Notes de cours Biostatistiques – MIV (L3)
Tests paramétriques

M. Bailly-Bechet

Université Claude Bernard Lyon 1 – France

1 Variable et test du χ^2

1.1 Variable du χ^2

On définit une variable du χ^2 à k degrés de liberté (d.d.l) comme la somme de k carrés des tirages indépendants d'une loi normale centrée réduite. Mathématiquement, on a :

$$\chi^2_k = \sum_{i=1}^{k} x_i^2, \tag{1}$$

avec chaque x_i une réalisation d'une variable normale centrée réduite. Sachant cela, on pourrait calculer la densité de la loi du χ^2 à k d.d.l à partir de la densité de probabilité de la loi normale. Le calcul est complexe, mais la densité de probabilité de la loi du χ^2 à k degrés de liberté est :

$$p(X = x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{\frac{-x}{2}}. \tag{2}$$

La fonction $\Gamma(r)$ est une fonction mathématique qui généralise de manière continue la fonction factorielle $n! = n \times (n-1) \times (n-2) \times \ldots \times 1$. Une de ses définitions est :

$$\Gamma(r) = \int_0^\infty e^{-\lambda x} \lambda^r x^{r-1} dx. \tag{3}$$

Cette fonction est donc définie par une intégrale, et n'a pas de forme analytique. Elle conserve et généralise à l'ensemble des réels les propriétés de la fonction factorielle [1].

Concernant le χ^2, on peut aisément remarquer que $\chi^2_k + \chi^2_l = \chi^2_{k+l}$, tous les x_i étant indépendants. Ceci permet de retrouver les propriétés générales du χ^2 à partir de celles du χ^2 à 1 d.d.l. Pour celui-ci – qui est donc simplement le carré d'une variable normale – on peut montrer que $\mathbb{E}(\chi^2_1) = 1$ et $\mathbb{V}(\chi^2_1) = 2$. Pour l'espérance, on écrit que $x^2 e^{-\frac{x^2}{2}} = x \times x e^{-\frac{x^2}{2}}$ et on intègre par parties :

$$\mathbb{E}(\chi^2_1) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \tag{4}$$

$$= \left[-x \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \quad = 1 \text{ car on reconnaît une intégrale gaussienne} \tag{5}$$

La démonstration pour la variance suit la même logique, mais est laissée aux étudiants. On peut déduire des deux propriétés précédentes et de la linéarité de l'espérance que l'espérance d'une v.a. du χ^2à k d.d.l vaut k, et sa variance $2k$.

1.2 Fonction génératrice des moments

Nous avons besoin pour la suite de la fonction génératrice des moments (FGM) du χ^2. Le calcul est direct, et passe par l'emploi de la définition de la fonction Γ. On va calculer la FGM d'une loi du χ^2 à k degrés de liberté

1. On peut facilement remarquer, par intégration par parties en prenant $u = x^r$ et $v' = e^{-\lambda x}\lambda^{r+1}$, que $\Gamma(r+1) = r\Gamma(r)$, ou encore que si r est entier, $\Gamma(r) = (r-1)!$

en passant par la formulation $M_X(t) = \mathbb{E}(e^{tx})$. On a :

$$\mathbb{E}(e^{tx}) = \int_0^\infty e^{tx} \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{\frac{-x}{2}} dx \tag{6}$$

$$= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^\infty e^{-\left(\frac{1}{2}-t\right)x} x^{\left(\frac{k}{2}-1\right)} dx \tag{7}$$

$$= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^\infty e^{-\lambda x} x^{r-1} dx \text{ en prenant } \lambda = \frac{1}{2} - t \text{ et } r = \frac{k}{2}, \tag{8}$$

$$= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \frac{1}{\lambda^r} \Gamma(r) \text{ en remplacant } \lambda \text{ et } r \text{ par leurs valeurs,} \tag{9}$$

$$= \frac{\Gamma\left(\frac{k}{2}\right)}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \frac{1}{\left(\frac{1}{2}-t\right)^{\frac{k}{2}}} \tag{10}$$

$$= \frac{1}{(1-2t)^{\frac{k}{2}}} \tag{11}$$

$$= (1-2t)^{-\frac{k}{2}}. \tag{12}$$

On peut vérifier les propriétés précédentes (espérance, variance) à partir de cette FGM.

