Statistique pour ingénieur

Thème 3 : Tests d'hypothèses, analyse de la variance

T. Verdel & M. Sauceau, 21 novembre 2016

Introduction

Le troisième thème que nous abordons est consacré à un mode de raisonnement nouveau grâce auquel nous allons chercher, à partir d'un ou plusieurs échantillons issus des populations étudiées à :

- comparer certaines caractéristiques de ces populations à des valeurs choisies a priori (tests de conformité à un standard, section 2),
- comparer entre elles certaines caractéristiques de 2 ou plus de 2 populations (tests de comparaison, section 3 et section 6 et analyse de la variance, section 7 et section 8),
- vérifier la compatibilité d'une population à une loi de probabilité choisie a priori (tests d'adéquation ou d'ajustement, section 4),
- se prononcer sur l'indépendance ou non de deux variables (test d'indépendance ou d'association, section 5).

Ce mode de raisonnement a été introduit au début du XIXe siècle, par William Gosset, statisticien anglais embauché par la brasserie Guinness, qui publiera ses travaux sous 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, l'espérance $\mathbb{E}(D)$ 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 $\mathbb{E}(D)=0$. Si elles sont incompatibles, l'hypothèse faite doit être remise en cause, et l'on peut conclure à l'influence de l'engrais... C'est ce raisonnement, théorisé plus tard par Neyman et Pearson, qui est appelé le test d'hypothèses.

On trouvera sur internet de nombreuses ressources à propos des tests statistiques qui permettrons au lecteur de compléter ses connaissances au delà du contenu de ce document.

Table des matières

1 Principe des tests d'hypothèses

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 $\mathcal{H}_0$,$\mathcal{H}_1$,...,$\mathcal{H}_n$ 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 $\mathcal{H}_0$ et $\mathcal{H}_1$. 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 :

— $\mathcal{H}_0$: le lot contient 5% de déchets

— $\mathcal{H}_1$: le lot contient 10% de déchets

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

— $\mathcal{H}_0$: le lot contient 5% ou moins de 5% de déchets

— $\mathcal{H}_1$: 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 $\mathcal{H}_0$ et $\mathcal{H}_1$ couvrant l'ensemble des possibilités ; cela veut dire que l'une ou l'autre des deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ est réalisée nécessairement. Dans ce cas, il est possible d'émettre l'un des deux jugements :

— $\mathcal{H}_0$ est vraie, donc $\mathcal{H}_1$ est fausse

— $\mathcal{H}_1$ est vraie, donc $\mathcal{H}_0$ est fausse.

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

	état réalisé	
jugement	$\mathcal{H}_0$ est réalisée	$\mathcal{H}_1$ est réalisée
$\mathcal{H}_0$	jugement correct	jugement faux
$\mathcal{H}_1$	jugement faux	jugement correct

Parmi les deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, 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 $\mathcal{H}_0$ et $\mathcal{H}_1$ de façon symétrique. Ainsi, on peut commettre deux types d'erreur :

— l'erreur de 1ère espèce qui est la probabilité de rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie ;

— l'erreur de 2e espèce qui est la probabilité d'accepter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie.

C'est exactement en ces termes que se posait le problème du contrôle abordé dans le thème 2.

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

— on dit que $\mathcal{H}_0$ est vraie si la valeur observée de X, soit x, se trouve dans un certain domaine ω, appelé *région d'acceptation* de l'hypothèse $\mathcal{H}_0$;

— on dit que $\mathcal{H}_1$ est vraie si la valeur observée appartient à $\overline{\omega}$, appelé *région critique* ou *région de rejet*.

Pour choisir le domaine ω, 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 ω 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 :

	état réalisé	
décision	$\mathcal{H}_0$ est réalisée	$\mathcal{H}_1$ est réalisée
$\mathcal{H}_0(X \in \omega)$	jugement correct $(1-\alpha)$	jugement faux (β)
$\mathcal{H}_1(X \notin \omega)$	jugement faux (α)	jugement correct $(1-\beta)$

Un tel mode de raisonnement est appelé test d'hypothèses. 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.

L'hypothèse $\mathcal{H}_0$ sur laquelle sera mené le test est appelée *l'hypothèse nulle*.

1.2 Détermination de la région d'acceptation

Si l'on note $f_0(x)$ et $f_1(x)$ les densités de probabilité d'un échantillon prélevé dans une population caractérisée par une variable aléatoire X, respectivement dans le cadre des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, les deux conditions précédentes s'expriment par les deux équations suivantes :

$$\int_\omega f_0(x)dx = 1-\alpha$$

$$\int_\omega f_1(x)dx = \beta \text{ minimum}$$

Théorème 1

Neyman et Pearson ont démontré qu'elles sont satisfaites s'il existe une constante positive λ et un domaine ω tels que pour x appartenant à ω :

$$f_1(x) \leqslant \lambda f_0(x)$$

$$\textit{sous la contrainte} \int_\omega f_0(x)dx = 1-\alpha$$

Appliquons ce résultat au test sur une proportion et au test sur une moyenne.

1.3 Test sur une proportion

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

— $\mathcal{H}_0$: la proportion de déchets est p_0, contre l'hypothèse

— $\mathcal{H}_1$: la proportion de déchets est p_1

Le nombre d'articles défectueux dans l'échantillon est une variable aléatoire définie par les probabilités $f_0(k)$ si $\mathcal{H}_0$ est vraie et $f_1(k)$ si c'est $\mathcal{H}_1$ avec :

$$f_0(k) = \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

$$f_1(k) = \binom{n}{k} p_1^k (1-p_1)^{n-k}$$

La première condition s'écrit :

$$\binom{n}{k} p_1^k (1-p_1)^{n-k} \leqslant \lambda \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

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

$$k \ln\left(\frac{p_0}{p_1}\right) + (n-k) \ln\left(\frac{1-p_0}{1-p_1}\right) + \ln(\lambda) > 0$$

Soit, pour $p_1 > p_0$:

$$k < \frac{n \ln\left(\frac{1-p_1}{1-p_0}\right) - \ln(\lambda)}{\ln\left(\frac{p_0}{p_1}\right) - \ln\left(\frac{1-p_0}{1-p_1}\right)} = k_c$$

L'inégalité se réduit donc à $k < k_c$. Pour déterminer k_c, il suffit d'utiliser la deuxième condition qui s'écrit :

$$\sum_{k=0}^{k_c} \binom{n}{k} p_0^k (1-p_0)^{n-k} = 1 - \alpha$$

On notera que la région d'acceptation ne dépend pas de la valeur p_1, c'est-à-dire de l'hypothèse $\mathcal{H}_1$. Par contre, le risque de deuxième espèce en dépend puisque :

$$\beta = \sum_{k=0} k_c \binom{n}{k} p_1^k (1-p_1)^{n-k}$$

1.4 Test sur une moyenne

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

— $\mathcal{H}_0 : \mu = \mu_0$

— $\mathcal{H}_1 : \mu = \mu_1$

Sous l'hypothèse $\mathcal{H}_0$, la densité de probabilité d'un tel échantillon $(X_1, X_2, \ldots, X_n)$ s'écrit, les variables X_i étant indépendantes et de même densité de probabilité que celle de X :

$$f_0(x_1) \times f_0(x_2) \times \cdots \times f_0(x_n)$$

Dès lors, X suivant une loi normale, la région d'acceptation est définie par :

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

En simplifiant par le facteur et en passant au logarithme on a :

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

Expression que l'on peut écrire aussi, après multiplication par $2\sigma^2$:

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

$$\sum_{i=1}^{n}{x_i}^2 - 2\mu_0\sum_{i=1}^{n}x_i + n{\mu_0}^2 - \left(\sum_{i=1}^{n}{x_i}^2 - 2\mu_1\sum_{i=1}^{n}x_i + n{\mu_1}^2\right) < 2\sigma^2\ln(\lambda)$$

Soit, en notant $\bar{x}$ la moyenne empirique :

$$2\mu_0 n\bar{x} + n{\mu_0}^2 + 2\mu_1 n\bar{x} - n{\mu_1}^2 < 2\sigma^2\ln(\lambda)$$

$$2\bar{x}(\mu_1-\mu_0) - \left({\mu_1}^2 - {\mu_0}^2\right) < \frac{2\sigma^2\ln(\lambda)}{n}$$

$$2\bar{x}(\mu_1-\mu_0) - (\mu_1-\mu_0)(\mu_1+\mu_0) < \frac{2\sigma^2\ln(\lambda)}{n}$$

Et en supposant que $\mu_1 > \mu_0$:

$$2\bar{x} - (\mu_1+\mu_0) < \frac{2\sigma^2\ln(\lambda)}{n(\mu_1-\mu_0)}$$

$$\bar{x} < \frac{\mu_1+\mu_0}{2} + \frac{\sigma^2\ln(\lambda)}{n(\mu_1-\mu_0)} = \bar{x}_c$$

Pour définir $\bar{x}_c$, il suffit alors d'écrire que :

$$\mathbb{P}\left(\overline{X} > \overline{x}_c\right) = \alpha \text{ si } X \sim \mathcal{N}(\mu_0,\sigma^2)$$

où $\overline{X}$, qui est donc notre variable décision, désigne la moyenne empirique d'un n-échantillon. Remarquons que, dans ce deuxième exemple aussi, la région d'acceptation ne dépend pas de l'hypothèse $\mathcal{H}_1$.

1.5 Cas des hypothèses composites

En réalité, très souvent, le problème n'est pas de choisir entre deux hypothèses simples $\mathcal{H}_0$ et $\mathcal{H}_1$, mais entre une hypothèse simple $\mathcal{H}_0$ et un ensemble plus ou moins vaste d'hypothèses $\mathcal{H}_1,\ldots,\mathcal{H}_i,\ldots,\mathcal{H}_n$ ou même à un ensemble continu d'hypothèses $\mathcal{H}$. Dans ce cas, on peut se ramener au problème précédent en comparant successivement $\mathcal{H}_0$ à chacune des hypothèses de l'ensemble $\mathcal{H}$. Si, par exemple, on compare $\mathcal{H}_0$ à $\mathcal{H}_i$, la méthode exposée plus haut permet de trouver une région ω_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 $\omega_1,\ldots,\omega_i,\ldots,\omega_n$ 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 ω_i ont une partie commune ω. Dans ce domaine ω, le test utilisé est dit uniformément le plus puissant (en abréviation de l'anglais : UMP). En effet, lorsque X tombe dans ω,

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 $\mathcal{H}$ 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 $\mathcal{H}_1$. Pas tout à fait cependant. Notons, en effet, que nous avons supposé, respectivement dans chacun des deux exemples, que $p_1 > p_0$ et que $\mu_1 > \mu_0$. Et nous avons abouti alors à des régions d'acceptation de la forme $k < k_c$ et $\bar{x} < \bar{x}_c$ telles que le risque α soit bloqué à l'une des extrémités de la distribution de la variable étudiée.

Si donc il s'agit de comparer deux hypothèses de la forme : $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta > \theta_0$, on est conduit à ce qu'on appelle un test *unilatéral à droite*, où le risque de première espèce est bloqué à droite (*cf.* figure 1).

