

PREMIÈRE PARTIE

INTRODUCTION À L'INFÉRENCE STATISTIQUE

Les méthodes de l'inférence statistique ont pour but de fournir des résultats sur des populations entières à partir de mesures statistiques réalisées sur des échantillons convenablement prélevés dans celles-ci. Ceci est d'autant plus intéressant que, dans la pratique, un travail de recherche se réalise le plus souvent sur des échantillons et son intérêt est subordonné au caractère plus ou moins général des conclusions auxquelles il permet d'aboutir.

Certaines conditions d'application dont dépendent les méthodes utilisées doivent être vérifiées au préalable. De même, les échantillons doivent être prélevés selon des règles bien précises, car la manière de choisir les échantillons est aussi importante que la manière de les analyser. A ce propos, il faut se rappeler que la plupart des méthodes de l'inférence statistique sont directement appliquées lorsque l'échantillonnage se réalise de manière aléatoire et simple.

Comme nous le verrons, la généralisation des conclusions obtenues à partir des échantillons aux différentes populations comporte certains risques d'erreurs. Ceux-ci peuvent être évalués en faisant appel à la théorie des probabilités. Le chercheur ne peut donc pas faire un jugement certain, mais seulement avec une probabilité assez importante.

D'une manière générale, les méthodes de l'inférence statistique nous permettent de traiter les problèmes d'estimation de paramètres de populations inconnus et les problèmes de tests d'hypothèses statistiques. Rappelons que les paramètres qu'on cherche à estimer sont des caractéristiques numériques d'une population. Il peut s'agir de la moyenne, de la proportion d'individus ayant une propriété donnée, de la corrélation de deux variables, etc.

1° Les problèmes d'estimation permettent d'obtenir des estimations ponctuelles pour les paramètres de la population à partir d'échantillons convenablement prélevés et de les entourer éventuellement d'intervalles de confiance (chapitre 1). C'est ainsi que, par exemple, à partir de la moyenne calculée à partir des prélèvements de sang réalisés sur un échantillon de bovins adultes d'une race donnée et conduits dans les mêmes conditions, on cherche à estimer la teneur moyenne théorique en protéines sériques totales du sang des bovins adultes de toute la population de la région, et à évaluer la précision de l'estimation obtenue.

2° Quant aux problèmes des tests d'hypothèses, ils ont pour but de contrôler, à partir de l'examen d'un ou de plusieurs échantillons, la validité d'hypothèses relatives à une ou plusieurs populations (chapitre 2). Les hypothèses portent sur les valeurs théoriques, tandis que les valeurs observées sur les échantillons servent à les tester. C'est le cas, par exemple, lorsqu'on souhaite comparer les teneurs moyennes en bêta globuline du sang de vaches adultes d'une race donnée chez les vaches indemnes et les vaches atteintes d'une certaine maladie. C'est aussi le cas lorsqu'on souhaite tester si l'oxygène total consommé par les truites saumonées dans des conditions standards varie avec leur poids frais de la même

manière dans deux bassins à 10 °C, le premier est soumis à courant d'eau fort et le second à courant d'eau faible.

Les deux premiers chapitres de ce document constituent la première partie qui donne les premiers éléments de base de la statistique inductive. Ils seront consacrés à des introductions respectivement à la théorie de l'estimation et aux tests d'hypothèses. Ils exposeront des notions de base qu'il faut connaître, le premier sur les qualités des estimateurs, les méthodes d'estimations et les précisions des estimations obtenues, et le second sur le procédé à suivre pour tester si une hypothèse est vraie ou fausse afin d'appréhender la réalité que l'on ne connaît pas. Nous verrons que les chapitres qui vont suivre ne seront que des applications des méthodes de l'inférence statistique.

CHAPITRE 1

INTRODUCTION À LA THÉORIE DE L'ESTIMATION

1.1. INTRODUCTION

Une population est caractérisée par un certain nombre de paramètres (moyenne, variance, coefficient de régression, etc.). Dans la plupart des cas, ces paramètres sont inconnus du chercheur. Les problèmes d'estimation permettent d'obtenir des estimations de ces paramètres à partir des valeurs obtenues d'échantillons convenablement prélevés dans cette population.

Supposons en effet que l'on s'intéresse à une variable aléatoire dont la distribution de probabilité connue analytiquement dépend d'un paramètre θ inconnu numériquement. Le problème de l'estimation consiste à construire une expression, fonction des réalisations de la variable dans un échantillon de données, permettant d'estimer le paramètre inconnu. Nous considérons que le paramètre inconnu θ est non aléatoire.

L'objectif de la recherche d'une bonne estimation d'un paramètre inconnu de la population permet par exemple de mieux comprendre un phénomène donné ou de pouvoir faire de la prévision en insérant la valeur estimée à la place du paramètre inconnu.