1.3 Test du χ^2

Le test du χ^2 existe en deux versions :

Le test de conformité où l'on vérifie si une table de contingence représente un tirage "conforme" à une loi de probabilité pour les différents évenements.

Le test d'indépendance où l'on vérifie que la table de contingence croisée de deux variables ou plus est bien le produit des tables de contingences de chaque variable, c-à-d que la valeur d'une variable n'influe pas sur le tirage de l'autre.

Dans les deux cas, la statistique à calculer est la même, et dépend de la comparaison des valeurs théoriques (T_i) et des valeurs observées (O_i) de la table de contingence, pour les n catégories possibles :

$$\chi^2_{obs} = \sum_{i=1}^{n} \frac{(O_i - T_i)^2}{T_i}, \tag{13}$$

que l'on va comparer à une valeur seuil du χ^2 à $n-1$ ddl.

Pour retrouver comment le test du χ^2 est relié aux variables du χ^2, on va prendre un cas simple : le test de conformité pour un tirage à deux urnes. On définit deux évenements complémentaires A et B de probabilité respectives p et $1-p$ (exemples : pièce à pile ou face, probabilité d'avoir un garçon à la naissance, ...). On observe une statistique de n tirages de ces évenements, composée de n_1 tirages de A et $n-n_1$ tirages de B. La question est : ce résultat est-il significativement différent de l'hypothèse nulle que l'on s'est fixée au départ, à savoir que les tirages se font avec des probabilités p et $1-p$? Dans le cas des naissances, on peut se poser la question de savoir si, dans une population particulière, le pourcentage observé de garçons à la naissance est le même que dans le reste de la population du pays, où la proportion de garçons vaut p.

La procédure habituelle pour vérifier cette hypothèse est la suivante :

1. Calculer la statistique observée du χ^2 : $\chi^2_{obs} = \frac{(n_1-np)^2}{np} + \frac{((n-n_1)-n(1-p))^2}{n(1-p)}$
2. Comparer sa valeur à celle de la loi du χ^2 à 1 d.d.l (1 ici car le nombre de d.d.l doit êtres égal au nombre de catégories possibles moins 1)
3. Conclure

Quel est le lien entre χ^2_{obs} et la loi du χ^2 ? Pourquoi $n-1$ d.d.l ? On peut répondre à ces questions par le calcul suivant :

$$\chi^2_{obs} = \frac{(n_1-np)^2}{np} + \frac{((n-n_1)-n(1-p))^2}{n(1-p)} \tag{14}$$

$$= \frac{(n_1-np)^2}{np} + \frac{(np-n_1)^2}{n(1-p)} \tag{15}$$

$$= \frac{(n_1-np)^2}{n}\left(\frac{1}{p} + \frac{1}{1-p}\right) \tag{16}$$

$$= \frac{(n_1-np)^2}{n}\left(\frac{1}{p(1-p)}\right) \tag{17}$$

$$= \left(\frac{n_1-np}{\sqrt{np(1-p)}}\right)^2 \tag{18}$$

Le terme de droite de la dernière ligne est une réalisation de n tirages qui a été centrée et réduite. En effet, le tirage de n éléments de probabilité p

suit une loi binomiale de moyenne np et d'écart-type $\sqrt{np(1-p)}$. Cette loi binomiale est elle-même la somme de n variables de Bernouilli ; si n est très grand, d'après le TCL on sait que l'on peut approximer cette loi binomiale exacte suivie par ce tirage par une loi normale de même espérance et variance. La variable $\frac{n_1-np}{\sqrt{np(1-p)}}$ est donc une variable normale centrée réduite ; mise au carré, c'est donc une variable du χ^2à 1 d.d.l. On obtient donc bien que, si le TCL est applicable (ici la seule condition étant que n soit assez grand), le χ^2_{obs} suit une loi du χ^2 à 1 d.d.l, ce qui explique pourquoi on la compare ensuite à cette loi.

