
- Avec $H_1 : \mu_D \neq 0$, on rejette H_0 si $|T| \geq t_{n-1,\frac{\alpha}{2}}$.

Ici la statistique de test est $T = \frac{\overline{D}}{S_D/\sqrt{n}}$.

Exemple 7 : Voici un exemple numérique illustratif. On veut comparer deux traitements. Les deux traitements sont utilisés sur 15 animaux malades. Chaque animal reçoit le traitement A sur une patte et le traitement B sur l'autre patte. Avec chaque animal, on lance

une pièce de monnaie pour déterminer quelle patte reçoit le traitement A et quelle patte reçoit le traitement B. Nous n'avons aucune raison de soupçonner que l'un ou l'autre des deux traitements est meilleur que l'autre. Nous allons donc faire un test bilatéral :

H_0 : Les deux traitements sont équivalents.

H_1 : Un des deux traitements est meilleur que l'autre.

Notre règle de décision sera basée sur les différences

$D_j =$ (réponse de l'animal j au traitement A) $-$ (réponse de l'animal j au traitement B).

Nous allons supposer que la distribution théorique des différences est la loi $N(\mu_D, \sigma_D^2)$. Nos hypothèses peuvent donc s'écrire sous la forme

$$H_0 : \mu_D = 0 \qquad \text{et} \qquad H_1 : \mu_D \neq 0.$$

Voici les données :

Animal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Trait. A	4.5	2.8	3.7	6.0	2.9	5.6	1.7	3.7	3.6	3.8	4.1	5.2	4.7	5.4	5.8
Trait. B	3.8	2.9	4.0	5.6	2.6	5.4	1.6	3.8	3.5	3.3	3.8	5.0	4.7	4.9	5.4
Dif.	0.7	-0.1	-0.3	0.4	0.3	0.2	0.1	-0.1	0.1	0.5	0.3	0.2	0.0	0.5	0.1

La moyenne échantillonnale des différences est $\overline{d} = 0.213$. L'écart-type échantillonnal des différences est $s_D = 0.270$. La valeur observée de notre statistique de test est donc

$$T_{obs} = \frac{\overline{d}}{s_D/\sqrt{n}} = \frac{0.213}{0.270/\sqrt{15}} = 3.065.$$

Le *p-value* est donc

$$\textit{p-value} = \text{deux fois la surface à droite de } 3.065 \text{ sous la densité } t_{14} = 0.0084.$$

Conclusion : au seuil 1% on rejette H_0.

Remarque : Que se serait-il passé si on n'avait pas tenu compte du fait que les données sont appariées et si on avait naïvement utilisé le test de Student présenté à la section 4.3 ? Faisons les calculs. On obtient $\overline{x}_A = 4.233$ et $s_A = 1.238$. De même on obtient $\overline{x}_B = 4.020$ et $s_B = 1.152$. La variance échantillonnale combinée est

$$s_c = \sqrt{\frac{(n_1-1)s_A^2 + (n_2-1)s_B^2}{n_1+n_2-2}} = \sqrt{\frac{(15-1)(1.238)^2 + (15-1)(1.152)^2}{15+15-2}} = 1.196.$$

La valeur observée de notre statistique de test serait donc

$$T_{obs} = \frac{\overline{x}_A - \overline{x}_B}{s_c\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{4.233 - 4.020}{1.196\sqrt{\frac{1}{15} + \frac{1}{15}}} = 0.489.$$

Le *p-value* serait donné par

$$\begin{aligned} \textit{p-value} &= \text{deux fois la surface à droite de } 0.489 \text{ sous la densité } t_{28} \\ &= 0.629. \end{aligned}$$

La conclusion serait de ne pas rejeter H_0.

4.6 Comparaison de deux proportions

On considère deux populations et on s'intéresse à une certaine variable dichotomique. Pour alléger le texte, supposons que les deux valeurs possibles de cette variable dichotomique sont *malade* et *en santé*. On s'intéresse aux proportions d'individus malades dans ces deux populations. On écrit p_1 pour la proportion d'individus malades dans la population 1 et p_2 pour la proportion d'individus malades dans la population 2.