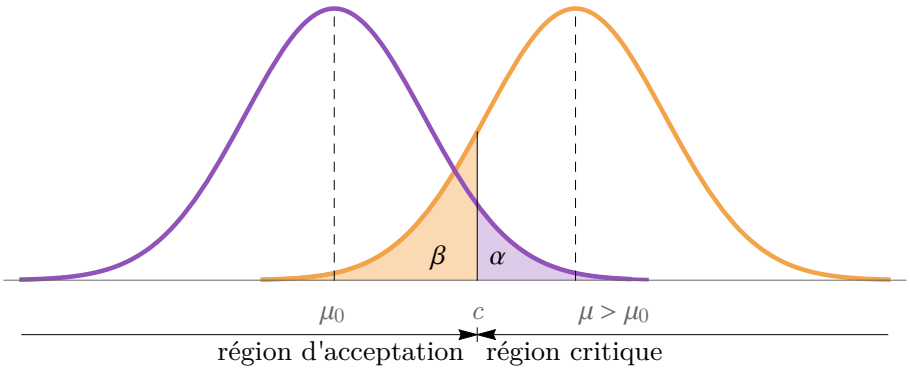

FIGURE 1 – Principe d'un test unilatéral à droite

De même, le test d'hypothèses de la forme $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta < \theta_0$, conduit à un test *unilatéral à gauche*.

Enfin, dans le cas d'hypothèses de la forme $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta \neq \theta_0$, il apparaît logique de répartir le risque α aux deux extrémités de la distribution. Le test est alors dit *symétrique* ou *bilatéral* (*cf.* figure 2).

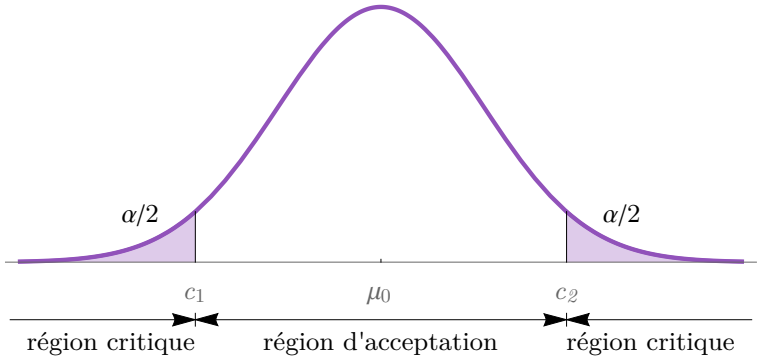

FIGURE 2 – Principe d'un test symétrique ou bilatéral

1.6 Démarche générale d'un test

En résumé, la démarche générale d'un test consiste à suivre les étapes suivantes :

1. Choisir et formuler les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$.
2. Déterminer la variable de décision, celle avec laquelle le test sera effectivement mené et qui sera évaluée sur un échantillon.
3. Choisir le risque α et calculer la région d'acceptation ou, préférentiellement, la région critique qui en dépend.
4. Calculer éventuellement la puissance du test $1 - \beta$
5. Calculer la valeur expérimentale de la variable de décision à partir d'un échantillon
6. La comparer à la région critique pour conclure en rejetant ou non l'hypothèse $\mathcal{H}_0$

On gardera à l'esprit, compte tenu de la dissymétrie des deux hypothèses testées, que la conclusion du test est plus forte quand on rejette l'hypothèse $\mathcal{H}_0$, le non-rejet ne valant pas vérité. Ainsi, le plus souvent, on choisira les hypothèses de telles sorte que $\mathcal{H}_1$ soit l'hypothèse pour laquelle on aimerait conclure.

2 Tests de conformité à un standard

2.1 Rappel des lois outils usuelles

La détermination des régions d'acceptation nécessite le choix d'une variable de décision et donc la mise en oeuvre de 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 les thèmes précédents.

2.1.1 Loi normale centrée réduite $\mathcal{N}(0,1)$

Théorème 2

Etant donnée une variable X qui suit une loi $\mathcal{N}(\mu,\sigma^2)$, on a :

$$U = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Corollaire 3

Ainsi, considérant la variable $\overline{X}$, moyenne empirique d'un n-échantillon, prélevé dans une population caractérisée par la variable X normale $\mathcal{N}(\mu,\sigma^2)$, on a :

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

2.1.2 Loi du χ^2

Définition 1

Etant données ν variables U_1,U_2,...,U_ν indépendantes et suivant des lois normales centrées réduites $\mathcal{N}(0,1)$, alors :

$$Z_\nu = U_1^2 + U_2^2 + \cdots + U_\nu^2 \sim \chi^2(\nu)$$

Corollaire 4

Il en résulte qu'étant donné un n-échantillon $(X_1,\ldots,X_i,\ldots,X_n)$, où les X_i suivent des lois $\mathcal{N}(\mu,\sigma^2)$ indépendantes, on a :

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

Théorème 5

Notant $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)$ la variance empirique d'un n-échantillon prélevé dans une population caractérisée par une variable $X \sim \mathcal{N}(\mu,\sigma^2)$, on a :

$$Z = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

2.1.3 Loi de Student $\mathcal{T}$

Définition 2

Etant données U et Z deux variables indépendantes suivant respectivement la loi $\mathcal{N}(0,1)$ et une loi $\chi^2(\nu)$ alors :

$$T = \frac{U}{\sqrt{Z/\nu}} \sim \mathcal{T}(\nu)$$

Corollaire 6

Il en résulte que $\overline{X}$ et S^2 étant respectivement la moyenne empirique et la variance empirique d'un n-échantillon, prélevé dans une population caractérisée par une variable $X \sim \mathcal{N}(\mu,\sigma^2)$, on a :

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n-1}} \sim \mathcal{T}(n-1)$$

2.2 Test de conformité sur la moyenne d'une variable aléatoire normale de variance connue

Ce test repose sur la variable $\overline{X}$, moyenne empirique d'un n-échantillon.
Nous allons procéder en 4 étapes :

1. Faisons l'hypothèse $\mathcal{H}_0 : \mu = \mu_0$, l'hypothèse alternative étant $\mathcal{H}_1 : \mu \neq \mu_0$
2. Il en résulte que $\overline{X}$ suit une loi $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$ et que, par conséquent :

$$U = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

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 (notre variable décision) 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. Autrement dit : si $\mathcal{H}_0$ est vraie, $\mathbb{P}(|U| > u_{\alpha/2}) = \alpha$ et négliger la probabilité α, c'est considérer qu'il est impossible de trouver U en dehors de l'intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$, si l'hypothèse est vraie (*cf.* figure 3).
4. A partir des données de l'échantillon effectivement obtenu $(x_1, \ldots, x_n)$, on calcule 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 gênant qu'il n'y paraît, 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.

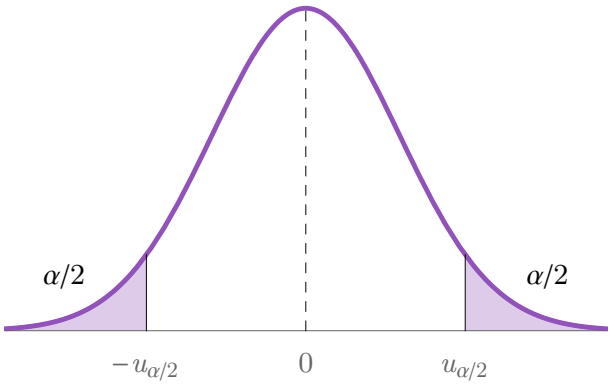

FIGURE 3 – Région critique (en couleur) correspondant à la zone de probabilité α, dans le cas d'un test bilatéral sur la moyenne d'une variable normale dont la variance est connue (loi $\mathcal{N}(0,1)$)

Exemple 1

Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir 2,5 g de substance active. 100 comprimés sont choisis au hasard dans la production et analysés. Ils contiennent en moyenne 2,6 g de substance active, l'écart-type étant connu égal à 0,4 g et la distribution normale. Le lot testé respecte-t-il les spécifications au risque 5% ?

On fait l'hypothèse $\mathcal{H}_0 : \mu = 2{,}5$. L'hypothèse alternative est $\mathcal{H}_1 : \mu \neq 2{,}5$. Il s'agit d'un test bilatéral. Sous $\mathcal{H}_0$, la variable $U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-2{,}5}{0.4/\sqrt{100}}$ suit une loi $\mathcal{N}(0,1)$. Au risque 5%, la région critique est $u < -1{,}96$ ou $u > 1{,}96$. On calcule $u = \frac{2{,}6-2{,}5}{0{,}4/\sqrt{100}} = 2{,}5$ qui se trouve donc dans la région critique. On peut donc rejeter l'hypothèse $\mathcal{H}_0$ avec un risque inférieur à 5% de la rejeter à tort. Au risque de 5%, on peut considérer que ce lot de médicaments ne respecte pas les spécifications.

2.3 Test de conformité sur la moyenne d'une variable normale de variance inconnue

Ce test repose sur les variables $\overline{X}$ et S^2, respectivement la moyenne et la variance empiriques d'un n-échantillon et nous avons vu au corollaire 6 que :

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \sim \mathcal{T}(n-1)$$

C'est notre variable de décision. Le test revient donc à placer la quantité $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n-1}}$ par rapport à l'intervalle $[-t_{\alpha/2}, t_{\alpha/2}]$ lu dans la table de Student à $(n-1)$ degrés de liberté et défini tel que : $\mathbb{P}(|T| > t_{\alpha/2}) = \alpha$ (*cf.* figure 4).

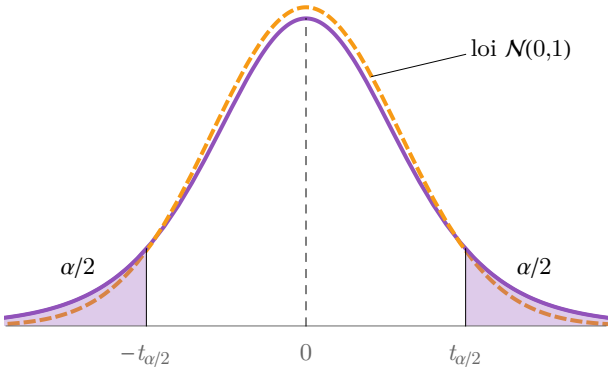

FIGURE 4 – Région critique (en couleur) dans un test bilatéral sur la moyenne d'une population normale dont la variance est inconnue (loi $\mathcal{T}(n-1)$).

Exemple 2

Reprenons les données de l'exemple 1 dans le cas où nous ne connaissons pas l'écart-type de la population et où l'écart-type mesuré sur notre échantillon vaut 0.6 g. Le lot testé respecte-t-il les spécifications au risque 5% ?

On fait l'hypothèse $\mathcal{H}_0 : \mu = 2{,}5$. L'hypothèse alternative est $\mathcal{H}_1 : \mu \neq 2{,}5$. Il s'agit d'un test bilatéral. Sous $\mathcal{H}_0$, la variable $T = \frac{\overline{X}-\mu}{S/\sqrt{n-1}} = \frac{\overline{X}-2.5}{S/\sqrt{99}}$ suit une loi $\mathcal{T}(99)$. Au risque

5%, la région critique est $t < -1{,}98$ ou $t > 1{,}98$. On calcule $t = \frac{2{,}6-2{,}5}{0{,}6/\sqrt{99}} = 1{,}66$ qui n'est pas dans la région critique. On ne peut donc pas rejeter l'hypothèse $\mathcal{H}_0$, mais cela ne signifie pas qu'elle soit vraie.

2.4 Test de conformité sur la variance d'une variable normale

Ce test repose sur la variable S^2, variance empirique d'un n-échantillon.
Faisant l'hypothèse $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$, nous avons vu au théorème 5 que :

$$Z = \frac{nS^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Il en résulte que, si l'hypothèse est vraie, Z a la probabilité $(1-\alpha)$ de tomber dans l'intervalle $[z_1,z_2]$ où z_1 et z_2 sont lus dans la table de la loi du χ^2 à $(n-1)$ degrés de liberté et tels que : $\mathbb{P}(Z < z_1) = \mathbb{P}(Z > z_2) = \alpha/2$.

