ECOLE DES MINES DE NANCY

Décision et Prévision Statistiques

Thierry Verdel (2007)

Chapitre 5

Comparaisons statistiques

Nous présentons dans ce chapitre un raisonnement nouveau. Son inventeur, au début de ce siècle, avait pris le pseudonyme de Student. Le problème qui lui était posé était le suivant : l'engrais a-t-il une influence sur le rendement des cultures de pomme de terre ? Pour le résoudre, Student imagine de choisir 4 parcelles. Chacune d'elles est divisée en deux, et on la cultive en traitant l'une des moitiés choisie au hasard, avec de l'engrais et l'autre non. Après la récolte, on calcule les rendements et, pour une parcelle donnée, la différence de rendements entre les deux moitiés avec engrais et sans engrais. Les 4 différences obtenues sont : {11, 30, -6, 13}. Student convient de considérer ces valeurs comme des réalisations d'une variable aléatoire D. Il fait alors l'hypothèse que l'engrais n'a pas d'influence. Si cette hypothèse est vraie, la moyenne E(D) de la variable D est nulle. La démarche se poursuit par une sorte de raisonnement par l'absurde, en vérifiant si les valeurs observées peuvent être considérées comme compatibles ou non avec E(D) = 0. Si elles sont incompatibles, l'hypothèse faite doit être remise en cause, et l'on peut conclure à l'influence de l'engrais ... Ce raisonnement, théorisé plus tard par Neyman et Pearson, est appelé le test d'hypothèse.

---

1. Tests d'hypothèse

1.1. Théorie de Neyman et Pearson

On suppose donnée une certaine variable aléatoire X dont la loi de probabilité dépend des hypothèses que l'on désire tester. Plus précisément, on suppose qu'il existe plusieurs hypothèses H_0, H_1, ..., H_n parfaitement connues (qui peuvent être en nombre fini ou non, dénombrable ou non) et que la loi de probabilité dépend de l'hypothèse vraie. Le test va permettre de porter un jugement sur l'hypothèse faite et d'évaluer le degré de validité du jugement, cela à partir de la valeur prise par X.

Nous étudierons d'abord le cas où l'on fait deux hypothèses simples H_0 et H_1. Une hypothèse est dite simple si elle définit complètement et d'une manière unique la loi de probabilité de X ; sinon, elle est dite composite. C'est ainsi, par exemple, qu'en présence d'un lot de pièces distinguées en " convenables " et " défectueuses ", les deux hypothèses :

H_0 : le lot contient 5 % de déchets

H_1 : le lot contient 10 % de déchets

sont des hypothèses simples puisque chacune d'elles définit entièrement le lot. Tandis que les deux hypothèses :

H_0 : le lot contient 5 % ou moins de 5 % de déchets

H_1 : le lot contient plus de 5 % de déchets

sont des hypothèses composites puisque ni l'une ni l'autre ne définit entièrement le lot.

Supposons donc qu'il existe deux hypothèses simples H_0 et H_1 couvrant l'ensemble des possibilités ; cela veut dire que l'une ou l'autre des deux hypothèses H_0 et H_1 est réalisée nécessairement. Dans ce cas, il est possible d'émettre l'un des deux jugements :

H_0 est vraie, donc H_1 est fausse,

H_1 est vraie, donc H_0 est fausse.

On peut symboliser cet ensemble par le tableau ci-dessous où figurent en lignes les états possibles et en colonnes les jugements portés. Le tableau contient les conséquences des différentes combinaisons.

Etat réalisé Jugement porté	H_0 est réalisé	H_1 est réalisé
H_0 est vraie	jugement correct	jugement faux
H_1 est vraie	jugement faux	jugement correct

Parmi les deux hypothèses H_0 et H_1, il en existe en général une dont le rejet à tort a des conséquences plus fâcheuses que pour l'autre. Il est donc normal de ne pas traiter H_0 et H_1 de

façon symétrique. Admettant alors que H_0 représente une circonstance favorable et H_1 une circonstance défavorable, on peut se tromper de deux manières :

- en considérant comme défavorable ce qui est favorable ; c'est l'*erreur de première espèce* ;

- en considérant comme favorable ce qui ne l'est pas ; c'est l'*erreur de deuxième espèce.*

C'est exactement en ces termes que se posait le problème du contrôle de réception, où ces deux types d'erreur correspondaient à des préoccupations toutes différentes : celle du fournisseur d'une part, et celle du client d'autre part.

Pour relier maintenant le jugement porté à l'observation de la variable X, on opère ainsi :

- on dit que H_0 est vraie si la valeur observée de X, soit x, se trouve dans un certain domaine w, appelé région d'acceptation de l'hypothèse H_0 ;

- on dit que H_1 est vraie si la valeur observée n'appartient pas à w.

Pour choisir le domaine w, on impose en général deux conditions :

- que la probabilité de commettre l'erreur de première espèce soit égale à un seuil déterminé α choisi a priori aussi faible qu'on le veut ;

- que la probabilité β de commettre l'erreur de deuxième espèce soit minimale.