On peut généraliser cette formule par récurrence à un test de conformité du χ^2 à k degrés de liberté, le raisonnement étant exactement le même ; la condition deviendra cependant que le nombre attendu de succès dans chaque catégorie soit suffisamment grand, pour que chacun des termes de la somme du χ^2 puisse être approximé par un terme normal centré réduit. Pour les tests d'indépendance, le principe est légèrement plus complexe, mais ne sera pas démontré ici.

2 Test de la moyenne

Il existe différents tests de la moyenne. Ceux-ci servent à comparer les moyennes de deux échantillons pour un paramètre quantitatif, ou à comparer la moyenne d'un échantillon à une valeur théorique. L'hypothèse nulle y est toujours : "Les deux échantillons ont même moyenne", ce que l'on interprète généralement par le fait que les deux échantillons viennent de la même population.

En fonction du fait que l'on connaisse ou non *a priori* la variance des populations de départ – cas en pratique très rare – le test se fait de différentes manières. On va les expliciter ici. Le cadre général est donc que l'on a deux échantillons, que l'on peut considérer comme deux séries de tirages de deux v.a. X et Y. On note respectivement μ_X et μ_Y les espérances des deux v.a. et σ_X^2 et σ_Y^2 leurs variances. On suppose que les tailles des échantillons sont n_X et n_Y. On note $\bar{x}$ et $\bar{y}$ les moyennes des deux échantillons. Le cas où l'on compare la moyenne d'un échantillon à une valeur théorique est très proche, et est laissé en exercice.

2.1 Variances connues

Si on connaît *a priori* les variances de X et Y, suite à des études populationelles précédentes, la construction du test est assez simple. On sait que la moyenne du premier échantillon $\bar{x}$ suit une loi qui vient directement de la loi de X, la loi de la population d'origine. On sait en effet que $\bar{x}$ suit une loi de même moyenne que X (par linéarité de l'espérance) et de variance $\frac{\sigma_X^2}{n_X}$, d'après les propriétés de la variance. Idem pour $\bar{y}$. Si X et Y sont indépendantes, on peut construire facilement la loi de la v.a. $\bar{x} - \bar{y}$: c'est une loi de moyenne $\mu_X - \mu_Y$ et de variance $\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$. La quantité

$$\epsilon = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \tag{19}$$

est donc une v.a. centrée réduite. Pour déterminer la loi suivie par cette v.a., nous avons plusieurs possibilités. Tout d'abord, si X et Y suivent chacun une loi normale, alors on peut montrer [2] que la loi de $x + y$, et a fortiori de $\bar{x} - \bar{y}$, est une loi normale. Si on ne connaît pas la loi suivie par X, Y, ou les deux, mais que l'on sait que n_X, n_Y ou les deux sont suffisament grands, on peut appliquer le TCL pour dire que $\bar{x}$, $\bar{y}$ ou les deux suivent une loi normale ; encore une fois dans ce cas la v.a. $\bar{x} - \bar{y}$ suivra elle aussi une loi normale. Donc pour conclure que ϵ suit une loi normale, il suffit que chaque échantillon ait une taille importante ou provienne d'une population normalement distribuée. Dans ces cas, ϵ suit une loi normale. Attention cependant, si les effectifs n_X et n_Y sont très faibles, il va être techniquement difficile de rejeter l'hypothèse de normalité avec un test de normalité (voir chapitres suivants). Dans le cas des petits effectifs, on se tournera donc en pratique vers des tests non paramétriques pour lesquelles la distribution des données de départ a moins d'importance.

Finalement, pour effectuer le test de comparaison de moyennes, on veut comparer l'hypothèse nulle H_0 $\mu_X = \mu_Y$ à l'hypothèse alternative $\mu_X \neq \mu_Y$. Si H_0 est vraie, alors on peut dire que la quantité

$$\epsilon_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \tag{20}$$

suit une loi normale. On va donc calculer cette quantité, et regarder la probabilité que le résultat obtenu (ou mieux) provienne d'une loi normale

2. Il s'agit de la stabilité par convolution de la loi normale

centrée réduite. Cette probabilité est une p-value, et on décidera en fonction de sa valeur d'accepter ou de rejeter l'hypothèse H_0. Notez bien que si H_0 est fausse, on ne peut plus rien dire sur notre statistique et sur la valeur attendue de ϵ_{obs} : on ne peut donc pas calculer de la même manière le risque de deuxième espèce.

