CENTRE DE RECHERCHE
CERVO
BRAIN RESEARCH CENTRE
QUÉBEC

Comparaisons statistiques de distributions

Quentin Perry-Auger (quentin.perry-auger.1@ulaval.ca)

Stagiaire CERVO
Été 2018

Résumé

La comparaison de deux échantillons pour savoir s'il y a une distinction entre les deux n'est pas toujours évidente. Ce document a pour but de détailler une approche par les tests statistiques pour effectuer la comparaison. La définition des tests d'hypothèse est d'abord donnée, puis quelques tests de normalité et de comparaison sont présentés.

1 Tests d'hypothèse et définitions de base

Bien qu'ils ne soient pas la seule méthode de comparaison statistique, les tests d'hypothèse sont utilisés dans de nombreux domaines. L'objet de ceux-ci est de tester une hypothèse nulle et de prendre une décision à savoir si cette hypothèse est statistiquement probable [1]. Dans le cas contraire, l'hypothèse alternative est plutôt acceptée. Par exemple, l'hypothèse nulle pourrait être qu'un échantillon provient d'une population, soit l'ensemble duquel l'échantillon a été prélevé, dont la distribution est normale. L'hypothèse alternative serait alors que l'échantillon provient d'une population dont la distribution n'est pas normale.

Afin de prendre une décision sur l'hypothèse à accepter, chaque test calcule une valeur appelée statistique de test. Cette statistique est calculée à partir des propriétés de l'échantillon et doit donner des valeurs différentes dans le cas où l'hypothèse nulle est vraie et dans celui où elle est fausse. Par exemple, pour le test statistique de Jarque-Bera présenté à la section 2.4, la statistique JB est donnée par [2]

$$\mathrm{JB} = n\left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2-3)^2}{24}\right) \tag{1}$$

où $\sqrt{b_1}$ est le coefficient d'asymétrie de l'échantillon, b_2 est son coefficient d'aplatissement et n est le nombre d'éléments dans l'échantillon. Cette statistique prend donc une valeur de 0 lorsque l'échantillon prend une forme normale et augmente dans le cas contraire.

La distribution de probabilité théorique de cette statistique est ensuite obtenue dans le cas où l'hypothèse nulle est vraie. Autrement dit, la distribution de probabilités d'obtenir chaque valeur de la statistique dans le cas où un échantillon serait prélevé d'une population qui satisfait l'hypothèse nulle. Dans l'exemple du test de Jarque-Bera, pour de grands échantillons, cette distribution prend la forme d'une loi de χ^2 à deux degrés de liberté [3]. Pour des plus petits échantillons (Matlab prend 2000 comme valeur limite [4]), la distribution de la statistique est obtenue par simulation de Monte-Carlo.

Enfin, la valeur p est calculée. Cette valeur est la probabilité d'obtenir, dans le cas où l'hypothèse nulle est vraie, une statistique de test aussi ou plus extrême que celle calculée à partir de l'échantillon. Autrement dit, il s'agit d'intégrer la distribution théorique à partir de la valeur de la statistique de test calculée jusqu'à l'infini. Il est important de mentionner que dans le cas d'un test bilatéral, c'est-à-dire lorsque la distribution possède un axe de symétrie, la valeur p est calculée en intégrant les deux côtés de la distribution [7]. Cette valeur p est ensuite comparée à un seuil de signification α (très souvent, la valeur de 0,05 est utilisée) et, si la valeur p est inférieure à ce seuil, l'hypothèse nulle est rejetée.

2 Tests de normalité

Avant de comparer des échantillons en utilisant des tests d'hypothèse, il est important de soulever les tests de normalité. Dans la plupart des tests d'hyptohèse, la normalité de la population est une condition nécessaire [2]. Il est donc important de savoir si la distribution des échantillons l'est bel et bien. Nombreux sont les tests de normalité disponibles pour accomplir ceci. Cette section est dédiée à en présenter quelques-uns.

2.1 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test non paramétrique qui peut être utilisé pour tester la normalité. Ce test utilise le concept de fonction de répartition empirique (FRE). Cette fonction est donnée par [2]

$$\mathrm{FRE}_n(x) = \frac{\text{Nombre d'observations} < x}{n}$$

où n est la taille de l'échantillon. Cette fonction est ensuite comparée à la fonction de répartition théorique (FRT) d'une distribution normale et la statistique de test est calculée. Cette statistique est donnée par la plus grande distance entre la FRE et la FRT [9] :

$$D = \max |\mathrm{FRE}_n(x) - \mathrm{FRT}_n(x)| \tag{2}$$

La valeur p est finalement obtenue en intégrant la distribution théorique, qui est une distribution de Kolmogorov-Smirnov, ou à l'aide d'une table de valeurs critiques. Il est important de noter que cette distribution de la statistique ne dépend pas du type de fonction testée [10]. Ainsi, ce test pourrait être utilisé pour savoir, par exemple, si un échantillon provient d'une population exponentielle décroissante et la distribution de Kolmogorov-Smirnov serait toujours valide.