Il suffit alors, comme précédemment, de calculer la valeur $z = ns^2/\sigma_0^2$ à partir des observations et de la placer par rapport à l'intervalle $[z_1,z_2]$ pour conclure (*cf.* figure 5).

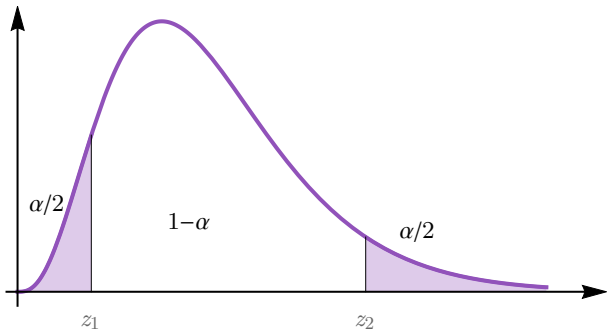

FIGURE 5 – Région critique (en couleur) correspondant à la zone de probabilité α, dans le cas d'un test bilatéral sur la variance d'une variable normale (loi $\chi^2(n-1)$).

Exemple 3

L'utilisateur d'un fil d'acier a l'impression que l'écart-type de sa résistance est plus grand que la valeur de 0,75 Newton, donnée par le fabricant. Pour le vérifier, il procède à un test de résistance sur plusieurs morceaux de ce fil et obtient les valeurs suivantes.

x	72,1	74,5	72,8	75,0	73,4	75,4	76,1	73,5	74,1

Faisant l'hypothèse que la résistance du fil est normalement distribuée, peut-on considérer que son écart-type dépasse la spécification au risque 5% ?

On fait l'hypothèse $\mathcal{H}_0 : \sigma^2 = 0{,}75^2 = 0{,}5625$. L'hypothèse alternative est $\mathcal{H}_1 : \sigma^2 > 0{,}5625$ car on veut éventuellement montrer que l'écart-type est plus grand que prévu. Il s'agit donc d'un test unilatéral à droite. On calcule aisément la variance et l'écart-type empiriques de l'échantillon : $s^2 = 1{,}4667$ et $s = 1{,}2111$.

Sous $\mathcal{H}_0$, la variable $Z = \frac{nS^2}{\sigma^2} = \frac{9S^2}{0{,}75^2}$ suit une loi $\chi^2(8)$. Au risque 5%, placé à droite, la valeur critique est 15,5. On calcule $z = \frac{9\times 1{,}4667}{0{,}75^2} = 23{,}47$ qui dépasse la valeur critique. On peut donc rejeter l'hypothèse $\mathcal{H}_0$ et conclure que la dispersion dans la résistance du fil est bien plus grande que telle que spécifiée par le fabricant.

2.5 Test des appariements ou test sur données appariées

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 :

articles	1	2	$\cdots$	i	$\cdots$	n
série 1	x_1	x_2	$\cdots$	x_i	$\cdots$	x_n
série 2	y_1	y_2	$\cdots$	y_i	$\cdots$	y_n

Soit $d_i = (y_i - x_i)$ et soient $\overline{d}$ et s_d la moyenne et l'écart-type des différences. On admet que les d_i sont des réalisations d'une variable D qui suit une loi normale (ce qui sous-entend la normalité des variables X_1 et X_2). Le test de l'hypothèse $\mathcal{H}_0 : \mathbb{E}(D) = 0$ (pas d'influence du traitement ou pas de différence entre les appareils de mesures) est le test présenté à la section 2.3 avec $\mu_0 = 0$.

Exemple 4

On dispose de 2 méthodes pour estimer la résistance au cisaillement de poutres en acier. Sur un échantillon de 9 poutres, on a obtenu les résultats suivants qui correspondent au rapport entre la valeur prédite par la méthode et la valeur effectivement mesurée. Au risque 5%, y-a-t-il une différence entre les 2 méthodes ?

méthode a	1,286	1,251	1,322	1,339	1,250	1,402	1,365	1,437	1,459
méthode b	1,061	1,092	1,363	1,362	1,265	1,278	1,237	1,386	1,352
$d = a - b$	0,225	0,159	−0,041	−0,023	−0,015	0,124	0,128	0,051	0,107

Notant $\mu_D = \mathbb{E}(D)$, on fait l'hypothèse $\mathcal{H}_0 : \mu_D = 0$. L'hypothèse alternative est $\mathcal{H}_1 : \mu_D \neq 0$. Il s'agit d'un test bilatéral. Sous $\mathcal{H}_0$, la variable $T = \frac{\overline{D} - \mu_D}{S/\sqrt{n-1}} = \frac{\overline{D}-0}{S/\sqrt{8}}$ suit une loi $\mathcal{T}(8)$. Au risque 5%, la région critique est $t < -2{,}306$ ou $t > 2{,}306$. On calcule $t = \frac{0{,}0794-0}{0{,}0865/\sqrt{8}} = 2{,}598$ qui appartient donc à la région critique. On peut alors rejeter l'hypothèse $\mathcal{H}_0$ et conclure à une différence significative entre les 2 méthodes au risque 5%.

3 Tests de comparaison de 2 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 apparaîtront dans un paragraphe suivant, la comparaison des variances doit précéder celle des moyennes.

3.1 Comparaison des variances

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 σ_1^2 et σ_2^2 des deux populations s'écrivent :

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

Dans l'hypothèse d'égalité des variances des deux populations ($\sigma_1^2 = \sigma_2^2$), ces deux estimations ne diffèrent qu'en raison des aléas de l'échantillonnage. Il en est de même de leur quotient $f = s_1^{*2}/s_2^{*2}$ qui ne diffère de 1 qu'à cause des aléas de l'échantillonnage.

Le statisticien Ronald Aylmer Fisher, biologiste et mathématicien britannique, auteur du test classique que nous allons présenter, a retenu cette forme et calculé la loi de probabilité de la variable :

$$F\,(\nu_1,\nu_2) = \frac{Z_1/\nu_1}{Z_2/\nu_2}$$

où Z_1 et Z_2 sont deux variables aléatoires indépendantes qui suivent des lois du χ^2 à respectivement ν_1 et ν_2 degrés de liberté.

Il en résulte le résultat suivant :

Théorème 7

Si l'on désigne par S_1^2 et S_2^2 les variances empiriques de deux échantillons extraits de populations normales, alors la variable :

$$F = \frac{\frac{n_1 S_1^2}{\sigma_1^2}/\,(n_1 - 1)}{\frac{n_2 S_2^2}{\sigma_2^2}/\,(n_2 - 1)} = \frac{S_1^{*2}/\sigma_1^2}{S_2^{*2}/\sigma_2^2}$$

suit une loi de Fisher à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté, notée $\mathcal{F}(n_1 - 1,n_2 - 1)$.

Corollaire 8

Dès lors, faisant l'hypothèse : $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, on a :

$$F = \frac{n_1 S_1^2/\,(n_1 - 1)}{n_2 S_2^2/\,(n_2 - 1)} = \frac{S_1^{*2}}{S_2^{*2}} \sim \mathcal{F}(n_1 - 1,n_2 - 1)$$

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 Fisher-Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté, les valeurs f_1 et f_2 correspondant à la figure 6.

Telles qu'elles sont présentées, les tables de la loi de Fisher-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. Bien entendu, les logiciels scientifiques fourniront directement les quantiles recherchés.

Plus souvent appelé test de Fisher, ce test prend parfois l'appellation de test de Fisher-Snedecor, test de Snedecor ou encore de F-test.

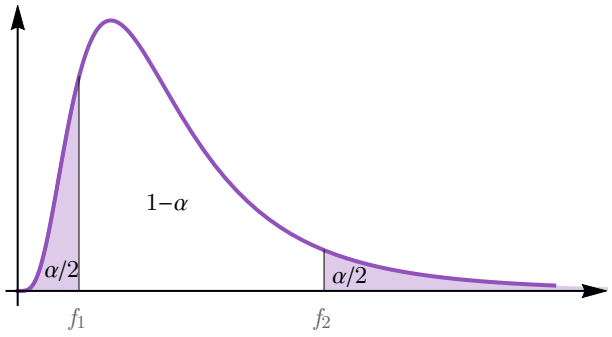

FIGURE 6 – Région critique (en couleur) dans un test bilatéral de comparaison de variances (loi de Fisher-Snedecor).

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 est utile d'estimer la valeur commune σ^2 des variances des deux populations. Nous en aurons notamment besoin pour le test de comparaison des moyennes présenté à la section suivante.

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 :

Théorème 9

Faisant l'hypothèse : $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, la variable :

$$S^{*2} = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

est un estimateur sans biais et convergent de σ^2.

Preuve. On a en effet :

$$\mathbb{E}\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right) = \sigma^2 \text{ et } \mathbb{V}\left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right) = \frac{2\sigma^4}{n_1 + n_2 - 2} \longrightarrow 0$$

□

3.3 Comparaison des moyennes

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 $\mathcal{P}_1$ et $\mathcal{P}_2$ caractérisées par des variables aléatoires normales X_1 et X_2 de moyennes respectives 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 dans chacune de ces deux populations ; soient $\overline{x}_1$ et $\overline{x}_2$ leurs moyennes et s_1^2 et s_2^2 leurs variances (*cf.* figure 7).

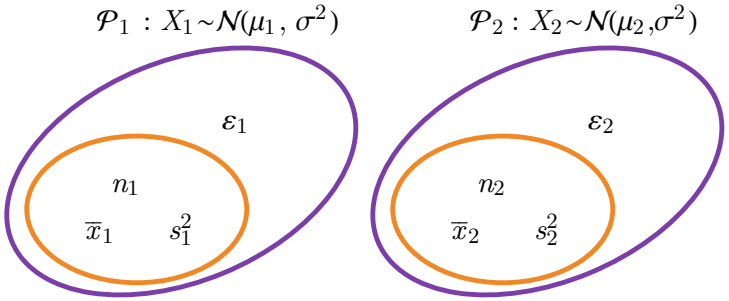

FIGURE 7 – Données pour la comparaison de deux populations normales

Dans ces conditions, il est permis de considérer que :
- $\overline{x}_1$ est une réalisation de la variable $\overline{X}_1$ qui suit une loi $\mathcal{N}(\mu_1, \sigma^2/n_1)$
- $\overline{x}_2$ est une réalisation de la variable $\overline{X}_2$ qui suit une loi $\mathcal{N}(\mu_2, \sigma^2/n_2)$
- s_1^2 et s_2^2 sont des réalisations des 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 $\chi^2(n_1 + n_2 - 2)$, indépendante de $\overline{X}_1$ et $\overline{X}_2$.

Faisons maintenant l'hypothèse $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu$. Il en résulte que la variable $\overline{X}_1 - \overline{X}_2$ suit une loi normale de *moyenne nulle* et de *variance égale à la somme des variances* de $\overline{X}_1$ et $\overline{X}_2$, c'est-à-dire à $\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Par conséquent,

Théorème 10

Etant données 2 variables $X_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, sous l'hypothèse $\mathcal{H}_0$: $\mu_1 = \mu_2$, on a :

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

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

$$T = \frac{\frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{\frac{n_1 {S_1}^2 + n_2 {S_2}^2}{\sigma^2}}{n_1 + n_2 - 2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 {S_1}^2 + n_2 {S_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\overline{X}_1 - \overline{X}_2}{S^*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

On a alors le résultat suivant :

Théorème 11

Etant données 2 variables $X_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, σ^2 étant inconnu. Sous l'hypothèse $\mathcal{H}_0 : \mu_1 = \mu_2$:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S^*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}(n_1 + n_2 - 2)$$

Par conséquent, sous l'hypothèse d'égalité des moyennes des deux populations considérées, la quantité :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

est la réalisation d'une loi de Student qu'il suffit, pour conclure, de placer par rapport à l'intervalle $\left[-t_{\alpha/2}, t_{\alpha/2}\right]$ correspondant au risque α choisi. Si t n'appartient pas à l'intervalle, on dira que la différence entre les moyennes observées est *significative* au risque α sinon, qu'elle n'est *pas significative*.