Les données :

- On obtient un échantillon aléatoire de taille n_1 à partir de la population 1 et on écrit $\hat{p}_1$ pour la proportion échantillonnale calculée à partir de cet échantillon.
- On obtient un échantillon aléatoire de taille n_2 à partir de la population 2 et on écrit $\hat{p}_2$ pour la proportion échantillonnale calculée à partir de cet échantillon.
- On suppose que ces deux échantillons sont indépendants l'un de l'autre.

La distribution de $\hat{p}_1 - \hat{p}_2$:

On sait que si n_1 et n_2 sont suffisamment grands, alors

$$\hat{p}_1 \approx N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \hat{p}_2 \approx N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

Il s'ensuit que si n_1 et n_2 sont suffisamment grands, alors

$$\hat{p}_1 - \hat{p}_2 \;\approx\; N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

et donc

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \;\approx\; N(0,1). \tag{4.12}$$

Tout ce qui suit est basé sur le résultat (4.12).

Estimation de $p_1 - p_2$:

Pour estimer la différence $p_1 - p_2$, on utilise l'estimateur $\hat{p}_1 - \hat{p}_2$. Cet estimateur est sans biais :

$$\mathbb{E}[\hat{p}_1 - \hat{p}_2] = \mathbb{E}[\hat{p}_1] - \mathbb{E}[\hat{p}_2] = p_1 - p_2.$$

La variance de l'estimateur $\hat{p}_1 - \hat{p}_2$ est donnée par

$$\mathbb{V}\text{ar}[\hat{p}_1 - \hat{p}_2] = \mathbb{V}\text{ar}[\hat{p}_1] + \mathbb{V}\text{ar}[\hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

On peut donc calculer l'erreur type associée à l'estimation $\hat{p}_1 - \hat{p}_2$ de la façon suivante :

$$\text{erreur type} \;= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

L'intervalle de confiance de niveau $1-\alpha$ pour $p_1 - p_2$ est donné par

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\, ,\, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right).$$

Notons que cet intervalle a la forme usuelle :

$$\text{estimation } \pm \text{ quelques fois l'erreur type.}$$

Tests d'hypothèses pour comparer p_1 et p_2 :

Nous voulons tester $H_0 : p_1 = p_2$ contre l'une ou l'autre des trois hypothèses alternatives usuelles. Lorsque H_0 est vraie, le résultat (4.12) prend la forme suivante :

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \;\approx\; N(0,1) \tag{4.13}$$

où p_0 dénote la valeur commune des deux proportions théoriques. En supposant que H_0 est vraie, comment estime-t-on ce p_0 ? Il suffit de prendre

$$\hat{p}_0 = \frac{\text{nombre total de malades dans les 2 échantillons}}{\text{nombre total d'observations}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

À partir du résultat (4.13), on obtient

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \;\approx\; N(0,1). \tag{4.14}$$

C'est à partir de ce résultat (4.14) que nous obtenons les règles de décision suivantes pour tester $H_0 : p_1 = p_2$:

- Avec $H_1 : p_1 > p_2$, on rejette H_0 si $Z \geq z_\alpha$.
- Avec $H_1 : p_1 < p_2$, on rejette H_0 si $Z \leq -z_\alpha$.
- Avec $H_1 : p_1 \neq p_2$, on rejette H_0 si $|Z| \geq z_{\frac{\alpha}{2}}$.

Ici Z est la statistique donnée par l'équation (4.14).