2.2 Variances inconnues : test de Fischer et test de Student

Dans de nombreux cas pratiques, on veut comparer les moyennes d'échantillons dont la variance est inconnue, et donc estimée à partir des données. Cette estimation de la variance conduit à faire des erreurs supplémentaires, et la loi suivie par la différence centrée réduite des moyennes ne va plus être une loi normale.

Soit Z une v.a normale de moyenne μ et de variance σ^2. Si on a un échantillon de Z composé de n tirages $z_1, z_2 \ldots, z_n$, on sait que la moyenne observée $\bar{z}$, d'après le TCL, suit une loi normale de moyenne μ et de variance $\frac{\sigma^2}{n}$. Qu'en est-il de la variance observée, s^2 ? On a :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2, \text{ par définition ; donc} \tag{21}$$

$$\frac{ns^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{z_i - \bar{z}}{\sigma}\right)^2 \tag{22}$$

$$= \sum_{i=1}^{n}\left(\frac{z_i - \mu + \mu - \bar{z}}{\sigma}\right)^2 \tag{23}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((z_i - \mu)^2 + 2\,(z_i - \mu)\,(\mu - \bar{z}) + (\mu - \bar{z})^2\right) \tag{24}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2 + 2\sum_{i=1}^{n}\left(z_i\mu - z_i\bar{z} - \mu^2 + \mu\bar{z}\right) + \sum_{i=1}^{n}\left(\mu^2 - 2\mu\bar{z} + \bar{z}^2\right) \tag{25}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2 + 2n\bar{z}\mu - 2n\bar{z}^2 - 2n\mu^2 + 2n\mu\bar{z} + n\mu^2 - 2n\mu\bar{z} + n\bar{z}^2 \tag{26}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2 - n\mu^2 + 2n\mu\bar{z} - n\bar{z}^2 \tag{27}$$

$$= \sum_{i=1}^{n}\left(\frac{z_i - \mu}{\sigma}\right)^2 - n\left(\frac{\bar{z} - \mu}{\sigma}\right)^2 \tag{28}$$

$$= \sum_{i=1}^{n}\left(\frac{z_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{z} - \mu}{\sqrt{\frac{\sigma^2}{n}}}\right)^2 \tag{29}$$

Les deux termes suivent une loi du χ^2, respectivement à n d.d.l et 1 d.d.l. Cependant, la loi suivie par la différence de deux termes du χ^2 n'est pas connue, contrairement au cas additif ; de plus, les deux termes ici ne sont pas indépendants (ils dépendent tous deux des z_i plus ou moins directement), et on ne peut donc *a priori* pas appliquer les relations sur la linéarité de l'espérance ou d'additivité des variances. Une interprétation géométrique (le modèle linéaire généralisé) permettrait d'arrêter les calculs et de conclure, masi sans passer par cela on peut néanmoins finir la démonstration avec des calculs plus pédestres.

Comme dans le cas du TCL, on va donc passer par les FGM pour trouver la loi suivie par la variance observée s^2. Cette variance, comme la moyenne $\bar{z}$, est en effet une v.a. qui dépend du tirage des z_i. On réécrit l'équation 29 et on prend sa FGM :

$$\frac{ns^2}{\sigma^2} + \left(\frac{\bar{z}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right)^2 = \sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right)^2 \tag{30}$$

$$\mathbb{E}\left(\exp\left\{t\frac{ns^2}{\sigma^2} + t\left(\frac{\bar{z}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right)^2\right\}\right) = \mathbb{E}\left(\exp\left\{t\sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right)^2\right\}\right) \tag{31}$$

$$\mathbb{E}\left(\exp\left\{t\frac{ns^2}{\sigma^2}\right\}\exp\left\{t\left(\frac{\bar{z}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right)^2\right\}\right) = \mathbb{E}\left(\exp\left\{t\sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right)^2\right\}\right) \tag{32}$$