3.4 Estimation de la différence des moyennes des populations

Si la différence observée entre les moyennes $\overline{x}_1$ et $\overline{x}_2$ 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 $\overline{X}_1 - \overline{X}_2$ 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 :

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \Delta}{S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}(n_1 + n_2 - 2)$$

On a, par conséquent, au risque α près :

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2} \times s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \Delta < (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2} \times s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Exemple 5

Reprenons l'exemple 4 dans lequel nous supposons maintenant que les valeurs du tableau sont issues de 2 échantillons distincts. Autrement dit, elles ne sont plus appariées. Au risque 5%, y-a-t-il une différence entre les 2 méthodes ? On rappelle ci-dessous les données :

méthode a	1,286	1,251	1,322	1,339	1,250	1,402	1,365	1,437	1,459
méthode b	1,061	1,092	1,363	1,362	1,265	1,278	1,237	1,386	1,352

Nous devons tout d'abord comparer les variances en faisant l'hypothèse : $\mathcal{H}_0 : \sigma_a^2 = \sigma_b^2$. L'hypothèse alternative est $\mathcal{H}_1 : \sigma_a^2 \neq \sigma_b^2$. Il s'agit donc d'un test bilatéral. On a :

$$F = \frac{n_a S_a^2/(n_a - 1)}{n_b S_b^2/(n_b - 1)} = \frac{9S_a^2/8}{9S_b^2/8} = S_a^2/S_b^2 \sim \mathcal{F}(8,8)$$

Au risque 5%, la région critique est donc $f < 0{,}22$ ou $f > 4{,}43$. On calcule $f = s_a^2/s_b^2 = 0{,}00522/0{,}01263 = 0{,}41$, qui ne se trouve pas dans la région critique. On ne peut donc pas rejeter l'hypothèse d'égalité des variances et, pour la suite de l'exercice, on calcule une estimation de la variance commune :

$$s^{*2} = \frac{n_a s_a^2 + n_b s_b^2}{n_a + n_b - 2} = \frac{9 \times 0{,}00522 + 9 \times 0{,}01263}{9 + 9 - 2} = 0{,}01.$$

On fait maintenant l'hypothèse d'égalité des moyennes, $\mathcal{H}_0 : \mu_a = \mu_b$. L'hypothèse alternative est $\mathcal{H}_1 : \mu_a \neq \mu_b$. Sous $\mathcal{H}_0$, on a :

$$T = \frac{\overline{X_a} - \overline{X_b}}{S^* \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} = \frac{\overline{X_a} - \overline{X_b}}{S^* \sqrt{\frac{1}{9} + \frac{1}{9}}} \sim \mathcal{T}(n_a + n_b - 2 = 16).$$

Au risque 5%, la région critique est $|t| > 2{,}12$. On calcule :

$$t = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{s^{*2}(\frac{1}{n_a} + \frac{1}{n_b})}} = \frac{1{,}346 - 1{,}266}{\sqrt{0{,}01 \times \frac{2}{9}}} = 1.68.$$

t étant inférieur à 2,12, il n'est donc pas dans la région critique. On ne peut donc pas rejeter l'hypothèse $\mathcal{H}_0$. On en conclu qu'il n'y a pas de différence significative entre les 2 moyennes. Contrairement à ce qu'on avait obtenu avec le test des appariements à l'exemple 4.

Ce résultat illustre ainsi en effet que le test des appariements, quand il peut être mis en oeuvre, est plus puissant que le test de comparaison des des moyennes précédé du test de comparaison des variances.

4 Tests d'ajustement

Nous avons vu dans les thèmes précédents certaines lois de probabilité susceptibles de constituer des modèles pour les populations de références. Il s'agit maintenant, en présence d'observations, de choisir le modèle adapté et de vérifier que les observations disponibles s'y raccordent bien.

Le problème, sous la forme la plus générale, consiste à caractériser à partir des données le type de la loi de référence, puis à préciser cette loi par estimation des paramètres qui la définissent complètement. En pratique cependant, on n'opère pas exactement ainsi. Les lois de référence s'identifiant le plus souvent aux lois de probabilité fondamentales (loi binomiale, normale, log-normale, ...), il s'avère plus simple de :

— rapprocher la distribution examinée de la loi de probabilité à laquelle il semble intuitivement (ou pour des raisons théoriques) qu'elle doive se raccorder ;

— vérifier ensuite la validité du rapprochement ainsi opéré.

Lorsque l'ajustement à l'une des lois fondamentales s'avère injustifié, il y a lieu de faire appel à d'autres lois de référence, et il en existe un nombre considérable (loi gamma, loi beta, loi de Pareto, loi de Gumbel, loi de Weibull, ...), ou d'en créer éventuellement pour la circonstance.

4.1 Détermination du type de la loi de référence

Il n'y a pas de recette particulière pour déterminer le type de la loi de référence à laquelle on soupçonne la distribution observée de se rattacher. En général, on se laisse guider par des considérations logiques ou bien on tente des rapprochements qui semblent résulter de la forme des distributions observées.

Dans le cas de distributions relatives à des variables discrètes, l'ajustement à une loi uniforme discrète, une loi binomiale ou une loi de Poisson s'impose de prime abord.

Dans le cas de variables continues, l'ajustement aux lois normale ou log-normale s'avère très souvent, mais pas toujours, légitime. Pour vérifier, avant tout calcul compliqué, que l'hypothèse de tels ajustements n'est pas a priori absurde, on dispose de moyens simples et rapides.

La loi normale est une loi symétrique. De plus, on a vu que l'intervalle $[\mu - u\sigma, \mu - u\sigma]$ comprend approximativement la probabilité 50% pour $u = 2/3$, 68% pour $u = 1$, 95% pour $u = 2$ et presque 100% pour $u = 3$. Donc, si une distribution observée est telle que les fréquences des observations comprises à l'intérieur de ces intervalles sont voisines de ces probabilités, il y a présomption de normalité.

On peut également vérifier cette présomption à l'aide d'une transformation connue sous l'appellation de *droite de Henry* qui opère une anamorphose de la fonction de répartition d'une loi normale en une droite.

On pourra de même tracer un *diagramme quantiles-quantiles* dans lequel on représente les quantiles de la distribution observée en fonction des mêmes quantiles d'une loi normale dont les paramètres sont estimés à partir de la distribution observée. Si la distribution est proche d'une loi normale, ce diagramme prend la forme d'une droite d'équation $y = x$. On pourra précéder de même avec d'autres lois de probabilité.

4.2 Estimation des paramètres de la loi de référence

La loi de référence dépend le plus souvent d'un certain nombre de paramètres qu'il est nécessaire d'estimer pour la définir complètement. Une loi binomiale est entièrement définie par la proportion p à laquelle elle correspond (n étant connu). Une loi normale est entièrement définie par sa moyenne μ et son écart-type σ. Il convient donc, à partir des données disponibles, d'estimer soit la proportion p, soit la moyenne μ et l'écart-type σ de la loi de référence binomiale ou normale, pour ne considérer que ces deux exemples.

4.3 Vérification de la légitimité d'un ajustement effectué

Soit un échantillon de taille n d'une variable aléatoire X, discrète ou discrétisée, fournissant les effectifs aléatoires $N_1, N_2, \ldots, N_k$ dans k classes. On se propose de tester l'hypothèse $\mathcal{H}_0$: « la variable aléatoire X suit la loi $\mathcal{L}$ », la loi de référence dont le choix a été discuté précédemment. Une fois les paramètres de cette loi de référence estimés, il est possible de calculer les probabilités théoriques $p_1, p_2, \ldots, p_k$ pour les k classes et d'en déduire les effectifs théoriques, égaux à np_i pour la classe i.

La comparaison des effectifs observés et théoriques met en évidence des différences plus ou moins fortes. Cela n'a rien d'étonnant puisque, dans l'hypothèse où l'ajustement opéré est justifié, la distribution des effectifs théoriques np_i n'est que la loi limite de la distribution des N_i. Il reste toutefois à savoir si les différences ainsi mises en évidence sont compatibles avec les seuls aléas de l'échantillonnage. Ce n'est en effet qu'à cette condition qu'on peut considérer l'ajustement opéré comme légitime. La vérification consiste à déterminer la loi d'une certaine fonction de l'ensemble des fluctuations entre effectifs observés et théoriques, *dans l'hypothèse où ces fluctuations ne sont effectivement dues qu'aux aléas de l'échantillonnage*. Dans ces conditions, il est relativement aisé de montrer que :

Théorème 12

Si n est grand, alors la variable :

$$D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

suit une loi du χ^2 à $k-1-\ell$ degrés de liberté.

Ici, ℓ est le nombre de paramètres qu'il a fallu estimer pour le calcul des effectifs théoriques.

À un seuil de probabilité α faible, pouvant être considéré comme négligeable, correspond une valeur c (ou χ^2_α ou d^2_α) telle que la probabilité d'observer $D^2 > c$ soit égale à α. Si justement la valeur observé d^2 est supérieure à c, il parait préférable de mettre en doute l'hypothèse de la légitimité de l'ajustement. Si, au contraire, la valeur de d^2 est inférieure à c, il n'y a pas de raison de mettre en doute cette hypothèse. Cela ne signifie malheureusement pas qu'elle soit vraie. Or, ce que l'on souhaiterait généralement, c'est confirmer la validité du modèle envisagé. L'aspect négatif du test statistique est gênant dans ce cas précis, dans le sens où il ne prend pas en compte le risque de conserver à tort l'hypothèse faite.

Notez que l'on a effectué un test unilatéral à droite, puisque ce sont des écarts importants entre effectifs observés et théoriques que l'on veut éventuellement détecter, donc une grande valeur du χ^2. Inversement, une distance trop petite (même si elle confirme un très bon ajustement) fera peser une doute sur la qualité de l'échantillonnage.

Ajoutons enfin une dernière remarque sur la mise en oeuvre du test. Pour que la loi de la quantité D^2 soit suffisamment voisine d'une loi du χ^2, il faut non seulement que n soit assez grand, mais encore que les effectifs théoriques np_i ne soient pas trop petits : en pratique, ils ne doivent pas être inférieurs à 5. Si certains d'entre eux sont trop petits, il est nécessaire de procéder à des regroupements de classes.

Exemple 6

Dans le contrôle de la fabrication d'un circuit intégré, on mesure la tension de sortie basse (V) en mV pour une entrée de 5 V. Le résultat des mesures effectuées sur un échantillon de taille 300 est représenté par le tableau ci-dessous :

V (mV)	*45 à 46*	*46 à 47*	*47 à 48*	*48 à 49*	*49 à 50*
Effectif	*20*	*59*	*136*	*74*	*11*

La distribution mesurée peut-elle être considérée comme suivant une loi normale ?

On trouve $\bar{x} = 47{,}49\ mV$ et $s^* = 0{,}927\ mV$. On se propose de tester l'hypothèse $\mathcal{H}_0$: « la variable aléatoire V suit la loi $\mathcal{N}(\bar{x}, s^{*2})$ ». Dans ce cas, on peut écrire, après centrage et réduction de la variable V :

$$p_1 = \mathbb{P}(45 < V \leq 46) = \mathbb{P}(-2{,}69 < U \leq -1{,}61) \simeq 5\%$$

où U est la variable centrée réduite qui suit la loi $\mathcal{N}(0,1)$. On peut ainsi construire le tableau suivant, où les effectifs théoriques sont calculés en multipliant, dans chaque case,

la probabilité p_i par l'effectif total observé $n = \sum_i N_i = 300$.