Exemple 8. On s'intéresse à une certaine maladie chez les perchaudes. On se demande si la proportion de perchaudes malades est la même dans le lac St-Augustin que dans le lac St-Pierre. Nous allons tester $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$. Ici p_1 dénote la proportion de perchaudes malades dans le lac St-Augustin et p_2 dénote cette même proportion dans le lac St-Pierre. Parmi 50 perchaudes tirées du lac St-Augustin, 23 sont malades. Parmi 60 perchaudes tirées du lac St-Pierre, seulement 12 sont malades. Que doit-on conclure ?

Solution. On a $n_1 = 50, \hat{p}_1 = 0.46, n_2 = 60$ et $\hat{p}_2 = 0.20$. On obtient $\hat{p}_0 = (23+12)/(50+60) = 35/110 = 0.3182$. On obtient donc

$$Z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.46 - 0.20}{\sqrt{0.3182(1-0.3182)\left(\frac{1}{50} + \frac{1}{60}\right)}} = 2.915.$$

Le *p-value* est donc

$$p\text{-}value \ = \mathbb{P}_{H_0}[|Z| \geq 2.915] = 2 \times \mathbb{P}_{H_0}[Z \geq 2.915] = 0.0036.$$

Conclusion : au seuil 1% on rejette H_0.

Exemple 9. Suite au résultat obtenu ci-dessus, on décide d'estimer $p_1 - p_2$ avec précision. Parmi 300 perchaudes tirées du lac St-Augustin (incluant les 50 perchaudes ci-dessus), 134 sont malades. Parmi 300 perchaudes tirées du lac St-Pierre (incluant les 60 perchaudes ci-dessus), 77 sont malades. Calculez l'estimation de $p_1 - p_2$. Quelle est l'erreur type associée à cette estimation ? Calculez un intervalle de confiance de niveau 95% pour $p_1 - p_2$.

Solution. L'estimation de $p_1 - p_2$ est $\hat{p}_1 - \hat{p}_2 = \frac{134}{300} - \frac{77}{300} = 0.190$. L'erreur type associée à cette estimation est

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.038.$$

L'intervalle de confiance de niveau 95% est

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

c'est-à-dire $0.190 \pm 1.96 \times 0.038$, c'est-à-dire 0.190 ± 0.075, c'est-à-dire $(0.115, 0.265)$.

4.7 Le test de Wilcoxon-Mann-Whitney [1]

4.7.1 Introduction

Dans tous les problèmes que nous avons considérés jusqu'à maintenant, la distribution de la variable d'intérêt était connue à un ou deux paramètres près. En fait, dans la plupart des cas, cette distribution était simplement la loi normale. On dit alors qu'on est en présence d'un modèle statistique *paramétrique* et les méthodes d'inférence statistique qu'on utilise sont dites *méthodes paramétriques*.

Dans certains problèmes, on préfère ne faire aucune hypothèse sur la forme de la distribution de la variable d'intérêt. On parle alors de modèle statistique *non paramétrique* et les méthodes statistiques propres à ces modèles sont dites *non paramétriques*. Dans la présente section, nous considérons une méthode non paramétrique appelée le *test de la somme des rangs* de Wilcoxon-Mann-Whitney.

4.7.2 Le scénario

On considère deux populations, disons la population 1 et la population 2, et on s'intéresse à une certaine variable numérique de type continu. On veut comparer les moyennes théoriques μ_1 et μ_2. On dispose d'échantillons aléatoires indépendants et on suppose que les deux distributions théoriques ont la même forme mais possiblement des moyennes différentes. On a donc

1. On peut omettre cette section si on manque de temps

(i) $X_{1,1}, X_{1,2}, ..., X_{1,n_1}$, un échantillon aléatoire de taille n_1 issu de la population 1.

(ii) $X_{2,1}, X_{2,2}, ..., X_{2,n_2}$, un échantillon aléatoire de taille n_2 issu de la population 2.

(iii) Ces deux échantillons sont indépendants l'un de l'autre.

(iv) Les distributions théoriques ont la même forme mais avec, possiblement, des moyennes différentes. Autrement dit, la densité de probabilité de la population 1 et la densité de probabilité de la population 2 sont identiques à une translation près.