Il importe de noter en effet que la première condition ne suffit pas, sauf cas très particulier, à définir w de façon unique.

Il est possible maintenant de compléter le tableau précédent en indiquant les règles de jugement et les probabilités pour qu'il soit correct ou faux :

Etat réalisé / Jugement porté	H_0 est réalisé	H_1 est réalisé
$X \in w$ H_0 est vraie	$1-\alpha$ jugement correct	β jugement faux
$X \notin w$ H_1 est vraie	α jugement faux	$1-\beta$ jugement correct

Un tel mode de raisonnement est appelé test d'hypothèse. Le complément à l'unité de β, soit $(1 - \beta)$ est appelé puissance du test : un test est d'autant plus puissant, pour un risque de première espèce fixé, que le risque de deuxième espèce est plus petit.

1.2. Détermination de la région d'acceptation

Si l'on note $p_0(x/H_0)$ et $p_1(x/H_1)$ les densités de probabilité de X, respectivement dans le cadre des hypothèses H_0 et H_1, les deux conditions précédentes s'expriment par les deux équations suivantes :

$$\int_w p_0(x)\,dx = 1-\alpha$$

$$\int_w p_1(x)\,dx = \beta \text{ minimum}$$

On démontre qu'elles sont satisfaites s'il existe une constante positive λ, telle que pour x appartenant à w :

$$p_1(x) < \lambda\, p_0(x) \qquad (1)$$

sous la contrainte :

$$\int_w p_0(x)\, dx = 1 - \alpha \qquad (2)$$

La démonstration qui suit n'est pas essentielle.

Supposons qu'une telle constante λ existe et considérons la quantité :

$$F(w) = \int_w p_1(x)\, dx - \lambda \int_w p_0(x)\, dx.$$

En appelant $I_w(x)$ la fonction indicatrice du domaine w, qui prend la valeur 1 *si x appartient à w et la valeur* 0 *sinon, on peut écrire $F(w)$ sous la forme :*

$$F(w) = \int I_w(x)\,(p_1(x) - \lambda\, p_0(x))\, dx.$$

On constate que $F(w)$ est négatif donc minimum pour :

$$I_w(x) = \begin{matrix} 0 & \text{si } p_1(x) - \lambda\, p_0(x) \geqslant 0 \\ 1 & \text{si } p_1(x) - \lambda\, p_0(x) < 0 \end{matrix}$$

Or, lorsque $F(w)$ est minimum sous la condition (2), la quantité $\int_w p_1(x)\, dx$, c'est-à-dire β, l'est évidemment aussi.

Appliquons ce résultat à deux exemples.

1.3. Test sur une proportion

Supposons qu'ayant prélevé un échantillon de n pièces dans un certain lot, on veuille tester l'hypothèse :

H_0 : la proportion de déchets est ϖ_0, contre l'hypothèse :

H_1 : la proportion de déchets est ϖ_1.

Le nombre de déchets dans l'échantillon est une variable aléatoire définie par les probabilités $p_0(k)$ si H_0 est vraie et $p_1(k)$ si c'est H_1 :

$$p_0(k) = C_n^k\, {\varpi_0}^k (1 - \varpi_0)^{n-k},$$
$$p_1(k) = C_n^k\, {\varpi_1}^k (1 - \varpi_1)^{n-k}.$$

La condition (1) s'écrit :

$$C_n^k\, {\varpi_1}^k (1 - \varpi_1)^{n-k} < \lambda\, C_n^k\, {\varpi_0}^k (1 - \varpi_0)^{n-k}.$$

Et, après simplification et passage aux logarithmes, on obtient :

$$k \log\left(\frac{\varpi_0}{\varpi_1}\right) + (n-k) \log\left(\frac{1-\varpi_0}{1-\varpi_1}\right) + \log(\lambda) > 0,$$

soit, pour $\varpi_1 > \varpi_0$:

$$k < \frac{n \log\left(\frac{1-\varpi_1}{1-\varpi_0}\right) - \log(\lambda)}{\log\left(\frac{\varpi_1}{\varpi_0}\right) - \log\left(\frac{1-\varpi_1}{1-\varpi_0}\right)} = k_s.$$

L'inégalité se réduit donc à $k < k_s$.