Il est important de noter que ce test perd énormément de son efficacité lorsque la même valeur est présente plusieurs fois dans l'échantillon. De plus, les paramètres de la population doivent être connus afin de trouver la FRT. Dans le cas où ces paramètres sont estimés, ce test devient conservateur. Si les paramètres sont inconnus, il est alors plus judicieux d'utiliser le test de Lilliefors (section 2.2) [8].

2.2 Test de Lilliefors

Le test de Lilliefors est une variante du test de Kolmogorov-Smirnov utilisée lorsque les paramètres de la population sont inconnus. Ceux-ci sont alors estimés à partir de l'échantillon. Par la suite, la démarche est la même que pour le test de Kolmogorov-Smirnov et la statistique de test est également la même. La seule différence est que la distribution théorique de cette statistique change, et suit plutôt la distribution de Lilliefors (ou la table de valeurs critiques de Lilliefors) [8].

2.3 Test d'Anderson-Darling

Le test de normalité d'Anderson-Darling est une autre modification du test de Kolmogorov-Smirnov. Toutefois, le test d'Anderson-Darling accorde plus de poids aux extrémités de la distribution [10]. Toutefois, le désavantage de ce test est que la distribution théorique de la statistique de test dépend du type de fonction testée. Ainsi, il faut la calculer (ou calculer les valeurs critiques) pour chaque famille de fonction (par exemple, fonctions normales, log-normales, de Weibull, exponentielles, etc.). La statistique du test d'Anderson-Darling est la suivante [10] :

$$A^2 = -n - \sum_{i=1}^{n} \frac{(2i-1)}{n} \left[\ln F(Y_i) + \ln(1 - F(Y_{n+1-i}))\right]$$

où n est la taille de l'échantillon, Y_i sont les données ordonnées de l'échantillon et $F(Y_i)$ est la fonction de répartition théorique du type de fonction testée.

2.4 Test de Jarque-Bera

Le test de Jarque-Bera utilise des concepts différents des autres tests de normalité. En effet, il considère plutôt le coefficient d'asymétrie $\sqrt{b_1}$ (*skewness*) et le coefficient d'aplatissement b_2 (*kurtosis*). Ces deux coefficients sont données par les troisième et quatrième moments centraux de l'échantillon, respectivement [11] :

$$\sqrt{b_1} = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}}$$

$$b_2 = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{2}}$$

où μ est le moment central, σ est la variance, n est la taille de l'échantillon et $\overline{x}$ est la moyenne de l'échantillon.

La statistique de Jarque-Bera est donnée à l'équation 1 de la section 1 et, tel qu'il l'est indiqué à cette section, cette statistique vaut 0 pour une fonction normale. Pour de grands échantillons, la distribution suit une loi de χ^2 à deux degrés de liberté. Toutefois, pour de plus petits échantillons, cette approximation n'est plus valide et la distribution doit être obtenue par simulations.

2.5 Test de Shapiro-Wilk

Le test de Shapiro-Wilk est dans une catégorie différente des autres tests, soit celle des tests de régression et de corrélation [2]. Celui-ci utilise les quantiles pour juger de la normalité et sa statistique est plus complexe à calculer puisqu'elle implique des vecteurs et des matrices. Cette statistique se calcule de la façon suivante [12] :

$$W = \frac{\left[\sum_{i=1}^{[n/2]} a_i (x_{n-i+1} - x_i)\right]^2}{\sum_i (x_i - \overline{x})^2}$$

où x_i sont les valeurs ordonnées de l'échantillon, $[n/2]$ est la partie entière de la taille de l'échantillon divisée par 2 et a_i sont des constantes obtenues avec la moyenne et la matrice de co-variance des quantiles d'un échantillon de la même taille que l'échantillon testé, mais ayant une distribution normale. Ce test est reconnu pour être efficace pour de petits échantillons ($n \leq 50$) [12].