On veut tester $H_0 : \mu_1 = \mu_2$ contre l'une ou l'autre des trois hypothèses alternatives usuelles. Il est important de noter que les quatre conditions énoncées ci-dessus sont presque les mêmes que les quatre conditions du test de Student pour comparer deux moyennes. Pour le test de Student on supposait que la loi normale était un bon modèle pour chacune de nos deux populations ; dans ce cas la condition (iv) revient à dire que ces deux lois normales ont la même variance.

4.7.3 La règle de décision

Pour fixer les idées, imaginez qu'on veuille tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. On procède de la façon suivante :

1. On place nos $n_1 + n_2$ observations en ordre croissant, de la plus petite à la plus grande.
2. On attribue à nos $n_1 + n_2$ observations les *rangs* 1 à $n_1 + n_2$ de la façon suivante : la plus petite observation reçoit le rang 1, la deuxième plus petite observation reçoit le rang 2, la troisième plus petite observation reçoit le rang 3, etc.
3. On pose $W =$ *la somme des rangs des observations issues de la population 1.*

Si H_0 est vraie, alors nos deux populations ont exactement la même distribution. Dans ce cas, il n'y a pas de raison pour que les observations issues de la population 1 aient tendance à recevoir des rangs plus élevés que celles issues de la population 2. Mais si c'est H_1 qui est vraie, alors les observations issues de la population 1 auront tendance à être plus grandes que celles issues de la populations 2. Elles auront donc tendance à recevoir des rangs plus élevés. Il est donc raisonnable d'utiliser le règle de décision suivante, proposée en 1945 par Frank Wilcoxon :

$$\text{on rejette } H_0 \text{ si } W \text{ est trop grand.} \tag{4.15}$$

Ici, « trop grand » veut dire « beaucoup plus grand que ce à quoi on devrait s'attendre si l'hypothèse nulle H_0 était vraie ». Mais alors, à quoi devrions-nous nous attendre si H_0 était vraie ? Nous répondons à cette question à la section suivante.

Remarque. La règle de décision (4.15) fut proposée en 1945 par Wilcoxon. La statistique W est appelée la statistique de la somme des rangs de Wilcoxon. En 1947, Mann et Whitney ont étudié le même problème et ont proposé une règle de décision qui à première vue semblait différente mais qui finalement s'est avérée être équivalente à celle de Wilcoxon. Pour cette raison, le test d'hypothèses décrit dans la présente section est parfois appelé le test de la somme des rangs de Wilcoxon-Mann-Whitney.

4.7.4 La distribution de la statistique de Wilcoxon

Que H_0 soit vraie ou non, il est facile de voir que la plus petite valeur possible de la statistique W de Wilcoxon est la valeur

$$w_{min} = 1 + 2 + 3 + \cdots + n_1 = \frac{n_1(n_1+1)}{2}.$$

Cette valeur est obtenue si ce sont les n_1 observations issues de la population 1 qui recoivent les rangs 1, 2, 3,..., n_1. Ceci survient si et seulement si les n_1 observations issues de la population 1 sont toutes plus petites que chacune des n_2 observations issues de la population 2. De même, la plus grande valeur possible de W est la valeur

$$w_{max} = (n_2+1) + (n_2+2) + (n_2+3) + \cdots + (n_2+n_1) = \frac{n_1(n_1+1)}{2} + n_1 n_2.$$

Cette valeur est obtenue si ce sont les n_1 observations issues de la population 1 qui recoivent les rangs n_2+1, n_2+2, n_2+3,..., n_2+n_1. Ceci survient si et seulement si les n_1 observations issues de la population 1 sont toutes plus grandes que chacune des n_2 observations issues de la population 2. Finalement, l'ensemble des valeurs possibles de W est simplement l'ensemble de tous les entiers compris entre w_{min} et w_{max}, incluant w_{min} et w_{max}.