V (mV)	45 à 46	46 à 47	47 à 48	48 à 49	49 à 50
N_i	20	59	136	74	11
p_i	5,03%	24,46%	41,03%	23,94%	4,82%
np_i	15,11	73,37	123,10	71,83	14,49

Comme les deux paramètres de la loi normale ont été estimés à partir des mesures, la variable D^2, définie dans le théorème 12, suit une loi du χ^2 à $5-1-2 = 2$ degrés de liberté. Finalement, on obtient $d^2 = 6{,}65$ et, à un seuil de probabilité $\alpha = 5\%$, $c = 5{,}99$. Comme nous l'avons vu précédemment, il parait donc préférable, au seuil de 5%, et puisque $d^2 > c$, de mettre en doute l'hypothèse $\mathcal{H}_0$ selon laquelle la distribution mesurée serait considérée comme suivant une loi normale.

5 Tests d'indépendance

Dans les tests vus à la section 3, nous nous sommes placés dans le cadre de populations décrites par des variables normales qui pouvaient donc être caractérisée par deux paramètres : leur moyenne et leur variance. Or, bien souvent, on est amené à prendre en considération des variables dont on ignore la loi de distribution. Pour lever cette difficulté, on s'est donc préoccupé de définir des tests, dits *non paramétriques*, ne faisant aucune hypothèse sur la nature des populations mises en jeu. Il existe une très grande variété de tels tests non paramétriques, dont certains reposent sur la prise en compte d'une même quantité suivant une loi du χ^2 que celle vue au test présenté à la section 4.

Dans la présente section, nous allons nous intéresser à la présentation du test d'indépendance entre deux variables. D'autres tests non paramétriques seront abordés à la section 6.

Soit $x_1, \ldots, x_i, \ldots, x_r$ et soit $y_1, \ldots, y_j, \ldots, y_s$ les modalités de deux variables X et Y. Un échantillon de n individus sur lesquels ont été repérées les valeurs prises simultanément par les deux variables a donné les résultats ci-dessous, qui se présentent sous la forme d'un tableau comprenant r lignes et s colonnes :

	y_1	$\ldots$	y_j	$\ldots$	y_s	total
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1s}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{is}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\vdots$	n_{rj}	$\vdots$	n_{rs}	$n_{r.}$
total	$n_{.1}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.s}$	n

On introduit les notations suivantes :

— n_{ij} est le nombre d'individus ayant présenté à la fois la modalité x_i de X et la modalité y_j de Y

— $n_{i.} = \sum_{j=1}^{s} n_{ij}$ représente le total de la ligne x_i

— $n_{.j} = \sum_{i=1}^{r} n_{ij}$ représente le total de la colonne y_j

Soient les probabilités suivantes :

— $p_{ij} = \mathbb{P}(X = x_i \text{ et } Y = y_j)$, la probabilité, pour un individu choisi au hasard, de se trouver dans la case (i,j) du tableau ;
— $p_{i.} = \mathbb{P}(X = x_i)$, la probabilité de posséder la modalité x_i de la variable X ;
— $p_{.j} = \mathbb{P}(Y = y_j)$, la probabilité de posséder la modalité y_j de la variable Y.

Nous ne connaissons pas ces probabilités, mais elles peuvent être estimées à partir des données de nos tableaux. En effet, grâce à la loi des grands nombres, on peut estimer $p_{i.}$ par la proportion d'individus ayant la modalité x_i, c'est-à-dire $n_{i.}/n$. De même, on peut estimer $p_{.j}$ par le rapport $n_{.j}/n$.

Faisons alors l'hypothèse $\mathcal{H}_0$: « les deux variables sont indépendantes ». Il s'ensuit que :

$$p_{ij} = p_{i.} \times p_{.j}$$

Sous l'hypothèse d'indépendance, l'effectif théorique t_{ij} correspondant à l'effectif observé n_{ij} est alors égal à :

$$t_{ij} = p_{ij} \times n = \frac{n_{i.} \times n_{.j}}{n}$$

Théorème 13

Si n est grand, et notant N_{ij} les variables dont les n_{ij} sont les réalisations, alors la variable :

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - t_{ij})^2}{t_{ij}}$$

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

Comme précédemment, pour un risque α choisi faible, on définit une valeur limite c telle que $\mathbb{P}(D^2 > c) = \alpha$. Si la valeur observée d^2 est supérieure à c, il parait préférable de mettre en doute l'hypothèse d'indépendance. Si, au contraire, la valeur de d^2 est inférieure à c, il n'y a pas de raison de la rejeter.

Exemple 7

Dans une usine, on a remplacé la commande manuelle de quelques presses par une commande automatique. On désire voir si cette modification a une influence sur les accidents du travail. On a relevé, pendant une période donnée, le nombre d'ouvriers qui ont eu ou non des accidents et on les a classés suivant qu'ils travaillaient sur des presses à commande manuelle ou à commande automatique. On a obtenu les résultats suivants :

	manuelle	*automatique*
accidentés	25	23
non accidentés	183	112

La modification du type de commande a-t-elle une influence sur le nombre d'accidents ?

Nous faisons l'hypothèse $\mathcal{H}_0$: « les deux variables sont indépendantes ». Sur la base du tableau précédent qui contient les effectifs n_{ij}, nous pouvons construire le tableau suivant

des effectifs théoriques t_{ij} :

t_{ij}	manuelle	automatique	total
accidentés	29,1	18,9	48
non accidentés	178,9	116,1	295
total	208	135	343

La variable D^2, définie dans le théorème 13, suit une loi du χ^2 à $(2-1)(2-1) = 1$ degré de liberté. Finalement, on obtient $d^2 = 1{,}72$ et, à un seuil de probabilité $\alpha = 5\%$, $c = 3{,}84$. Au seuil de 5%, et puisque $d^2 < c$, on ne peut pas rejeter l'hypothèse d'indépendance entre le type de commande et le nombre d'accidents.

6 Autres tests non paramétriques

Il existe une très grande variété d'autres tests non paramétriques que ceux vus à la section 4 et à la section 5. Nous nous limiterons ici à la présentation de ceux qui sont les plus utilisés et qui se trouvent reposer sur la prise en compte d'une même quantité suivant une loi du χ^2 vue précédemment.

6.1 Test de comparaison de plusieurs populations qualitatives

Soient r populations $\mathcal{P}_1, \ldots, \mathcal{P}_i, \ldots, \mathcal{P}_r$, dont les individus sont distingués suivant s catégories $C_1, \ldots, C_j, \ldots, C_s$, qui peuvent être les modalités d'une variable qualitative (ou les classes d'une variable quantitative). Pour deux lots de pièces, par exemple, classées en bonnes ou mauvaises, on a $r = 2$ et $s = 2$.

On a prélevé un échantillon dans chacune de ces populations. Soient $n_1, \ldots, n_i, \ldots, n_r$, leurs tailles et soit n_{ij} le nombre d'individus qui proviennent de la population $\mathcal{P}_i$ et qui appartiennent à la catégorie C_j.

	C_1	$\cdots$	C_j	$\cdots$	C_s	Total
$\mathcal{P}_1$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	n_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathcal{P}_i$	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	n_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathcal{P}_r$	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	n_r

Si l'on fait l'hypothèse que les populations sont *identiques*, alors les probabilités d'appartenir à chacune des classes sont les mêmes pour toutes les populations, soit p_1, …, p_j, …, p_s, et l'on peut définir des effectifs théoriques dans chaque classe et pour chaque population : $t_{ij} = n_i p_j$ pour la classe C_j de la population $\mathcal{P}_i$.

Théorème 14

On montre alors que sous l'hypothèse $\mathcal{H}_0$: « les populations sont identiques », on a :

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - t_{ij})^2}{t_{ij}} \sim \chi^2((r-1)(s-1))$$

les n_{ij} étant les réalisations des variables N_{ij}.

On peut alors mettre en oeuvre le même test que celui vu à la section 5.

6.2 Test de la médiane

Etant donnés les résultats fournis par deux échantillons de taille n_1 et n_2 :

échantillon 1	x_1	x_2	$\cdots$	x_{n_1}
échantillon 2	y_1	y_2	$\cdots$	y_{n_1}

Arrangeons l'ensemble de ces résultats selon une même suite croissante, par exemple :

$$x_1,y_1,y_2,x_2,y_3,x_3,\ldots$$

et désignons la médiane de cette suite par M.

Après décompte des observations au-dessus et en dessous de M, le tableau des données peut être résumé ainsi :

Effectifs observés	$> M$	$< M$	Total
échantillon 1	n_{11}	n_{12}	n_1
échantillon 2	n_{21}	n_{22}	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	n_1+n_2

Dans l'hypothèse où les deux populations sont identiques, la proportion théorique des observations au-dessus et en dessous de la médiane est dans tous les cas 1/2. Au tableau précédent correspond le tableau théorique ci-après, et on est en définitive ramené à un test du χ^2 avec 1 degré de liberté comme celui de la section 5.

Effectifs théoriques	$> M$	$< M$	Total
échantillon 1	$\frac{n_1}{2}$	$\frac{n_1}{2}$	n_1
échantillon 2	$\frac{n_2}{2}$	$\frac{n_2}{2}$	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	n_1+n_2

6.3 Test des signes

Ce test s'applique à des observations appariées. Sur un même individu i on a effectué deux mesures x_i et y_i et on s'intéresse aux différences $d_i = y_i - x_i$. Dans le test classique, on prenait en compte les valeurs de ces différences, mais dans le test des signes on ne retiendra que les signes, *plus* ou *moins*, de ces différences. Il y a donc une perte d'information.

S'il n'y a pas de différence entre les mesures, la probabilité d'un signe *plus* est égale à celle d'un signe *moins* et égale à 0,5. S'il y a n individus dans l'échantillon, les effectifs théoriques sont égaux à $0{,}5n$ et on est encore ramené à un test du χ^2 avec 1 degré de liberté sur la quantité :

$$\frac{(n_+ - 0{,}5n)^2}{0{,}5n} + \frac{(n_- - 0{,}5n)^2}{0{,}5n}$$

On peut aussi mettre en oeuvre ce test en mobilisant une loi binomiale (voir Saporta. En effet, si on note K le nombre de différences positives. Sous l'hypothèse nulle d'absence de différences entre les moyennes des deux populations, il y a une chance sur deux qu'une différence soit positive ou négative. K suit donc une loi binomiale $\mathcal{B}(n; 0{,}5)$. Il suffit alors

de vérifier que la valeur calculée k appartient ou non à la région critique, cette dernière étant définie par les quantiles de la loi $\mathcal{B}(n; 0{,}5)$ aux niveaux $\alpha/2$ et $(1-\alpha/2)$.

Saporta [1] présente également le test des rangs signés de Wilcoxon, plus puissant que le test des signes pour les données appariées et insuffisament connu. On pourra en retrouver une description en anglais sur internet.

7 Analyse de la variance

Imaginons le cas suivant : un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament (une variable qualitative) sur la durée de vie des ampoules fabriquées (une variable quantitative). Pour ce faire, il va faire fabriquer 4 échantillons de plusieurs ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus. La technique statistique permettant cette comparaison est appelée l'analyse de la variance. L'objectif de cette section est donc de présenter la technique de l'analyse de la variance pour l'étude de l'influence d'un facteur, puis de plusieurs facteurs.

Comme nous allons le voir, cette technique constitue une extension du test de comparaison de moyennes que nous avons vu à la section 3 mais appliquée au cas de plus de 2 populations normales.