3 Comparaisons de deux échantillons indépendants

Lorsque deux échantillons doivent être comparés afin d'avoir une idée de si ceux-ci proviennent du même type de population, il est important de faire la distinction entre des échantillons dépendants et indépendants. Cette section traite du deuxième cas, c'est-à-dire lorsque les échantillons sont indépendants. Cette condition est obtenue lorsque le coefficient de corrélation de Pearson des deux échantillons est nul [13]. Autrement dit, il ne doit pas y avoir de lien possible entre les deux échantillons. La meilleure façon d'obtenir l'indépendance est de prendre deux échantillons aléatoires. À l'inverse, un cas ou il y a une dépendance entre les échantillons pourrait être, par exemple, lorsque le même échantillon est observé avant et après un traitement quelconque.

Plusieurs tests de comparaison nécessitent l'indépendance des données. Trois tests seront détaillés ici, soient le test t de Student, le test de Kolmogorov-Smirnov à deux échantillons ainsi que le test de la somme des rangs de Wilcoxon.

3.1 Test t à deux échantillons

Les tests t de Student sont une famille de tests statistiques qui permettent de comparer la moyenne d'un échantillon à une valeur connue, les moyennes de deux échantillons indépendants ou encore la moyenne de deux échantillons dépendants (voir plutôt la section 4.1 pour ce dernier cas). Le test t à deux échantillons présenté ici est pour le cas d'échantillons indépendants. Ce test requiert également la normalité.

La statistique du test t à deux échantillons est la suivante [14] :

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \tag{3}$$

où $\overline{Y}_1$ et $\overline{Y}_2$ sont les moyennes des échantillons, s_1^2 et s_2^2 sont leur variance et N_1 et N_2 sont leur taille. On remarque donc que plus les moyennes sont proches, plus la statistique tend vers 0 (les moyennes sont égales) et qu'il en est de même lorsque les variances augmentent (il est plus difficile de déterminer avec précision les moyennes). Les valeurs critiques sont ensuite calculées à l'aide de la distribution de Student, aussi appelée distribution t [14]. Il est également important de noter que ce test tient plus ou moins compte de la forme des distributions (à part pour la variance) et seulement une comparaison des moyennes est effectuée.

3.2 Test de Kolmogorov-Smirnov à deux échantillons

Le test de Kolmogorov-Smirnov à deux échantillons est très semblable au test du même nom à un seul échantillon. La statistique de test est la même (équation 2), mais plutôt que de comparer la fonction de répartition empirique à une fonction de répartition théorique, la fonction de répartition empirique du premier échantillon est comparée à celle du deuxième échantillon. La distance maximale entre ces fonctions est obtenue et est comparée aux mêmes valeurs critiques que pour le test à un échantillon [15]. Les mêmes valeurs critiques peuvent être utilisées puisque la distribution de la statistique de test ne dépend pas de la forme des fonctions ou des échantillons étudiés.

3.3 Test de la somme des rangs de Wilcoxon

Si le test t est un test de moyenne, le test de la somme des rangs de Wilcoxon, aussi appelé test de Mann-Whitney, est plutôt un test de médiane, puisqu'il tient compte de l'ordre des données. La première étape de ce test est de placer toutes les données (les deux échantillons confondus) en ordre croissant et d'accorder un rang à chaque valeur (la plus petite valeur à un rang de 1, la deuxième un rang de 2, etc.). Par la suite, les échantillons sont de nouveaux séparés et la somme des rangs est effectuée pour chaque échantillon. La statistique de test est ensuite calculée. Celle-ci correspond à la plus petite valeur parmi les deux valeurs de U suivantes [16] :

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

où les n sont les tailles des échantillons et les R sont la somme des rangs pour chaque échantillon. Cette valeur est ensuite comparée à des tables afin de déterminer si l'hypothèse nulle est rejetée ou non.

Un problème de ce test concerne les petits échantillons. En raison de la nature finie des échantillons étudiés, le nombre de valeurs possibles pour U est également fini. Ainsi, pour de petits échantillons, peu de valeurs de U sont possibles et les écarts entre ces valeurs deviennent important. De la précision est donc perdue dans ces cas [17].

4 Comparaison de deux échantillons dépendants

Dans le cas ou les échantillons sont dépendants, par exemple si le même échantillon est étudié plusieurs fois, les tests présentés à la section 3 ne peuvent plus être utilisé et il devient nécessaire d'en trouver de plus adéquats. Plus précisément, on s'intéresse ici aux échantillons appariés, c'est-à-dire lorsqu'on peut relier les deux échantillons en paires de valeurs (par exemple, avant et après traitement). Deux tests seront présentés dans ce rapport : le test t de Student pour échantillons appariés et le test des rangs signés de Wilcoxon.