Que peut-on dire de plus dans le cas où H_0 est vraie ? Bien qu'elle soit difficile à calculer, la distribution, sous H_0, de la statistique W de Wilcoxon est facile à décrire grâce au résultat suivant.

THÉORÈME. Si H_0 est vraie, alors la distribution de la statistique W de Wilcoxon est la même que la distribution de la somme des résultats de n_1 tirages sans remise fait à partir d'un panier contenant $n_1 + n_2$ boules numérotées 1, 2, 3,..., $n_1 + n_2$.

Voici quelques conséquences de ce théorème. Les démonstrations sont omises.

1. $\mathbb{E}_{H_0}[W] = \frac{n_1(n_1+n_2+1)}{2}$
2. $\mathbb{V}\text{ar}_{H_0}[W] = \frac{n_1 n_2(n_1+n_2+1)}{12}$
3. Sous H_0, la distribution de W est symétrique.
4. Si n_1 et n_2 sont grands, disons $n_1 \geq 10$ et $n_2 \geq 10$, alors, toujours sous H_0,

$$\frac{W - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \approx N(0,1)$$

c'est-à-dire

$$W \approx N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right). \tag{4.16}$$

QUELQUES REMARQUES :

- Lorsqu'on calcule un *p-value* à l'aide du résultat (4.16), on utilise toujours la correction pour la continuité.
- Dans le cas où n_1 et n_2 sont petits, il existe des tables pour la distribution de W sous H_0 et ces tables nous permettent d'obtenir nos *p-values*.
- Dans le logiciel R, la commande `dwilcox(k,`n_1`,`n_2`)` nous donne la probabilité $\mathbb{P}_{H_0}[W = k]$ et la commande `pwilcox(k,`n_1`,`n_2`)` nous donne la probabilité $\mathbb{P}_{H_0}[W \leq k]$.

4.7.5 Un premier exemple illustratif

Voici un exemple élémentaire pour illustrer les résultats de la section précédente. Supposons que $n_1 = 3$ et $n_2 = 4$. La plus petite valeur possible et la plus grande valeur possible de W sont

$$w_{min} = 1 + 2 + 3 = 6 \qquad \text{et} \qquad w_{max} = 5 + 6 + 7 = 18.$$

L'ensemble des valeurs possibles de W est donc l'ensemble

$$\{6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}.$$

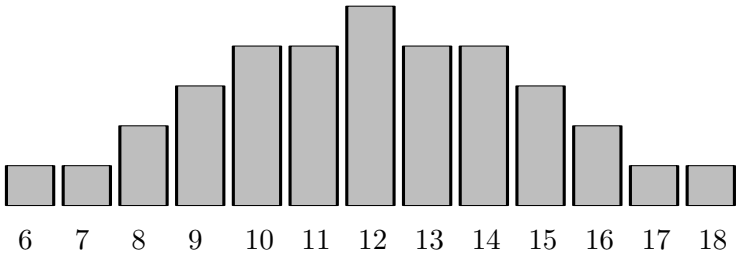
Sous H_0, l'espérance et la variance de W sont données par

$$\begin{aligned} \mathbb{E}_{H_0}[W] &= \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{3(3 + 4 + 1)}{2} = 12, \\ \mathbb{V}\text{ar}_{H_0}[W] &= \frac{n_1 n_2(n_1 + n_2 + 1)}{12} = \frac{(3 \times 4)(3 + 4 + 1)}{12} = 8. \end{aligned}$$