7.1 Recherche de l'influence d'un facteur

Nous noterons A le facteur et appellerons $A_1, \ldots, A_i, \ldots, A_p$ ses p modalités. Le problème est l'étude de l'influence du facteur A sur la variable quantitative Y. L'expérimentation disponible a consisté à réaliser, pour chaque modalité A_i du facteur, un certain nombre n_i de mesures de la variable Y étudiée de sorte de disposer d'un tableau comme le suivant, où $\overline{y_1}, \ldots, \overline{y_i}, \ldots, \overline{y_p}$ sont les moyennes des colonnes :

A_1	$\ldots$	A_i	$\ldots$	A_p
y_{11}		y_{i1}		y_{p1}
$\vdots$		$\vdots$		$\vdots$
$\vdots$		y_{ij}		$\vdots$
y_{in_1}		$\vdots$		$\vdots$
		$\vdots$		y_{pn_p}
		y_{in_i}		
$\overline{y_1}$	$\ldots$	$\overline{y_i}$	$\ldots$	$\overline{y_p}$

On pourra plus généralement interpréter les colonnes du tableau comme des échantillons issus de plusieurs populations qu'il s'agit de comparer.

7.2 La relation d'analyse de la variance

Appelons $\bar{y}$ la moyenne générale des mesures :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij} \text{ avec } n = \sum_{i=1}^{p} n_i$$

1. Gilbert Saporta, *Probabilités, analyse des données et Statistique*, Editions Technip, 2006

Effectuons alors la décomposition :

$$(y_{ij} - \bar{y}) = (\overline{y_i} - \bar{y}) + (y_{ij} - \overline{y_i})$$

En élevant au carré et en sommant, le double produit est nul. En effet, par définition des moyennes $\overline{y_i}$, on peut écrire :

$$2\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{y_i} - \bar{y})(y_{ij} - \overline{y_i}) = 2\sum_{i=1}^{p}(\overline{y_i} - \bar{y})\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i}) = 0$$

On obtient donc :

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} n_i (\overline{y_i} - \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

relation appelée d'analyse de la variance, qui décompose la somme des carrés totale :

$$\text{SCT}^{\,2} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

en une somme des carrés mesurant la variabilité *intercolonnes* (c'est-à-dire l'influence du facteur ou la différence entre les populations) :

$$\text{SCA}^{\,3} = \sum_{i=1}^{p} n_i (\overline{y_i} - \bar{y})^2$$

et une somme des carrés mesurant la variabilité *intracolonnes* (somme des carrés résiduelle) :

$$\text{SCR}^{\,4} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

Notons la grande généralité de cette relation puisqu'elle a été établie sans faire aucune hypothèse sur les données. Cependant, la structure de la relation de base : $(y_{ij} - \bar{y}) = (\overline{y_i} - \bar{y}) + (y_{ij} - \overline{y_i})$ revient à admettre implicitement l'*additivité* de l'influence du facteur $(\overline{y_i} - \bar{y})$ et d'un résidu $(y_{ij} - \overline{y_i})$.

On peut associer des degrés de liberté à ces sommes qui correspondent au nombre de termes indépendants dans chacune d'elles :

— pour SCT : $(n-1)$ degrés de liberté (nombre de valeurs - 1)
— pour SCA : $(p-1)$ degrés de liberté (nombres de modalités du facteur - 1)
— pour SCR : $(n-p)$ par différence des valeurs précédentes $(n-1)-(p-1) = (n-p)$

7.3 Le modèle

Pour permettre l'inférence statistique, il est nécessaire de poser un certain nombre d'hypothèses. Le modèle de base de l'analyse de la variance s'écrit :

$$Y_i = \mu_i + \varepsilon_i = \mu + \alpha_i + \varepsilon_i$$

2. SCT : Somme des Carrés Totale
3. SCA : Somme des Carrés expliqués par le facteur A ou par les variations du facteur A
4. SCR : Somme des Carrés Résiduelle

Les α_i sont des quantités inconnues, mais certaines, qui mesurent l'influence du facteur A. Pour lever leur indétermination à une constante près, on a l'habitude de poser :

$$\sum_{i=1}^{p} n_i \alpha_i = 0$$

Les ε_i représentent les fluctuations aléatoires correspondant aux erreurs de mesure ou à l'influence des facteurs non contrôlés. Nous poserons qu'il n'y a pas d'erreur systématique, ou qu'elle est contenue dans μ, donc que $\mathbb{E}(\varepsilon_i) = 0$.

Les hypothèses suivantes stipulent que les ε_i :

— sont indépendants, ce qui entraîne : $\mathbb{C}\text{ov}(\varepsilon_i,\varepsilon_j) = 0$ pour $i \neq j$
— ont même variance (homoscédasticité) : $\forall i, \mathbb{V}(\varepsilon_i) = \sigma^2$
— suivent des lois normales.

Parmi ces hypothèses, la plus restrictive est certainement la seconde d'après laquelle l'erreur sur la variable Y est indépendante de la valeur prise par Y, c'est-à-dire notamment, qu'elle n'est pas de type multiplicatif. Pour vérifier si elle est légitime, on dispose de plusieurs tests dont le plus connu est celui de *Bartlett* mais ce dernier est très sensible à l'hypothèse de normalité.

7.4 Test d'analyse de la variance

Faisons l'hypothèse $\mathcal{H}_0$ que le facteur A n'a pas d'influence sur la variable Y. Cela signifie que les Y_i ont toutes la même moyenne μ et donc que $\alpha_1 = \cdots = \alpha_i = \cdots = \alpha_p = 0$. Sous $\mathcal{H}_0$, on peut alors montrer que $\frac{\text{SCA}}{\sigma^2} \sim \chi^2(p-1)$. Cela implique que $\frac{\text{SCA}}{p-1}$ est un estimateur de σ^2. Comme d'autre part $\frac{\text{SCR}}{\sigma^2} \sim \chi^2(n-p)$, la variable $\frac{\text{SCR}}{n-p}$ est aussi un estimateur de σ^2. Il en résulte que :

Théorème 15

Sous $\mathcal{H}_0 : \forall i, \alpha_i = 0$ (absence d'influence du facteur A), on a :

$$F = \frac{SCA/(p-1)}{SCR/(n-p)} \sim \mathcal{F}(p-1,n-p)$$

et qu'à ce titre, sa valeur est proche de 1.

Si la valeur f calculée est supérieure au seuil f_α lu dans la table de Fisher-Snedecor et tel que $\mathbb{P}(F > f_\alpha) = \alpha$, on pourra rejeter l'hypothèse $\mathcal{H}_0$ au risque α de la rejeter à tort et conclure en faveur de l'hypothèse alternative $\mathcal{H}_1 : \exists i,j, \alpha_i \neq \alpha_j$ qui traduit une influence du facteur A. Si elle est inférieure, l'information disponible ne permet pas de conclure à une influence du facteur A. Il importera d'effectuer un *test unilatéral à droite*. En effet, les faibles valeurs de f correspondent à des différences faibles entre les moyennes $\overline{y_i}$ des colonnes, alors que le test vise à mettre en évidence des différences fortes.

7.5 Calcul pratique

On calcule :

$$\text{SCT} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}^2 - n\overline{y}^2$$
$$\text{SCA} = \sum_{i=1}^{p} n_i(\overline{y_i}-\overline{y})^2 = \sum_{i=1}^{p} n_i\overline{y_i}^2 - n\overline{y}^2$$

et, par différence :

$$\text{SCR} = \text{SCT} - \text{SCA}$$

On constitue alors le tableau suivant :

Variation	SC	Degrés de liberté	f calculé	Valeur critique
Facteur	SCA	$p-1$	$\frac{\text{SCA}/(p-1)}{\text{SCR}/(n-p)}$	$f_\alpha(p-1,n-p)$
Résiduelle	SCR	$n-p$		
Totale	SCT	$n-1$		

Exemple 8

Le tableau suivant donne la durée de vie (en heures, au delà de 1000 heures) de plusieurs échantillons d'ampoules de 60W provenant de 3 marques différentes, la mesure ayant été réalisée dans des conditions contrôlées et identiques pour toutes les ampoules.

Marque 1	16	15	13	21	15	$\overline{y_1}=16$	$s_1^2=7{,}2$
Marque 2	18	22	20	16	24	$\overline{y_2}=20$	$s_2^2=8$
Marque 3	26	31	24	30	24	$\overline{y_1}=27$	$s_3^2=8{,}8$

Faisant l'hypothèse que la durée de vie des ampoules obéit à une loi normale, les variances étant les mêmes d'une marque à l'autre, peut-on considérer qu'il y a une différence significative de durée de vie entre les marques, au risque 1%.

On calcule :
- $\text{SCT} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}^2 - n\overline{y}^2 = 7045 - 15\times 21^2 = 430$
- $\text{SCA} = \sum_{i=1}^{p} n_i(\overline{y_i}-\overline{y})^2 = 5(16-21)^2 + 5(20-21)^2 + 5(27-21)^2 = 310$
- $\text{SCR} = \text{SCT} - \text{SCA} = 430 - 310 = 120$

Puis on construit le tableau d'analyse de la variance classique :

Variation	SC	Degrés de liberté	f calculé	Valeur critique
Facteur	310	$3-1=2$	$\frac{\text{SCA}/(p-1)}{\text{SCR}/(n-p)} = \frac{310/2}{120/12} = 15{,}5$	$f_{1\%}=6{,}92$
Résiduelle	120	$15-3=12$		
Totale	430	$15-1=14$		

Au risque 1%, la valeur de f calculée est supérieure à la valeur critique. On peut donc rejeter l'hypothèse que les 3 marques sont équivalentes du point de vue de la durée de vie des ampoules et conclure en faveur d'une influence du facteur étudié, ici la marque.

8 Etude de l'influence de deux facteurs

Imaginons que le fabricant d'ampoules évoqué plus haut, se préoccupe d'étudier l'influence, sur la durée de vie des ampoules, non seulement du type de filament utilisé, mais également de la nature du gaz de remplissage.

Il pourrait évidemment faire, d'une part, une première étude ⟨filament⟩ en utilisant l'analyse de la variance à un facteur, puis procéder, d'autre part, à une étude ⟨gaz⟩ en tous points analogue. Cela fait, il lui resterait à rapprocher les résultats de ces deux études pour se faire une idée de l'influence des deux facteurs étudiés. Mais en procédant de la sorte, il postulera implicitement l'additivité des influences ⟨filament⟩ et ⟨gaz⟩, ce qui n'est pas acquis.

L'analyse de la variance à deux facteurs va permettre de traiter globalement le problème, et de mettre éventuellement en évidence ce qu'il est convenu d'appeler les *interactions* des facteurs étudiés.

8.1 Plan factoriel

Soit, d'une façon générale, A et B les deux facteurs dont on se propose d'étudier l'influence sur une variable quantitative Y. Nous appellerons $A_1,\ldots,A_i,\ldots,A_p$, les p modalités du facteur A, et $B_1,\ldots,B_j,\ldots,B_q$, les q modalités du facteur B. La mise en oeuvre de l'analyse de la variance à deux facteurs nécessite de disposer d'au moins une mesure de Y pour toute combinaison (A_i,B_j) des modalités des facteurs.

Nous admettrons que l'expérimentation a permis de réaliser r répétitions, c'est-à-dire r mesures pour chacune des pq combinaisons des modalités des facteurs. Le cas où il n'y a pas de répétitions ($r = 1$) fera l'objet d'un paragraphe particulier à la section 8.5

Les essais sont donc menés de façon à obtenir le tableau de mesures ci-dessous, une des difficultés de l'expérimentation étant d'éviter les mesures manquantes.