Pour déterminer k_s, il suffit d'utiliser la condition (2) qui s'écrit :

$$\sum_{k=0}^{k_s} C_n^k \, \varpi_0{}^k (1-\varpi_0)^{n-k} = 1-\alpha.$$

On notera que la région d'acceptation ne dépend pas de la valeur ϖ_1, c'est-à-dire de l'hypothèse H_1. Par contre, le risque de deuxième espèce en dépend puisque :

$$\beta = \sum_{k=0}^{k_s} C_n^k \, \varpi_1{}^k (1-\varpi_1)^{n-k}.$$

1.4. Test sur une moyenne

Soit un échantillon de taille n prélevé dans une population normale d'écart-type σ connu, mais de moyenne μ inconnue. Considérons les hypothèses :

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu = \mu_1$$

La région d'acceptation est définie par :

$$\frac{1}{(2\pi)^{n/2}\,\sigma^n} \, e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu_1)^2}{\sigma^2}} < \frac{\lambda}{(2\pi)^{n/2}\,\sigma^n} \, e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu_0)^2}{\sigma^2}},$$

expression que l'on peut écrire aussi :

$$\sum_{i=1}^{n} (x_i-\mu_0)^2 - \sum_{i=1}^{n} (x_i-\mu_1)^2 < 2\,\sigma^2 \log(\lambda),$$

soit, en notant m la moyenne empirique $m = \frac{\sum_{i=1}^{n} x_i}{n}$ et en supposant que $\mu_1 > \mu_0$:

$$m < \frac{\mu_0+\mu_1}{2} + \frac{\sigma^2 \log(\lambda)}{n(\mu_1-\mu_0)} = m_s.$$

Pour définir m_s, il suffit alors d'écrire que :

$$\text{Prob}\ \{M_n > m_s \,/\, \mu = \mu_0\} = \alpha,$$

où M_n désigne la variable aléatoire moyenne d'un échantillon de taille n. Remarquons que, dans ce deuxième exemple aussi, la région d'acceptation ne dépend pas de l'hypothèse H_1.

1.5. Cas d'hypothèses composites

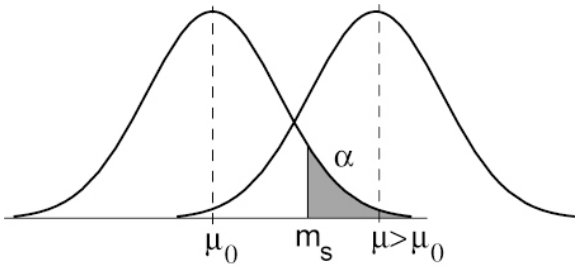
En réalité, très souvent, le problème n'est pas de choisir entre deux hypothèses simples H_0 et H_1, mais entre une hypothèse simple H_0 et un ensemble plus ou moins vaste d'hypothèses H_1, ..., H_i, ..., H_n, ou même à un ensemble continu d'hypothèses H.

Dans ce cas, on peut se ramener au problème précédent en comparant successivement H_0 à chacune des hypothèses de l'ensemble H. Si, par exemple, on compare H_0 à H_i, la méthode exposée plus haut permet de trouver une région w_i telle que le risque de première espèce soit égal à α et que le risque de deuxième espèce β_i soit minimum. On obtient ainsi un ensemble de régions d'acceptation w_1, ..., w_i, ..., w_n et, dans le cas général, on ne peut pas aller plus loin.

Mais il existe un cas particulier très intéressant, celui où les différentes régions w_i ont une partie commune w. Dans ce domaine w, le test utilisé est dit uniformément le plus puissant (en abréviation de l'anglais : UMP). En effet, lorsque X tombe dans w, on est sûr que le risque de première espèce est égal à α et que le risque de deuxième espèce est minimum, quelle que soit

l'hypothèse H vérifiée. Les deux exemples précédents constituent une illustration de ce cas, la région d'acceptation étant, comme nous l'avons souligné, indépendante de l'hypothèse H_1. Pas tout à fait cependant.

Notons, en effet, que nous avons supposé, respectivement dans chacun des deux exemples, que $\varpi_1 > \varpi_0$ et que $\mu_1 > \mu_0$. Et nous avons abouti alors à des régions d'acceptation de la forme $k < k_s$ et $m < m_s$ telles que le risque α soit *bloqué* à l'une des extrêmités de la distribution de la variable étudiée.

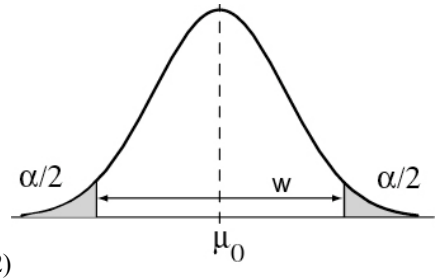


(5.1)

Si donc il s'agit de comparer deux hypothèses de la forme : $H_0 : \theta = \theta_0$ et $H_1 : \theta > \theta_0$, on est conduit à ce qu'on appelle un test à droite, où le risque de première espèce est bloqué à droite.

Le test d'hypothèses de la forme $H_0 : \theta = \theta_0$ et $H_1 : \theta < \theta_0$, conduit à un test appelé test à gauche .

Dans le cas, enfin, d'hypothèses de la forme $H_0 : \theta = \theta_0$ et $H_1 : \theta \neq \theta_0$, il apparait logique de répartir le risque α aux deux extrêmités de la distribution. Le test est alors un test symétrique.