Notez que $\mathbb{E}_{H_0}[W]$ est exactement à mi-chemin entre w_{min} et w_{max}, conformément avec le point 3 de la page précédente. On peut trouver la distribution exacte de W sous H_0. Il suffit de considérer les $\binom{7}{3} = 35$ façons différentes de choisir les 3 rangs qui seront attribués aux 3 observations issues de la population 1. C'est ce que nous faisons dans le tableau de la page suivante. Sous H_0, ces 35 cas sont équiprobables. Pour chacun des 35 cas on calcule la valeur de W. On en déduit ensuite la distribution de W. Par exemple, dans le tableau présenté à la page suivante, on note qu'il y a quatre cas pour lesquels on obtient $W = 10$. On obtient donc $\mathbb{P}_{H_0}[W = 10] = 4/35$. De la même façon on obtient $\mathbb{P}_{H_0}[W = k]$ pour chaque valeur possible k. Voici donc, sous forme de tableau, la distribution de W (sous l'hypothèse nulle) :

k	6	7	8	9	10	11	12	13	14	15	16	17	18
$\mathbb{P}_{H_0}[W = k]$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{5}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{3}{35}$	$\frac{2}{35}$	$\frac{1}{35}$	$\frac{1}{35}$

On note que cette distribution est bel et bien symétrique, conformément au point 3 de la page précédente. Voici un graphique de cette distribution :



Même si n_1 et n_2 ne sont pas très grands, on note que la distribution a quand même une forme de cloche symétrique, conformément au point 4 ci-dessus.

Voici le tableau qui nous a permis d'obtenir la distribution de W sous H_0 :

Rangs échantillon 1	Probabilité	Valeur de W
$1-2-3$	$1/35$	6
$1-2-4$	$1/35$	7
$1-2-5$	$1/35$	8
$1-2-6$	$1/35$	9
$1-2-7$	$1/35$	10
$1-3-4$	$1/35$	8
$1-3-5$	$1/35$	9
$1-3-6$	$1/35$	10
$1-3-7$	$1/35$	11
$1-4-5$	$1/35$	10
$1-4-6$	$1/35$	11
$1-4-7$	$1/35$	12
$1-5-6$	$1/35$	12
$1-5-7$	$1/35$	13
$1-6-7$	$1/35$	14
$2-3-4$	$1/35$	9
$2-3-5$	$1/35$	10
$2-3-6$	$1/35$	11
$2-3-7$	$1/35$	12
$2-4-5$	$1/35$	11
$2-4-6$	$1/35$	12
$2-4-7$	$1/35$	13
$2-5-6$	$1/35$	13
$2-5-7$	$1/35$	14
$2-6-7$	$1/35$	15
$3-4-5$	$1/35$	12
$3-4-6$	$1/35$	13
$3-4-7$	$1/35$	14
$3-5-6$	$1/35$	14
$3-5-7$	$1/35$	15
$3-6-7$	$1/35$	16
$4-5-6$	$1/35$	15
$4-5-7$	$1/35$	16
$4-6-7$	$1/35$	17
$5-6-7$	$1/35$	18

EXEMPLE NUMÉRIQUE : On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. La règle de décision nos dit de rejeter H_0 si W est trop grand. On obtient les données suivantes :

échantillon 1	20.6	43.7	18.5	
échantillon 2	12.7	41.7	15.5	17.4

Voici les rangs :

échantillon 1	20.6	43.7	18.5	
rang	5	7	4	
échantillon 2	12.7	41.7	15.5	17.4
rang	1	6	2	3

La valeur observée de notre statistique W est donc

$$W_{obs} = 5 + 7 + 4 = 16.$$

Notre *p-value* est donc

$$p\text{-}value \ = \mathbb{P}_{H_0}[W \geq 16] = \frac{2}{35} + \frac{1}{35} + \frac{1}{35} = \frac{4}{35} \approx 0.1143.$$

Il n'y a pas lieu de rejeter H_0.