4.1 Test t pour échantillons appariés

Lorsque le même échantillon est mesuré deux fois et que l'on veut savoir si un changement s'est produit entre les mesures, le test t pour échantillons appariés peut être utilisé. La première étape ici est de calculer la différence entre les valeurs des deux mesures et ainsi d'obtenir un nouvel "échantillon" de même taille. Par la suite, un test t à un seul échantillon est utilisé pour déterminer si la moyenne de ce nouvel échantillon est 0 (dans un tel cas, aucun changement n'est décelé). La statistique d'un test t à un échantillon est très similaire à celle du test à deux échantillons (équation 3) [18] :

$$t = \frac{\mu}{s/n}$$

où μ est la moyenne de la différence entre les échantillons, s est son écart type et n est sa taille. Cette statistique est comparée à la distribution de Student pour trouver les valeurs critiques. De plus, comme tous les tests t, celui-ci requiert également la normalité.

4.2 Test des rangs signés de Wilcoxon

Le test des rangs signés de Wilcoxon est un test similaire au test de la somme des rangs de Wilcoxon puisque lui aussi utilise l'ordre des valeurs des échantillons. La première étape de ce test est la même que pour le test t pour échantillons appariés (section 4.1), c'est-à-dire qu'on prend la différence, valeur par valeur, entre les deux échantillons. Toutefois, pour ce test, on en prend la valeur absolue. Ensuite, on place les valeurs en ordre croissant. Pour chaque valeur, en fonction de si elle était positive ou négative avant la valeur absolue, son rang est multiplié par 1 ou -1, respectivement. Enfin, la statistique W est simplement la somme des rangs et sa distribution est normale pour des de grands échantillons ($n > 10$) [19]. Ce test étant médian, il ne nécessite pas la normalité.

5 Tests statistiques par ordinateur

Plutôt que de calculer les valeurs p à la main, les tests sont souvent effectués par ordinateur. Plusieurs langages de programmation permettent de les effectuer (Matlab, Python, R, etc.). Des fonctions Matlab pour effectuer divers tests statistiques sont disponible sur le Github de DCC-Lab dans le repository CodeAutofluorescence2018.

Références

[1] Statistics How To, *Hypothesis Testing*, En ligne, http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/#WhatisHT

[2] Yap, B.H. et Sim, C.H., *Comparisons of various types of normality tests*, Journal of Statistical Computation and Simulation (2011).

[3] National Institute of Standards and Technology, *JARQUE BERA TEST*, En ligne, https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/jarqbera.htm

[4] Mathworks, *jbtest*, En ligne, https://www.mathworks.com/help/stats/jbtest.html

[5] Nornadiah, M.R. et Yap, B.H. *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests*, Journal of Statistical Modeling and Analytics Vol. 2 No. 1 (2011).

[6] Feigelson, E. et Babu, G.J., *Beware the Kolmogorov-Smirnov test !*, En ligne, https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test

[7] Investopedia, *Two-Tailed Test*, En ligne, https://www.investopedia.com/terms/t/two-tailed-test.asp

[8] Real Statistics Using Excel, *Lilliefors Test for Normality*, En ligne, http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/lilliefors-test-normality/

[9] Real Statistics Using Excel, *Kolmogorov-Smirnov Test for Normality*, En ligne, http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/

[10] Engineering Statistics Handbook, *Anderson-Darling Test*, En ligne, https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

[11] Wikipedia, *Jarque-Bera Test*, En ligne, https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

[12] Ricco Rakotomalala *Tests de normalité : Techniques empiriques et tests statistiques*, Université Lumière Lyon 2 (2011).

[13] Krzywinski, M. et Altman, N. *Comparing Samples - Part I*, Nature Methods Vol. 11 No. 3 (2014).

[14] Engineering Statistics Handbook, *Two-Sample t-Test for Equal Means*, En ligne, https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

[15] National Institute of Standards and Technology, *KOLMOGOROV SMIRNOV TWO SAMPLE*, En ligne, https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/ks2samp.htm

[16] Wayne W. Lamorte, *Mann Whitney U Test (Wilcoxon Rank Sum Test)*, En ligne, http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html

[17] Krzywinski, M. et Altman, N. *Comparing Samples - Part I*, Nature Methods Vol. 11 No. 5 (2014).

[18] Rosie Shier, *Paired t-tests*, En ligne, http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf

[19] Statistics Solutions, *How to Conduct the Wilcoxon Sign Test*, En ligne, http://www.statisticssolutions.com/how-to-conduct-the-wilcox-sign-test/