(5.2)

2. Tests usuels de comparaison à un standard

2.1. Rappel des lois outils usuelles

La détermination des régions d'acceptation nécessite la mise en oeuvre des lois de probabilité caractéristiques des échantillons prélevés dans des populations de référence spécifiées. D'où l'extrême importance d'une connaissance précise des lois de probabilité usuelles définies dans le chapitre précédent, mais que nous allons reprendre ici.

2.1.1. Loi normale réduite

Etant donnée une variable qui suit une loi normale de moyenne μ et d'écart-type σ, la variable :

$$U = \frac{X-\mu}{\sigma}$$

est distribuée suivant une loi normale réduite (moyenne nulle et écart-type égal à 1).

Etant donnée la variable $M_n = \frac{\sum_{i=1}^{n} X_i}{n}$, moyenne d'un échantillon de taille n prélevé dans une population normale $\mathcal{N}(\mu, \sigma)$, elle suit une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Il en résulte que la variable : $\frac{M_n - \mu}{\sigma/\sqrt{n}}$ suit une loi normale réduite.

2.1.2. Loi du χ^2

Etant données ν variables $U_1, U_2, \ldots, U_\nu$ indépendantes et suivant des lois normales réduites, la variable :

$$\chi^2_\nu = {U_1}^2 + {U_2}^2 + \ldots + {U_\nu}^2$$

suit une loi du χ^2 à ν degrés de liberté.

Il en résulte qu'étant donné un échantillon $(X_1, \ldots, X_i, \ldots, X_n)$, prélevé dans une population normale $\mathcal{N}(\mu, \sigma)$, la variable :

$$\chi^2_n = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

suit une loi du χ^2 à n degrés de liberté.

Appelant $S^2 = \frac{\sum_{i=1}^{n} (X_i - M)^2}{n}$ la variance de l'échantillon, la variable :

$$\chi^2_{n-1} = \frac{\sum_{i=1}^{n} (X_i - M)^2}{\sigma^2} = \frac{n S^2}{\sigma^2}$$

suit une loi du χ^2 à $(n - 1)$ degrés de liberté.

2.1.3. Loi de Student

Etant données $(\nu + 1)$ variables normales, réduites, indépendantes, la variable :

$$T_\nu = \frac{U}{\sqrt{\frac{\sum_{i=1}^{\nu} {U_i}^2}{\nu}}}$$

suit une loi de Student à ν degrés de liberté.

Il en résulte qu'étant données M et S^2 la moyenne et la variance d'un échantillon de taille n prélevé dans une population normale $\mathcal{N}(\mu, \sigma)$, la variable :

$$T_{n-1} = \frac{M - \mu}{\sqrt{\frac{\frac{n}{n-1} S^2}{n}}}$$

(où $\frac{n}{n-1} S^2$ est l'estimateur sans biais de σ^2) suit une loi de Student à $(n - 1)$ degrés de liberté.

2.2. Comparaison de la moyenne d'une population normale de variance σ^2 connue à une valeur donnée μ_0

Nous allons procéder en 4 étapes.

1) Faisons l'hypothèse que la moyenne de la population est égale à μ_0 :

$H_0 : \mu = \mu_0$, l'hypothèse alternative étant :

$H_1 = \mu \neq \mu_0$.

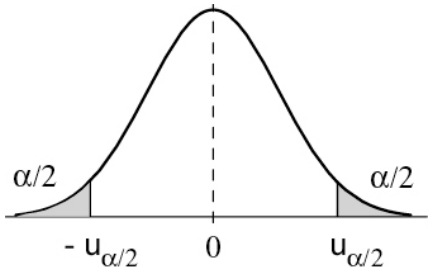
2) Il en résulte que la moyenne M d'un échantillon de taille n suit une loi normale de moyenne μ_0 et de variance $\frac{\sigma^2}{n}$ et que, par conséquent, la variable :

$$U = \frac{M-\mu_0}{\sigma/\sqrt{n}}$$

suit une loi normale réduite.

3) Fixons nous un risque α que nous conviendrons de considérer comme négligeable.

Il en résulte un certain intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$ dans lequel la variable U a une probabilité $(1 - \alpha)$ de tomber si l'hypothèse est exacte et, par conséquent, hors duquel U a une probabilité α petite de tomber. Négliger cette probabilité α, c'est considerer qu'il est impossible de trouver U en dehors de l'intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$, si l'hypothèse est vraie.



(5.3)

4) On calcule à partir des données de l'échantillon effectivement obtenu $(x_1, \ldots, x_n)$ la valeur u de U et on la situe par rapport à l'intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$. On conclut alors de la façon suivante :

- si u tombe à l'extérieur de l'intervalle, on préfère rejeter l'hypothèse, en sachant toutefois qu'on assume le risque α de la rejeter à tort.