	A_1	$\cdots$	A_i	$\cdots$	A_p
B_1	$y_{111}\cdots y_{11r}$	$\cdots$	$y_{i11}\cdots y_{i1r}$	$\cdots$	$y_{p11}\cdots y_{p1r}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_j	$y_{1j1}\cdots y_{1jr}$	$\cdots$	$y_{ij1}\cdots y_{ijr}$	$\cdots$	$y_{pj1}\cdots y_{pjr}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_q	$y_{1q1}\cdots y_{1qr}$	$\cdots$	$y_{iq1}\cdots y_{iqr}$	$\cdots$	$y_{pq1}\cdots y_{pqr}$

Le plan d'expérience ainsi réalisé est appelé *plan factoriel*. Il est dit *équilibré* parce qu'il y a le même nombre de mesures dans chaque case du tableau. Il existe d'autres plans d'expériences équilibrés qui évitent le principal inconvénient du plan factoriel, qui est d'être très coûteux du point de vue du nombre de mesures à effectuer.

8.2 Modèle additif et modèle avec interaction

Le modèle le plus général, en admettant l'additivité des erreurs, est le suivant :

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

En explicitant μ_{ij}, un modèle couramment utilisé est le modèle additif :

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

On suppose ainsi qu'il y a *additivité* des effets : l'action conjuguée des modalités A_i et B_j est la somme des actions isolées de A d'une part et de B d'autre part.

Si l'on ne suppose pas réalisée cette hypothèse restrictive d'additivité, on adopte le modèle avec *interaction* :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Il n'y a plus additivité des effets car, aux actions directes de A et B, s'ajoute le terme γ_{ij} qui traduit un effet supplémentaire du à la conjonction des modalités A_i et B_j.

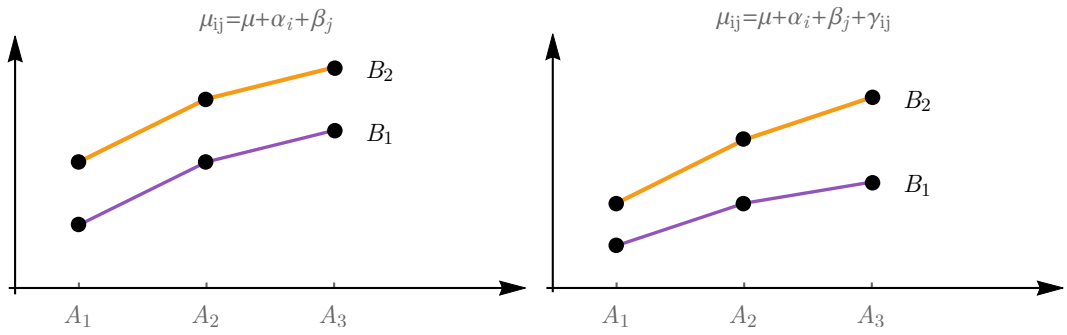

FIGURE 8 – Additivité seulement (à gauche), additivité avec interaction (à droite)

On dit que $(\alpha_1, \ldots, \alpha_p)$ et $(\beta_1, \ldots, \beta_q)$ sont les actions des facteurs A et B, tandis que $(\gamma_{11}, \ldots, \gamma_{pq})$ sont les interactions du couple (A,B). On peut encore dire que le modèle avec interaction traduit le fait que l'action du facteur A, par exemple, dépend des modalités du facteur B, comme l'illustre la figure 8.

Pour lever l'indétermination de μ, on pose les relations suivantes :

$$\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = 0$$

$$\sum_{i=1}^{p} \gamma_{ij} = 0 \text{ pour tout } j \quad \text{et} \quad \sum_{j=1}^{q} \gamma_{ij} = 0 \text{ pour tout } i$$

8.3 Relation d'analyse de la variance

Appelons :

— $\overline{y_i}$ la moyenne d'une colonne du tableau des mesures : $\overline{y_i} = \frac{1}{qr} \sum_{jk} y_{ijk}$

— $\overline{y_j}$ la moyenne d'une ligne du tableau : $\overline{y_j} = \frac{1}{pr} \sum_{ik} y_{ijk}$

— $\overline{y_{ij}}$ la moyenne d'une case du tableau : $\overline{y_{ij}} = \frac{1}{r} \sum_{k} y_{ijk}$

— $\bar{y}$ la moyenne générale des mesures : $\bar{y} = \frac{1}{pqr} \sum_{ijk} y_{ijk}$

Effectuons alors la décomposition :

$$(y_{ijk} - \bar{y}) = (\overline{y_i} - \bar{y}) + (\overline{y_j} - \bar{y}) + [(\overline{y_{ij}} - \bar{y}) - (\overline{y_i} - \bar{y}) - (\overline{y_j} - \bar{y})] + (y_{ijk} - \overline{y_{ij}})$$

En élevant au carré et en sommant, les doubles produits s'annulent par définition des différentes moyennes, à la condition stricte que le tableau soit complet, c'est-à-dire qu'il

n'y ait aucune mesure manquante. On obtient par conséquent :

$$\sum_{ijk} (y_{ijk} - \bar{y})^2 = qr \sum_i (\overline{y_i} - \bar{y})^2 + pr \sum_j (\overline{y_j} - \bar{y})^2 + r \sum_{ij} [(\overline{y_{ij}} - \bar{y}) - (\overline{y_i} - \bar{y}) - (\overline{y_j} - \bar{y})]^2 + \sum_{ijk} (y_{ijk} - \overline{y_{ij}})^2$$

que nous noterons symboliquement :

$$\text{SCT} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCR}$$

C'est la relation d'analyse de la variance. Elle permet de décomposer la somme des carrés totale en quatre sommes. Les deux premières correspondent respectivement aux *actions* de A et de B. La troisième correspond à l'*interaction* de A et B. La dernière est la somme des carrés *résiduelle*.

8.4 Les tests d'analyse de la variance

Admettons, comme dans le cas d'un seul facteur, que les εij sont des variables aléatoires indépendantes suivant toutes la loi $\mathcal{N}(0,\sigma^2)$. Il est alors possible d'effectuer une inférence statistique à partir des observations, et de tester :

— la présence d'une interaction,
— l'influence d'un facteur.

8.4.1 Test de l'interaction

Faisons l'hypothèse qu'il n'y a pas d'interaction des facteurs A et B. Cela signifie que $\forall i,j : \gamma_{ij} = 0$. On montre alors que $\frac{\text{SCAB}}{\sigma^2} \sim \chi^2((p-1)(q-1))$. Comme d'autre part, $\frac{\text{SCR}}{\sigma^2} \sim \chi^2(n - pq = pq(r-1))$, il en résulte que :

Théorème 16

Sous $\mathcal{H}_0 : \forall i,j,\gamma_{ij} = 0$ *(absence d'interaction) on a :*

$$F_{AB} = \frac{SCAB/((p-1)(q-1))}{SCR/(pq(r-1))} \sim \mathcal{F}((p-1)(q-1),pq(r-1))$$

8.4.2 Test de l'influence d'un facteur

Faisons l'hypothèse que le facteur A, par exemple, n'a pas d'influence sur la variable Y. Cela signifie que $\forall i : \alpha_i = 0$. On montre alors que $\frac{\text{SCA}}{\sigma^2} \sim \chi^2(p-1)$. Par conséquent :

Théorème 17

Sous $\mathcal{H}_0 : \forall i,j,\alpha_i = 0$ *(absence d'influence du facteur A), on a :*

$$F_A = \frac{SCA/(p-1)}{SCR/(pq(r-1))} \sim \mathcal{F}(p-1,pq(r-1))$$

8.4.3 Exécution des calculs

On calcule SCA, SCB, SCAB et SCR par les formules suivantes :

$$\text{SCA} = qr \sum_i \overline{y_i}^2 - pqr\bar{y}^2$$

$$\text{SCB} = pr \sum_j \overline{y_j}^2 - pqr\bar{y}^2$$

$$\text{SCAB} = r \sum_{ij} \overline{y_{ij}}^2 - pqr\bar{y}^2 - \text{SCA} - \text{SCB}$$

$$\text{SCT} = \sum_{ijk} {y_{ijk}}^2 - pqr\bar{y}^2$$

puis, par différence :

$$\text{SCR} = \text{SCT} - \text{SCA} - \text{SCB} - \text{SCAB}$$

On dresse enfin le tableau :

Variation	SC	DL	f calculé	valeur critique
Facteur A	SCA	$p-1$	$f_A = \frac{\text{SCA}/(p-1)}{\text{SCR}/(pq(r-1))}$	$f_\alpha(p-1, pq(r-1))$
Facteur B	SCB	$q-1$	$f_B = \frac{\text{SCB}/(q-1)}{\text{SCR}/(pq(r-1))}$	$f_\alpha(q-1, pq(r-1))$
Interaction	SCAB	$(p-1)(q-1)$	$f_{AB} = \frac{\text{SCAB}/((p-1)(q-1))}{\text{SCR}/(pq(r-1))}$	$f_\alpha((p-1)(q-1), pq(r-1))$
Résiduelle	SCR	$pq(r-1)$		
Totale	SCT	$pqr-1$		

Exemple 9

On fait passer un test de connaissance à des personnes novices ou expertes d'un certain domaine. Le test est calibré pour donner une loi normale (variable Y) et on obtient les résultats suivants :

y	*hommes*				*femmes*			
novices	15	16	17	18	26	27	28	29
experts	20	21	22	23	31	32	33	34

$$\textit{avec : } \sum_{i,j} y_{ij} = 392 \textit{ et } \sum_{i,j} y_{ij}^2 = 10208$$

Que peut-on en conclure ?

Les femmes semblent se débrouiller mieux que les hommes, indépendamment du niveau d'expertise. Pour le confirmer, on peut faire une ANOVA [5] à deux facteurs (*genre* et *expertise*) et construire le tableau des moyennes où A désigne le facteur *genre* et B le facteur *expertise* (modalités dans le même ordre que dans le tableau de données) :

$\overline{y_{ij}}$	A_1	A_2	$\overline{y_j}$
B_1	16,5	27,5	22
B_2	21,5	32,5	27
$\overline{y_i}$	19	30	24,5

On peut alors calculer les grandeurs classiques :

5. ANOVA : de l'anglais « ANalysis Of VAriance », c'est-à-dire *Analyse de la variance*

— $\text{SCT} = \sum_{ijk} (y_{ijk} - \overline{y}) = \sum_{ijk} y_{ijk}^2 - n\overline{y}^2 = 10208 - 16 \times \frac{392}{16}^2 = 604$
— $\text{SCA} = \sum_i n_i \left(\overline{y_i} - \overline{y}\right)^2 = 8(19 - 24{,}5)^2 + 8(30 - 24{,}5)^2 = 484$
— $\text{SCB} = \sum_j n_j \left(\overline{y_j} - \overline{y}\right)^2 = 8(22 - 24{,}5)^2 + 8(27 - 24{,}5)^2 = 100$
— $\text{SCAB} = \sum_{ij} n_{ij} \left[(\overline{y_{ij}} - \overline{y}) - (\overline{y_i} - \overline{y}) - (\overline{y_j} - \overline{y})\right]^2 = \sum_{ij} n_{ij} \left[(\overline{y_{ij}} - \overline{y_i} - \overline{y_j} + \overline{y})\right]^2 = 4(16{,}5 - 22 - 19 + 24{,}5)^2 + 4(21{,}5 - 27 - 19 + 24{,}5)^2 + 4(27{,}5 - 22 - 30 + 24{,}5)^2 + 4(32{,}5 - 27 - 30 + 24{,}5)^2 = 0$
— $\text{SCR} = \text{SCT} - \text{SCA} - \text{SCB} - \text{SCAB} = 604 - 484 - 100 - 0 = 20$

Puis on construit le tableau de synthèse suivant :

Variation	SC	DL	f calculé	$f_{5\%}$
Facteur A	484	1	$f_A = \frac{\text{SCA}/1}{\text{SCR}/12} = 290{,}4$	4,74
Facteur B	100	1	$f_B = \frac{\text{SCB}/1}{\text{SCR}/12} = 60$	4,74
Interaction	0	$1 \times 1 = 1$	$f_{AB} = \frac{\text{SCAB}/1}{\text{SCR}/12} = 0$	4,74
Résiduelle	20	$2 \times 2 \times (4 - 1) = 12$		
Totale	SCT	$pqr - 1$		

Les valeurs de f calculées correspondant aux facteurs A et B étant supérieures aux valeurs critiques, on peut conclure, au risque 5%, à l'influence significative de ces deux facteurs sur les résultats au test de connaissance. Par contre, pour l'interaction, le f calculé est inférieur à la valeur critique, ce qui ne permet pas de conclure à une interaction entre les facteurs. En conclusion, avec moins de 5 chances sur 100 de se tromper, on peut dire que les *experts* sont meilleurs que les *novices*, les *femmes* sont meilleures que les *hommes*. Par ailleurs, on ne peut pas rejeter l'hypothèse d'une absence d'interaction entre *expertise* et *genre*.