4.7.6 Un deuxième exemple illustratif

On veut tester $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. Attention : il s'agit d'un test bilatéral ! La règle de décision sera donc de la forme *on rejette H_0 si W est ou bien trop petit, ou bien trop grand*. On utilise des échantillons indépendants de tailles $n_1 = 10$ et $n_2 = 10$. Voici les observations et leurs rangs :

échantillon 1	15	55	45	11	26	66	12	27	69	39
rang	3	11	10	1	5	13	2	6	14	9
échantillon 2	81	84	61	17	97	72	28	33	85	73
rang	17	18	12	4	20	15	7	8	19	16

La valeur observée de notre statistique de Wilcoxon W est

$$W_{obs} = 3 + 11 + 10 + 1 + 5 + 13 + 2 + 6 + 14 + 9 = 74.$$

Sous H_0, on a

$$\begin{aligned} \mu_W &= \mathbb{E}_{H_0}[W] = \frac{n_1(n_1 + n_2 + 1)}{2} = 105 \\ \sigma_W &= \sqrt{\mathbb{V}\mathrm{ar}_{H_0}[W]} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{175} = 13.23. \end{aligned}$$

Comme on fait un test bilatéral, et comme la distribution de W sous H_0 est symétrique, notre *p-value* est donné par

$$p\text{-}value \ = 2 \times \mathbb{P}_{H_0}[W \leq 74].$$

À l'aide d'une table de la distribution du W de Wilcoxon, on obtient un *p-value* de 0.01854. Pour construire une telle table à la main, il faudrait procéder comme à la section 4.7.5. Pas facile ! Le tableau de la section 4.7.5 comprenait $\binom{7}{3} = 35$ lignes. Avec $n_1 = n_2 = 10$, le tableau compterait maintenant $\binom{20}{10} = 184\,756$ lignes ! Heureusement de telles tables existent.

On peut aussi obtenir une bonne approximation du *p-value* à l'aide de l'approximation gaussienne (4.16) :

$$W \ \approx \ N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

Avec $n_1 = 10$ et $n_2 = 10$, ce résultat nous donne, toujours sous H_0, $W \approx N(105, 175)$. On obtient donc

$$\begin{aligned} p\text{-}value &= 2 \times \mathbb{P}_{H_0}[W \leq 74] \\ &\approx 2 \times \mathbb{P}\left[Z \leq \frac{74.5 - 105}{\sqrt{175}}\right] \\ &= 2 \times \mathbb{P}[Z \leq -2.3056] \\ &= 2 \times 0.0106 \ = \ 0.0212. \end{aligned}$$

Cette approximation du *p-value* est très proche de la valeur exacte de 0.01854 obtenue à partir d'une table.

Enfin, on peut faire le test de Wilcoxon avec le logiciel R. On tape la commande

```
wilcox.test(c(15,55,45,11,26,66,12,27,69,39),c(81,84,61,17,97,72,28,33,85,73))
```

et le logiciel R nous donne le *p-value* 0.01854.

Conclusion. Est-ce qu'on accepte ou est-ce qu'on rejette H_0 ? Pas facile ! Nous sommes ici dans la zone grise : au seuil 5% on rejette H_0 alors qu'au seuil 1% on accepte H_0.

4.8 Exercices

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 la moyenne, l'écart-type, le 5^e centile et le 95^e centile.

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 $N(2.080, (0.050)^2)$. Celle 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 $\overline{x}_{\mathrm{A}}$, s_{A}, $\overline{x}_{\mathrm{B}}$, s_{B}.

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

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

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) Que vaut $\mathbb{P}[N \geq 4]$?

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 écarts-types théoriques σ_A/σ_B.

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 ?

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.

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 $\mu_{\mathrm{A}} - \mu_{\mathrm{B}}$.

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

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

(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) ?

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 ce σ^2.

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 nouveau-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. Ici μ_1 est la réponse moyenne à la crème A et 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 une 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.

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.

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^o C à -2^o 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 $\mu_A - \mu_B$.

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 et 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.

NUMÉRO 12. À l'Université de Montréal, deux types d'étudiants prennent le cours IFT-12550 *Introduction au langage 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* ?

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.

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 ?

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 1%, 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.

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.

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

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.

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

$$\text{on rejette } H_0 \text{ si } W \text{ est trop grand.}$$

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 ?

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.