- si u tombe à l'intérieur de l'intervalle, cela ne signifie nullement, hélas, que l'hypothèse faite est vraie, mais seulement que les données recueillies *ne sont pas en contradiction avec cette hypothèse*. Autrement dit, on est dans l'incapacité de conclure ni en faveur, ni en défaveur de l'hypothèse. On verra que dans les applications pratiques, cela est généralement moins génant qu'il n'y parait, parce que c'est contre un rejet, fait à tort, de l'hypothèse qu'il faut se prémunir, la conservation de l'hypothèse correspondant au statu quo.

2.3. Comparaison de la variance d'une population normale à une valeur donnée σ_0^2

Faisant l'hypothèse :

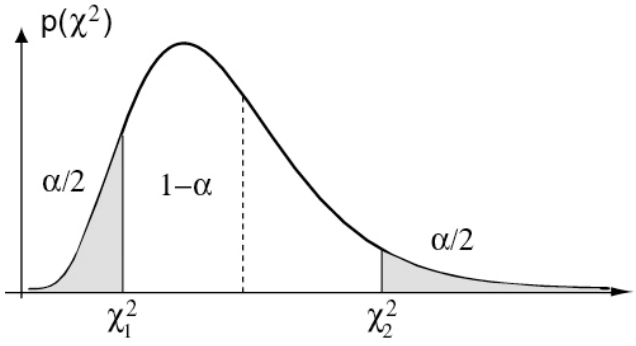
$H_0 = \sigma^2 = {\sigma_0}^2$,

la quantité :

$$\chi^2 = \frac{n S^2}{{\sigma_0}^2} = \frac{\sum_{i=1}^{n} (X_i - M)^2}{{\sigma_0}^2}$$

suit une loi du χ^2 à $(n - 1)$ degrés de liberté.

Il en résulte que, si l'hypothèse est vraie, $\frac{n S^2}{{\sigma_0}^2}$ a la probabilité $(1 - \alpha)$ de tomber dans l'intervalle $[\chi_1^2, \chi_2^2]$ où χ_1^2 et χ_2^2 sont lus dans la table de la loi du χ^2 à $(n - 1)$ degrés de liberté. Il suffit alors, comme précédement, de calculer la valeur $\frac{n s^2}{{\sigma_0}^2}$ à partir des observations, de la placer par rapport à l'intervalle $[\chi_1^2, \chi_2^2]$ et enfin de conclure.



(5.4)

2.4. Comparaison de la moyenne d'une population normale (de variance inconnue) à une valeur donnée μ_0

Faisant l'hypothèse :

$$H_0 : \mu = \mu_0,$$

la quantité :

$$T = \frac{M - \mu_0}{\sqrt{\frac{n}{n-1} S^2 / n}}$$

suit une loi de Student à $(n - 1)$ degrés de liberté.

Le test revient à placer la quantité :

$$t = \frac{m - \mu_0}{\sigma^* / \sqrt{n}} \text{ (où } \sigma^{*2} = \frac{n s^2}{n-1}),$$

par rapport à l'intervalle $[-t_{\alpha/2}, t_{\alpha/2}]$ lu dans la table de Student à $(n - 1)$ degrés de liberté.

2.5. Test des appariements

Nous avons présenté, dans l'introduction du chapitre, le dispositif expérimental qui consiste, disposant de n parcelles, à diviser chacune de ces parcelles en deux, et à cultiver chaque parcelle en soumettant l'une des moitiés à un certain traitement et l'autre moitié à un autre traitement. A chaque parcelle correspondront, en fin de culture, deux rendements *appariés*.

Imaginons un autre exemple, dans lequel on veuille confronter deux appareils de mesure et que, pour ce faire, on utilise n supports en procédant, sur chacun d'eux, à deux mesures à l'aide des deux appareils soumis à examen. Les deux mesures seront dites *appariées* et les résultats obtenus se présenteront, en définitive, comme suit :

mesures 1 : $x_1, x_2, \ \ldots, \ x_i, \ \ldots, \ x_n$

mesures 2 : $y_1, y_2, \ \ldots, \ y_i, \ \ldots, y_n$

Soit d_i la différence $d_i = (y_i - x_i)$ et soient m_d et $\sigma_d{}^*$ la moyenne et l'écart-type estimés des différences. On admet que les d_i sont des réalisations d'une variable D qui suit une loi normale. Le test de l'hypothèse H_0 : $E(D) = 0$ (pas d'influence du traitement ou pas de différence entre les appareils de mesures) est le test présenté au paragraphe précédent avec $\mu_0 = 0$.

3. Comparaison sur échantillons de deux populations normales

3.1. Comparaison des variances de deux populations normales

La comparaison de deux populations normales revient à se demander si elles ont *même moyenne* et *même variance* puisque ces deux paramètres suffisent à déterminer entièrement une distribution normale. Pour des raisons théoriques qui apparaitront dans un paragraphe suivant, la comparaison des variances doit précéder celle des moyennes.