Par indépendance de s^2 et $\bar{z}$ (indépendance qui découle de la normalité de X), on peut utiliser l'égalité $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ sur le membre gauche de l'équation 32. Notez bien que l'on n'aurait pas pu faire cette manipulation si on avait laissé le terme en $\frac{\bar{z}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ à droite, car son indépendance du terme en $\sum_{i=1}^{n}$ est non triviale, contrairement à l'indépendance de s^2 et $\bar{z}$. On obtient alors, en reconnaissant les FGM de variables du χ^2 à n et 1 d.d.l :

$$\mathbb{E}\left(\exp\left\{t\frac{ns^2}{\sigma^2}\right\}\right)\mathbb{E}\left(\exp\left\{t\left(\frac{\bar{z}-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right)^2\right\}\right) = \mathbb{E}\left(\exp\left\{t\sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right)^2\right\}\right) \tag{33}$$

$$\mathbb{E}\left(\exp\left\{t\frac{ns^2}{\sigma^2}\right\}\right)(1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}} \tag{34}$$

$$\mathbb{E}\left(\exp\left\{t\frac{ns^2}{\sigma^2}\right\}\right) = (1-2t)^{-\frac{(n-1)}{2}} \tag{35}$$

La v.a. $\frac{ns^2}{\sigma^2}$ a donc pour FGM $(1-2t)^{-\frac{(n-1)}{2}}$, ce qui est la FGM d'une v.a. du χ^2 à $n-1$ d.d.l. On en déduit donc, par le théorème de Lévy, que $\frac{ns^2}{\sigma^2}$

suit une loi du χ^2 à $n-1$ d.d.l. Cette égalité nous sera très utile par la suite, pour retrouver les lois de Fisher et Student. De plus, on peut remarquer, en prenant l'espérance de $\frac{ns^2}{\sigma^2}$, que l'on a $\mathbb{E}(\frac{ns^2}{\sigma^2}) = n-1$, formule que l'on voit souvent réécrite de manière simplifiée comme $\hat{\sigma}^2 = \frac{n}{n-1}s^2$. Cette dernière formule nous donne la valeur de $\hat{\sigma}^2$, l'estimateur de σ^2, mais contient moins d'information, puisque l'on perd la notion de la loi suivie par s^2 dans ce cas.

2.2.1 Égalité des variances et test de Fisher

Reprenons les v.a. X et Y de la section 2.1 et les notations associées. Quand les variances sont estimées, il faut avant de comparer les moyennes de X et Y vérifier si leurs variances sont significativement différentes ou non. Le test associé à cette question est le test de Fisher, d'hypothèse nulle H_0 : "Les variances de X et Y sont égales". Pour effectuer ce test, on a besoin d'utiliser la loi de Fisher, qui est définie à partir de la loi du χ^2. En effet, on définit la loi de Fisher de paramètres a et b comme le rapport de lois du χ^2 normalisés :

$$F(a,b) = \frac{\frac{\chi_a^2}{a}}{\frac{\chi_b^2}{b}} \tag{36}$$

Ici encore, on pourrait calculer la densité de probabilité de la loi de Fisher à partir de celle du χ^2, mais ce ne sera pas nécessaire en pratique.

Sachant cela, on peut dire que le rapport des variances observées des v.a. X et Y va suivre une loi de Fisher. En effet on a :

$$\frac{\frac{n_X s_X^2}{\sigma_X^2(n_X-1)}}{\frac{n_Y s_Y^2}{\sigma_Y^2(n_Y-1)}} \to F(n_X-1, n_Y-1), \tag{37}$$

ou plus simplement

$$\frac{\frac{\hat{\sigma}_X^2}{\sigma_X^2}}{\frac{\hat{\sigma_Y}^2}{\sigma_Y^2}} \to F(n_X-1, n_Y-1). \tag{38}$$

Sous l'hypothèse H_0, on a $\sigma_X^2 = \sigma_Y^2$ et alors $\frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \to F(n_X-1, n_Y-1)$. C'est cette statistique que l'on va calculer, puisque l'on ne dispose pas de σ_X^2 ou σ_Y^2. Si H_0 est vraie, on peut estimer la probabilité des déviations de la valeur observée de F par rapport à la loi de Fisher. Si cette déviation est

trop grande ou trop faible, on rejettera H_0 et concluera que les variances sont différentes [3].