Dans ce chapitre, nous considérerons d'abord l'estimation ponctuelle d'un paramètre. Nous définirons à cette occasion la notion d'estimateur (paragraphe 1.2), les différents critères pour juger les qualités d'un estimateur (paragraphe 1.3) et les méthodes d'estimation les plus connues (paragraphe 1.4). Ensuite, nous présenterons l'estimation par intervalle de confiance pour exprimer la précision de l'estimation obtenue (paragraphe 1.5).

1.2. NOTION D'ESTIMATEUR

1.2.1. Définition d'un estimateur

Soit un phénomène modélisé par une variable aléatoire X dont la distribution de probabilité dépend d'un paramètre inconnu θ et soit (X_1, X_2, \dots, X_n) un échantillon aléatoire de taille n et (x_1, x_2, \dots, x_n) les valeurs observées correspondantes. Les variables aléatoires X_1, X_2, \dots, X_n sont de même loi que X et indépendantes (i.i.d¹).

On appelle **estimateur** ponctuel du paramètre inconnu θ toute fonction $G(X_1, X_2, \dots, X_n)$ de l'échantillon (X_1, X_2, \dots, X_n) ne dépendant pas du paramètre inconnu θ et servant à l'estimer. Il s'agit d'un procédé de calcul permettant d'obtenir une approximation de θ .

¹ i.i.d est une abréviation souvent utilisée pour indiquer que des variables aléatoires sont indépendantes et identiquement distribuées.

Toute valeur numérique $G(x_1, x_2, \dots, x_n)$ correspondant à la réalisation (x_1, x_2, \dots, x_n) de l'échantillon (X_1, X_2, \dots, X_n) est une **estimation** de θ ; c'est la valeur de l'estimateur. Une estimation se réfère donc à la valeur numérique utilisée pour l'approximation. Les estimations seront, dans la suite, notées g_1, g_2, g_3 , etc.

Remarques :

- l'estimateur est une variable aléatoire ; la valeur estimée, obtenue par l'estimateur, peut donc varier d'un échantillon de données à un autre ;
- il existe plusieurs estimateurs d'un paramètre d'une population.

Exemple 1.1.

Reprenez l'exemple de la teneur en protéines sériques totales du sang de bovins adultes cité lors de l'introduction de cette première partie, en considérant que cette variable est distribuée selon une loi normale de moyenne et de variance inconnues μ et σ^2 . Cherchez une estimation de la teneur moyenne en protéines sériques totales du sang de la population (μ), sachant que l'on dispose des valeurs des teneurs en protéines sériques totales du sang d'un échantillon aléatoire et simple de dix bovins (tableau 1.1).

Tableau 1.1. Teneurs en protéines sériques totales du sang de 10 bovins adultes (en g/100 g).

N° vache	1	2	3	4	5	6	7	8	9	10
Teneur	7.2	6.8	7.3	8.1	8.5	8.6	8.2	7.5	7.3	6.7

Solution 1.1

Dans cet exemple :

- \mathbf{X} est la variable aléatoire qui modélise la teneur en protéines sériques totales ;
- Le paramètre inconnu à estimer θ est la moyenne μ de la population (teneur moyenne théorique en protéines sériques totales) ;
- \mathbf{X}_i est la variable aléatoire qui modélise la teneur en protéines sériques totales de la i -ème vache, tandis que \mathbf{x}_i désigne la valeur observée correspondante ;
- Un des estimateurs de μ est :

$$G(X_1, X_2, \dots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i ;$$

- \bar{X} est un estimateur de la moyenne μ ;
- Une estimation de μ est fournie par les données de l'échantillon :

$$\hat{\mu} = g_1 = \bar{x}_1 = 7.62 \text{ g/100 g}$$

où l'accent circonflexe indique qu'il s'agit d'une estimation.

La moyenne μ de la population est ici estimée par la moyenne de l'échantillon, mais nous pouvons aussi l'estimer par :

- la **médiane observée** (7.40 g/100 g),
- la **première valeur** de l'échantillon 7.2 g/100 g ,
- etc.

1.2.2. Notations

Pour les besoins des paragraphes et des chapitres suivants, nous aurons besoins de distinguer entre les paramètres de la population, c'est-à-dire ceux qu'on cherche à estimer, et les paramètres observés, c'est-à-dire ceux qu'on calcule à partir d'un échantillon. Ces notations sont reprises au tableau 1.2.

Tableau 1.2. Notations adoptées pour quelques paramètres de la population et de l'échantillon.

Paramètre	Population	Échantillon
Moyenne	μ	\bar{x}
Variance	σ^2	s^2
Écart-type	σ	s
Coefficient de corrélation	ρ	r
Coefficient de régression	β	b

1.3. QUALITÉS D'UN ESTIMATEUR

1.3.1. Généralités

En présence de plusieurs estimateurs d'un même paramètre, on est tenté de choisir celui qui semble être le "*meilleur*". Dans ce cas, on est amené à examiner différents critères qui définissent les qualités d'un estimateur (paragraphe 1.3.2 à 1.3.7). Certaines de ces propriétés s'appliquent aux estimateurs quelle que soit la taille de l'échantillon, d'autres s'appliquant lorsque la taille de l'échantillon est élevée.