Soient n_1 et ${s_1}^2$ la taille et la variance de l'échantillon extrait de la première population, et soient n_2 et ${s_2}^2$ la taille et la variance de l'échantillon extrait de la deuxième population. Nous savons que les estimations sans biais des variances ${\sigma_1}^2$ et ${\sigma_2}^2$ des deux populations s'écrivent :

$$\sigma_1^{*2} = \frac{n_1 {s_1}^2}{n_1 - 1} \text{ et } \sigma_2^{*2} = \frac{n_2 {s_2}^2}{n_2 - 1}.$$

Dans l'hypothèse d'égalité des variances des deux populations : ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$, ces deux estimations ne diffèrent qu'en raison des aléas de l'échantillonnage. Il en est de même de leur quotient $f = \frac{\sigma_1^{*2}}{\sigma_2^{*2}}$ qui ne diffère de 1 qu'à cause des aléas de l'échantillonnage.

Le statisticien Snedecor, auteur du test classique que nous allons présenter, a retenu cette forme et calculé la loi de probabilité de la variable :

$$F(\nu_1, \nu_2) = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

où χ_1^2 et χ_2^2 sont deux variables aléatoires indépendantes qui suivent des lois du χ^2 à ν_1 et ν_2 degrés de liberté.

Dans l'hypothèse d'égalité des variances des deux populations, si l'on désigne par ${S_1}^2$ et ${S_2}^2$ les variables, dont les variances des échantillons qui en sont extraits au hasard, sont des réalisations, $\frac{n_1 {S_1}^2}{\sigma^2}$ et $\frac{n_2 {S_2}^2}{\sigma^2}$ sont indépendantes et suivent des lois du χ^2 à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté. Il en résulte, par définition de cette variable, que le quotient :

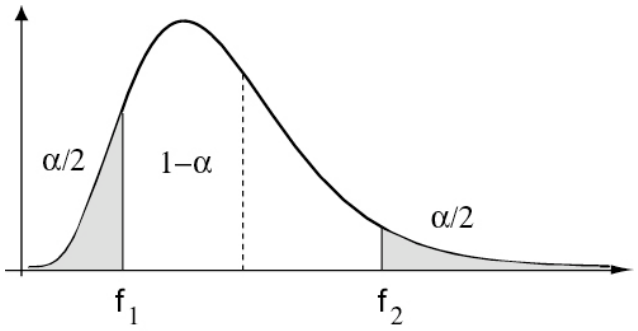
$$F = \frac{\frac{n_1 {S_1}^2}{\sigma^2}/(n_1 - 1)}{\frac{n_2 {S_2}^2}{\sigma^2}/(n_2 - 1)} = \frac{n_1 {S_1}^2/(n_1 - 1)}{n_2 {S_2}^2/(n_2 - 1)} \text{ (après simplification par } \sigma^2)$$

suit une loi de Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté. Par conséquent, la quantité :

$$f = \frac{\sigma_1^{*2}}{\sigma_2^{*2}}$$

est une réalisation, si l'hypothèse d'égalité des variances est vérifiée, d'une loi de Snedecor.

Cette loi définie, la suite des opérations est maintenant bien connue. Se fixant un seuil de probabilité α négligeable, on lit dans la table de Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté les valeurs f_1 et f_2 correspondant au dessin ci-dessous.



(5.5)

Telles qu'elles sont présentées, les tables de la loi de Snedecor portent, en tête de colonnes, le nombre de degrés de liberté du numérateur ν_1 et, en tête de lignes, celui du dénominateur ν_2; elles fournissent, à l'intersection de la colonne ν_1 et de la ligne ν_2, la limite supérieure f_2 de l'intervalle d'acceptation. Elles fournissent donc, à l'intersection de la colonne ν_2 et de la ligne ν_1, la valeur $1 / f_1$ de l'intervalle d'acceptation.

3.2. Estimation de σ^2

En admettant que le résultat du test précédent ne s'oppose pas à l'hypothèse d'égalité des variances, il peut s'avérer utile d'estimer la valeur commune σ^2 des variances des deux populations.

Puisque, dans l'hypothèse d'égalité des variances, $\frac{n_1 S_1^2}{\sigma^2}$ et $\frac{n_2 S_2^2}{\sigma^2}$ sont des variables indépendantes qui suivent des lois du χ^2, respectivement à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté leur somme $\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ suit une loi du χ^2 à $(n_1 + n_2 - 2)$ degrés de liberté, dont la moyenne et la variance sont respectivement $(n_1 + n_2 - 2)$ et $2(n_1 + n_2 - 2)$.

Il en résulte que la variable $\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$ est un estimateur sans biais et convergent de σ^2, puisque $E\left[\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right] = \sigma^2$ et $Var\left[\frac{n_1 S_1^2 + n_2 S_2^2}{(n_1 + n_2 - 2)}\right] = \frac{2\sigma^4}{n_1 + n_2 - 2} \longrightarrow 0$.