8.5 Analyse de la variance sans répétitions

Supposons qu'on n'ait réalisé qu'une seule mesure y_{ij} pour chaque couple de modalités (A_i,B_j), conformément au tableau ci-après.

	A_1	$\cdots$	A_i	$\cdots$	A_p
B_1	y_{11}	$\cdots$	y_{i1}	$\cdots$	y_{p1}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_j	y_{1j}	$\cdots$	y_{ij}	$\cdots$	y_{pj}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
B_q	y_{1q}	$\cdots$	y_{iq}	$\cdots$	y_{pq}

L'équation d'analyse de la variance s'écrit alors :

$$\sum_{ij} (y_{ij} - \bar{y})^2 = q \sum_i (\overline{y_i} - \bar{y})^2 + p \sum_j (\overline{y_j} - \bar{y})^2 + \sum_{ij} \left[(y_{ij} - \bar{y}) - (\overline{y_i} - \bar{y}) - (\overline{y_j} - \bar{y})\right]^2$$

soit, avec les notations habituelles :

$$\text{SCT} = \text{SCA} + \text{SCB} + \text{SCAB}$$

Il devient impossible de tester l'interaction, puisqu'on ne dispose plus d'une quantité telle que SCR permettant, par division, d'éliminer σ^2 et d'obtenir une loi de Fisher-Snedecor. Il est donc nécessaire dans ce cas de faire l'hypothèse (impossible à vérifier)

qu'il n'y a pas d'interaction. On doit donc adopter le modèle additif :

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Sous cette condition, et quelles que soient les actions des facteurs A et B, on montre, comme dans le cas général, que $\frac{\text{SCAB}}{\sigma^2} \sim \chi^2((p-1)(q-1))$.

Dès lors, pour tester l'influence de A, par exemple, faisons l'hypothèse que les α_i sont tous nuls. Elle entraîne que $\frac{\text{SCA}}{\sigma^2} \sim \chi^2(p-1)$ et, par conséquent, que :

$$F_A = \frac{\frac{\text{SCA}}{(p-1)}}{\frac{\text{SCAB}}{(p-1)(q-1)}} \sim \mathcal{F}((p-1),(p-1)(q-1))$$

Exemple 10

Un fabricant d'ordinateur souhaite comparer la vitesse de 4 compilateurs. A cet effet, il choisit 5 programmes différents et les fait compiler par chacun des 4 compilateurs. Il obtient les résultats suivants (en millisecondes) où PX désignent les programmes et CX les compilateurs, $\overline{y}$ et s respectivement les moyennes et les écart-types par ligne et par colonne. Que peut-on en conclure au risque 1% ?

	$C1$	$C2$	$C3$	$C4$	$\overline{y}$	s
$P1$	29,21	28,25	28,2	28,62	28,57	0,4036
$P2$	26,18	26,02	26,22	25,56	25,995	0,2621
$P3$	30,91	30,18	30,52	30,09	30,425	0,3227
$P4$	25,14	25,26	25,20	25,02	25,155	0,0887
$P5$	26,16	25,14	25,26	25,46	25,505	0,3951
$\overline{y}$	27,52	26,97	27,08	26,95	27,13	
s	2,1752	1,9554	2.0334	2,0261		2,0620

Avec les notations que nous utilisées dans le présent document, on calcule laborieusement SCT = 85,0388, SCA = 83,0404, SCB = 1,063 et SCAB = 0,9354 où A est le facteur « programmes » et B le facteur « compilateurs ». En l'absence de répétitions, je ne dispose pas d'une somme de carrés résiduelle pour tester l'interaction. Je fais donc l'hypothèse qu'il n'y en a pas et je retiens le modèle linéaire suivant : $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$.

Pour tester l'influence de la variable « programmes », je fais l'hypothèse $\mathcal{H}_0 : \forall i, \alpha_i = 0$, alors la quantité $f_A = \frac{\text{SCA}/(p-1)}{\text{SCAB}/(p-1)(q-1)} = 266{,}326$ (où $p = 5$ et $q = 4$) doit être la réalisation d'une loi $\mathcal{F}(4{,}12)$. Or la région critique au risque 1% est l'intervalle $]5{,}41, +\infty[$. La valeur calculée étant dans la région critique, on rejette l'hypothèse et on conclut à une influence très significative du facteur « programme » sur la durée de compilation.

On teste ensuite l'influence de la variable « compilateurs ». Sous l'hypothèse $\mathcal{H}_0$: $\forall j, \beta_j = 0$, la quantité $f_B = \frac{\text{SCB}/(q-1)}{\text{SCAB}/(p-1)(q-1)} = 4{,}555$ (où $p = 5$ et $q = 4$) doit être la réalisation d'une loi $\mathcal{F}(3{,}12)$. Or la région critique au risque 1% est l'intervalle $]5{,}95, +\infty[$. La valeur calculée n'étant pas dans la région critique, je ne peux pas rejeter l'hypothèse et on conclut que le facteur « compilateurs » n'a pas d'influence significative sur la durée de compilation.

9 Exercices

Exercice 1 : Rappels sur les tests

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 $s^2 = 6$.

1. Calculer une estimation sans biais de σ^2 et en déduire un intervalle de confiance de σ au risque 5%.

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

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

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

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

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

Exercice 2 : Test sur valeurs appariées

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

1. Que peut-on conclure de ces résultats ?

Exercice 3 : Test de comparaison de 2 populations

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

— ligne 1 : $\bar{x} = 2800$ et $\sum(x - \bar{x})^2 = 103600$

— ligne 2 : $\bar{y} = 2680$ et $\sum(y - \bar{y})^2 = 76400$

On admettra que les variances σ_x^2 et σ_y^2 sont inconnues mais égales.

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

2. Quel est l'intervalle de confiance à 95% de la différence ?

Exercice 4 : Test sur des variances

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'extérieur qu'à l'intérieur. 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 $\sum y^2 = 84$.

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

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

Exercice 5 : Test d'ajustement à une loi uniforme discrète

Dans une étude portant sur l'orientation spatiale chez les souris, les animaux de l'expérience ont été placés un par un au centre d'un labyrinthe radiaire comportant 8 allées orientées dans les huit directions de la rose des vents. Chaque animal s'est échappé par l'une de ces allées. Les expériences ont porté sur des souris sauvages récemment capturées en un lieu situé au nord-est du laboratoire. Les répartitions des directions de fuite sont données dans le tableau ci-dessous.

Directions	N	NO	O	SO	S	SE	E	NE
Nombre de souris	26	17	9	2	3	16	33	54

1. Au seuil de 5%, est-ce que le choix de la direction de fuite se fait au hasard ?

Exercice 6 : Test d'ajustement à une loi de Poisson

Soit la variable aléatoire X correspondant au nombre annuel de tués lors de l'absorption d'un médicament destiné aux nouveaux nés. 200 observations faites sur plusieurs années dans différents hopitaux ont donné le tableau suivant.

Nombre de tués x_i	0	1	2	3	4
Effectifs observés n_i	109	65	22	3	1

On souhaite tester pour X l'hypothèse d'une loi de Poisson de paramètre λ.

1. Montrer que l'estimation du paramètre λ est égale à 0,61.

2. Que peut-on conclure sur l'hypothèse au risque 5% ? (remarque : les effectifs des 3 classes les moins élevés seront rassemblés pour que l'effectif résultant soit supérieur à 5).

Exercice 7 : Test d'ajustement à une loi normale

Un correcteur rend ses 100 copies au secrétariat d'un concours. Par souci d'équité, la consigne est de noter les copies de manière telle que la distribution des notes soit normale avec une moyenne de 10 et un écart-type de 4. Le secrétariat a établi la distribution suivante :

Notes	moins de 4	de 4 à 8	de 8 à 12	de 12 à 16	plus de 16
Effectifs	8	25	45	10	12

1. Montrer si la distribution observée est cohérente avec celle visée ($\mu = 10$ et $\sigma = 4$). Pour cela, justifier que le nombre de degrés de liberté est égal à 4.

Exercice 8 : Test d'indépendance entre 2 variables

On étudie l'action d'un insecticide sur une culture. On considère deux parcelles : une est non traitée et sert de témoin, alors que l'autre est soumise à un traitement par l'insecticide. Pour chaque parcelle, on recense le nombre de pieds indemnes, le nombre de pieds malades mais vivants et le nombre de pieds morts.

	témoin	traitée
Malades	516	71
Morts	481	82
Indemnes	437	103

1. Que peut-on conclure sur l'efficacité de l'insecticide ?

Exercice 9 : Test d'indépendance entre 2 variables

Pour tester l'efficacité d'un médicament en injection intraveineuse directe, on forme deux groupes de cent malades. Au premier (le groupe A), on injecte du sérum physiologique, et au second (le groupe B), le médicament en question. Les résultats sont les suivants :

	guéris	non guéris
Groupe A	75	25
Groupe B	65	35

1. Testez au seuil de risque de 5% si la guérison dépend ou non de la prise du médicament.

Exercice 10 : Analyse de la variance à 1 facteur

Un laboratoire utilise 4 thermomètres de façon interchangeable pour faire des mesures de température. Pour étudier si les résultats diffèrent suivant les thermomètres, ces derniers ont été placés dans un récipient maintenu à température constante. Trois lectures ont été faites avec chaque thermomètre. Les résultats en degrés centigrades ont été les suivants :

Thermo 1	Thermo 2	Thermo 3	Thermo 4
0,9	0,3	$-0{,}6$	0
1,2	$-0{,}2$	$-1{,}0$	0,4
0,8	0,1	$-0{,}7$	0,5

On donne également :

$$\sum_{i,j} y_{ij} = 1{,}7 \quad \sum_{i,j} {y_{ij}}^2 = 5{,}29 \quad 3\sum_{i} \overline{y_i}^2 = 4{,}85.$$

1. Peut-on conclure à une influence du thermomètre sur les mesures ?

Exercice 11 : Analyse de la variance à 2 facteurs

On réalise une expérience visant à étudier la déformation de plaques de cuivre en fonction de la teneur en cuivre et de la température de ces plaques. La variable étudiée est une mesure de la déformation. On obtient alors le tableau suivant :

	teneur en cuivre (%)			
température (C)	40	60	80	100
50	17	19	23	29
75	12	15	18	27
100	14	19	22	30
125	17	20	22	30

1. Posez toutes les hypothèses nécessaires à l'analyse

2. Analysez l'influence des facteurs