2.2.2 Construction du test de Student

Si les deux variances sont égales, on définit la variance commune estimée par la formule

$$\hat{\sigma}^2 = \frac{n_X s_X^2 + n_Y s_Y^2}{n_X + n_Y - 2}. \tag{39}$$

L'espérance de $\hat{\sigma}^2$ vaut 1 ; en effet, de la même manière que plus haut dans le cas d'une seule variance, on peut également démontrer que la variance commune observée de deux variables suit une loi du χ^2 à $n_X + n_Y - 2$ d.d.l ; plus précisément on a :

$$\frac{n_X s_X^2 + n_Y s_Y^2}{\sigma^2} \to \chi^2_{n_X+n_Y-2} \ . \tag{40}$$

La démonstration est la même que plus haut et passe par les FGM. La combinaison de ces deux équations nous donne :

$$\frac{\hat{\sigma}^2}{\sigma^2} (n_X + n_Y - 2) \to \chi^2_{n_X+n_Y-2} \ . \tag{41}$$

Si on reprend maintenant le raisonnement vu pour le cas des variances connues, toujours en supposant que X et Y suivent une loi normale (ici cette hypothèse est nécessaire et ne peut pas être remplacée par de grands échantillons comme plus haut, notamment pour pouvoir montrer les convergences des variances observées vers des lois de χ^2), on va pouvoir écrire que :

$$\frac{\bar{x} - \bar{y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \to \mathcal{N}(0,1) \tag{42}$$

Le problème vient du fait que l'on ne dispose que de $\hat{\sigma}^2$ et pas de σ^2. On réécrit alors l'équation précédente sous la forme :

3. En pratique, on choisit souvent de faire un test de Fisher unilatéral en divisant la variance la plus grande par la plus faible, et en ne regardant que les grandes valeurs de F comme significatives.

$$\frac{\bar{x} - \bar{y} - (\mu_X - \mu_Y)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \to \mathcal{N}(0,1) \tag{43}$$

En remplaçant le terme sous la racine par l'équation 41, et en supposant H_0 vraie (donc en supprimant le terme $\mu_X - \mu_Y$), on obtient la convergence en loi suivante :

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \to \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n_X+n_Y-2}}{n_X+n_Y-2}}} \tag{44}$$

La statistique de t_{obs} suit donc celle du rapport d'une loi normale centrée réduite et de la racine d'un χ^2 normalisé. C'est ainsi que l'on *définit* la loi de Student à $n_X + n_Y - 2$ d.d.l. Ici encore, il serait possible de trouver une expression analytique pour la densité de probabilité de la loi de Student, mais en pratique elle est inutile, et soit le calcul direct sur ordinateur, soit les tables statistiques sont utilisées. Comme précedemment, si H_0 est fausse, on ne peut rien dire en termes statistiques, et notamment on ne peut pas estimer l'erreur faite si on accepte H_0 à partir des calculs précédents.

Si les deux variances sont inégales, les choses deviennent plus complexes. En effet, on ne peut pas dans ce cas là calculer une statistique exacte, se ramenant à une distribution connue comme celle de Student. Dans la pratique, on calculera

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma_X}^2}{n_X} + \frac{\hat{\sigma_Y}^2}{n_Y}}} \tag{45}$$

et on fera l'hypothèse que, si H_0 est vraie, cette quantité suit une loi de Student. Comme l'on sait que cette hypothèse est partiellement inexacte, on la corrige en disant que le nombre de d.d.l n'est pas $n_X + n_Y - 2$ comme précedemment, mais plutôt par

$$\nu = \frac{\left(\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}\right)^2}{\frac{s_X^4}{N_X^2 \cdot (N_X - 1)} + \frac{s_Y^4}{N_Y^2 \cdot (N_Y - 1)}} \tag{46}$$

Cette correction permet de se rapprocher le plus possible de la distribution de Student pour la quantité calculée. On parle alors de test du t de Welch (employé par défaut sous R quand un `t.test` est appelé sans l'option `var.equal=TRUE`).