Par conséquent, la quantité :

$$\sigma^{*2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

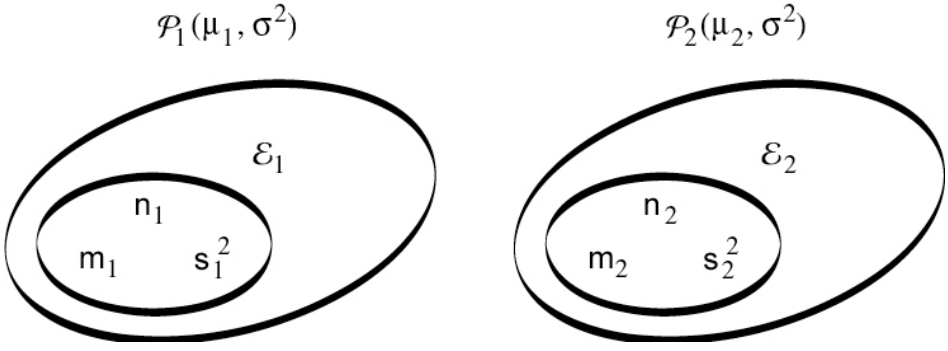
calculée à partir des observations, est une estimation sans biais de σ^2, la variance commune aux deux populations.

3.3. Comparaison des moyennes de deux populations normales

Dans l'hypothèse de populations normales, une fois testée l'égalité des variances, il suffit de tester l'égalité des moyennes pour pouvoir considérer que les populations sont identiques. Les raisons théoriques qui conduisent à présenter la comparaison des variances avant celle des moyennes peuvent, à ce stade, être explicitées. En effet, le test de comparaison des variances ne faisait aucune hypothèse sur l'égalité des moyennes. Par contre, le test d'égalité des moyennes implique l'égalité des variances. Il est donc nécessaire de vérifier cette égalité avant de s'intéresser aux moyennes.

Cela étant, soient deux populations normales $\mathcal{P}_1$ et $\mathcal{P}_2$ de moyennes μ_1 et μ_2, mais de *même variance* σ^2. Soient n_1 et n_2 les tailles de deux échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ prélevés au hasard respectivement dans chacune de ces deux populations ; soient m_1 et m_2 leurs moyennes, et soient $s_1{}^2$ et $s_2{}^2$ leurs variances.

(5.6)



Dans ces conditions, il est permis de considérer que :

- m_1 est une réalisation d'une variable M_1 normale, de moyenne μ_1 et de variance $\frac{\sigma^2}{n_1}$,

- m_2 est une réalisation d'une variable M_2 normale, de moyenne μ_2 et de variance $\frac{\sigma^2}{n_2}$,

- $s_1{}^2$ et $s_2{}^2$ sont des réalisations de variables $S_1{}^2$ et $S_2{}^2$ telles que la variable $\frac{n_1 S_1{}^2 + n_2 S_2{}^2}{\sigma^2}$ suit une loi du χ^2 à $(n_1 + n_2 - 2)$ degrés de liberté et est indépendante de M_1 et M_2.

Faisons maintenant l'hypothèse que $\mu_1 = \mu_2 = \mu$. Il en résulte que la variable $(M_1 - M_2)$ suit une loi normale de *moyenne nulle* et de variance égale à la somme des variances de M_1 et M_2, c'est-à-dire à $\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Par conséquent, la variable :

$$U = \frac{M_1 - M_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi normale réduite.

Pour éliminer la quantité σ inconnue, il suffit de considérer le quotient :

$$T = \frac{\frac{M_1 - M_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{\frac{n_1 S_1{}^2 + n_2 S_2{}^2}{\sigma^2}}{n_1 + n_2 - 2}}} = \frac{M_1 - M_2}{\sqrt{\frac{n_1 S_1{}^2 + n_2 S_2{}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

qui suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté. Pour simplifier l'écriture, on peut tenir compte de ce que figure, au dénominateur, l'expression de l'estimateur sans biais de σ^2. Par conséquent $t = \frac{m_1 - m_2}{\sigma^*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ est une réalisation d'une loi de Student qu'il suffit, pour conclure, de placer par rapport à l'intervalle $[-t_{\alpha/2}, t_{\alpha/2}]$ correspondant au risque α choisi.

Si t n'appartient pas à l'intervalle, on dit souvent que la différence entre les moyennes observées est *significative* au risque α et, sinon, qu'elle n'est *pas significative*.

3.4. Estimation de la différence des moyennes des populations

Si la différence observée entre les moyennes m_1 et m_2 des échantillons est *significative* (d'une différence entre les moyennes μ_1 et μ_2 des populations), il peut s'avérer utile d'estimer la différence $\Delta = \mu_1 - \mu_2$. La variable $(M_1 - M_2)$ est évidemment un estimateur sans biais de Δ. Quant à la détermination de l'intervalle de confiance, elle repose sur la prise en compte de la variable :

$$T = \frac{(M_1 - M_2) - \Delta}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

qui suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté.

On a, par conséquent, au risque α près :

$$(m_1 - m_2) - t_{\alpha/2}\, \sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \Delta < (m_1 - m_2) + t_{\alpha/2}\, \sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Exercices du chapitre 5

Exercice 1

On a prélevé, au hasard dans une population normale de moyenne μ et d'écart-type σ, un échantillon de taille $n = 10$. La moyenne et la variance calculées sur cet échantillon sont respectivement $m = 4$ et $s^2 = 6$.

a) Calculer une estimation sans biais de σ et son intervalle de confiance au risque 5%.

b) Tester l'hypothèse $\sigma = 2$ au risque 5%.

c) En admettant σ connu égal à 2, tester l'hypothèse $\mu = 3$ au risque 5%.

d) Tester, au risque 5%, l'hypothèse $\mu = 3$ sans faire aucune hypothèse sur la valeur de σ.

e) Calculer une estimation sans biais de μ et son intervalle de confiance au risque 5% sans faire aucune hypothèse sur la valeur de σ.

f) En admettant μ connu égal à 3, est-il possible d'envisager un test plus efficace que celui mis en oeuvre en b) pour tester l'hypothèse $\sigma = 2$?

Exercice 2

Pour comparer les rendements de deux variétés de blé A et B, on a ensemencé 10 couples de deux parcelles voisines, l'une en variété A, l'autre en variété B, les 10 couples étant répartis dans des localités différentes. On a obtenu les résultats suivants :

couple n°	1	2	3	4	5	6	7	8	9	10
récolte A	45	32	56	49	45	38	47	51	42	38
récolte B	47	34	52	51	48	44	45	56	46	44

Que peut-on conclure de ces résultats ?

Exercice 3

On donne ci-après les pourcentages de matière grasse dans un aliment, déterminés sur 10 échantillons par deux méthodes d'analyse différentes A et B.

échantillon n°	1	2	3	4	5	6	7	8	9	10
méthode A	24.6	25.3	25.3	25.6	25.6	25.9	26	27	27.3	27.7
méthode B	24.9	25.6	25.8	26.2	26.1	26.7	26.3	26.9	28.4	27.1

Comparer ces deux méthodes.

Exercice 4

On a prélevé au hasard un échantillon $\mathcal{E}_1$ de taille $n_1 = 10$ dans une population normale $\mathcal{P}_1$ de moyenne μ_1 et d'écart-type σ_1. La moyenne et la variance calculées sur cet échantillon sont respectivement $m_1 = 4$ et ${s_1}^2 = 6$.

On préleve au hasard un échantillon $\mathcal{E}_2$ de taille $n_2 = 15$ dans une population normale $\mathcal{P}_2$ de moyenne μ_2 et d'écart-type σ_2. La moyenne et la variance calculées sur cet échantillon sont respectivement $m_2 = 7$ et ${s_2}^2 = 20$.

a) Tester l'hypothèse $\sigma_2 = \sigma_1$, au risque 5%.

b) Tester l'hypothèse $\sigma_2 = 2\,\sigma_1$, au risque 5%.

c) En admettant que $\sigma_2 = 2\,\sigma_1$, calculer une estimation sans biais de σ_1, à partir des deux échantillons, et son intervalle de confiance au risque 5%.

d) Utiliser un test du χ^2 pour tester simultanément les hypothèses $\sigma_2 = 4$ et $\sigma_1 = 2$.

e) En admettant que $\sigma_2 = 2\,\sigma_1 = 4$, tester, au risque 5%, l'hypothèse $\mu_2 = 2\,\mu_1$.

f) Calculer une estimation de μ_1 à partir des deux échantillons, en admettant que $\mu_2 = 2\,\mu_1$ et son intervalle de confiance au risque 5%.

Exercice 5

Il y a des raisons de penser que l'épaisseur de la cire dont sont enduits des sacs en papier est plus irrégulière à l'intérieur qu'à l'extérieur. Pour le vérifier 75 mesures de l'épaisseur ont été faites et ont donné les résultats suivants :

- surface intérieure : $\sum x = 71.25$ et $\sum x^2 = 91$

- surface extérieure : $\sum y = 48.75$ et $\sum y^2 = 84$.

a) Faire un test pour déterminer, au risque 5%, si la variabilité de l'épaisseur de la cire est plus grande à l'intérieur qu'à l'extérieur des sacs.

b) Revenant à la loi de F, calculer l'intervalle de confiance à 95% du rapport des variances.

Exercice 6

Deux chaines de fabrication produisent des transistors. Des relevés effectués pendant 10 jours ont donné les résultats suivants :

- ligne 1 : $m_x = 2800$ et $\sum (x - m_x)^2 = 103600$

- ligne 2 : $m_y = 2680$ et $\sum (y - m_y)^2 = 76400$

On admettra que les écarts-type σ_x et σ_y sont inconnus mais égaux.

a) Peut-on conclure, au risque de 5%, à une différence entre les productions moyennes des deux lignes ?

b) Quel est l'intervalle de confiance à 95% de la différence ?