1.3.2. Estimateur sans biais

Un estimateur est dit sans biais, ou non-biaisé, s'il ne surestime ou sous-estime pas de manière systématique le paramètre inconnu θ . En moyenne, on doit retrouver la vraie valeur du paramètre θ . L'estimateur $G(X_1, X_2, \dots, X_n)$ de θ est donc sans biais si :

$$E(G) = \theta ,$$

où $E(G)$ est l'espérance mathématique de l'estimateur. Le biais est donné par :

$$\boxed{\text{biais} = E(G) - \theta}.$$

Un biais positif signifie que l'estimation, en moyenne, surestime le paramètre inconnu θ , alors qu'un biais négatif le sous-estime.

La figure 1.1a montre que l'espérance mathématique, ou la moyenne de la distribution, de l'estimateur U, est égale à la vraie valeur du paramètre inconnu θ , tandis que l'espérance mathématique de l'estimateur V ne coïncide pas avec la véritable valeur de θ (figure 1.1b). L'estimateur U est donc sans biais alors que l'estimateur V est biaisé.

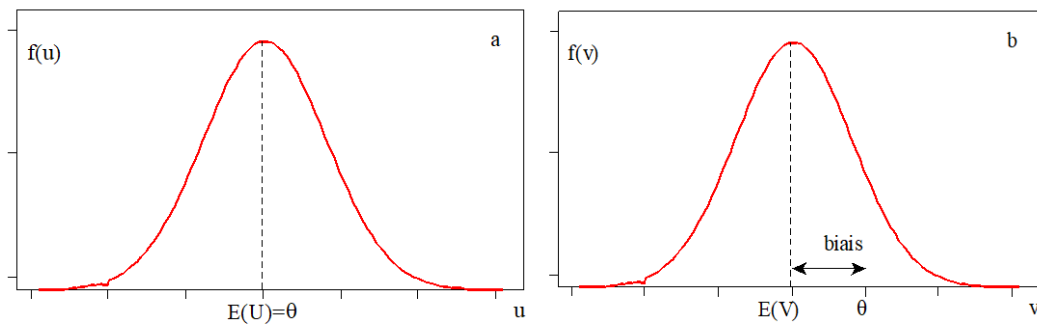


Figure 1.1. Illustration d'un estimateur U sans biais (a) et d'un estimateur V avec biais (b).

Exemple 1.2.

Vérifiez si \bar{X} est un estimateur sans biais de la moyenne μ de la population.

Solution 1.2.

a) **Méthode** : calcul du biais d'un estimateur.

b) **Calcul** :

Le biais est :

$$\begin{aligned} E(\bar{X}) - \mu &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) - \mu = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) - \mu \\ &= \frac{1}{n} \left(\sum_{i=1}^n E(X_i)\right) - \mu = \frac{1}{n} nE(X_i) - \mu = 0 . \end{aligned}$$

Puisque le biais est nul, il en résulte que \bar{X} est un estimateur sans biais de la moyenne. Dans la pratique on utilise en effet souvent la moyenne de l'échantillon pour estimer la moyenne de la population :

$$\hat{\mu} = \bar{x} .$$

On peut remarquer qu'on peut aussi obtenir un estimateur non biaisé du paramètre μ en prenant une valeur de l'échantillon prélevé, puisque $E(X_i) = \mu$, où i est le numéro d'observation dans l'échantillon. Ce dernier estimateur apparaît moins intéressant que la moyenne car cette dernière contient plus d'information sur la population. Il est donc clair que nous aurons besoin d'autres critères pour comparer les estimateurs.

Exemple 1.3.

Vérifiez si S^2 est un estimateur sans biais de la variance σ^2 de la population.

Solution 1. 3.

a) **Méthode** : calcul du biais d'un estimateur.

b) **Calcul** :

Le biais est donné par :

$$\begin{aligned} E(S^2) - \sigma^2 &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) - \sigma^2 \\ &= \frac{(n-1)\sigma^2}{n} - \sigma^2 = \frac{-\sigma^2}{n}. \end{aligned}$$

Le biais est donc de $-\sigma^2/n$. Il en résulte que S^2 est un estimateur biaisé de la variance de la population σ^2 : la variance de l'échantillon est en moyenne inférieure à la variance de la population.

C'est pourquoi on apporte une correction en multipliant la variance de l'échantillon par $n/(n-1)$ pour obtenir l'estimateur **non biaisé** suivant :

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Dans la pratique, une estimation sans biais de la variance de la population σ^2 peut donc être obtenue à partir de l'échantillon en utilisant l'expression suivante :

$$\hat{\sigma}^2 = \frac{ns^2}{n-1} = \frac{\text{SCE}_x}{n-1}$$

où s^2 est la variance de l'échantillon calculée par l'expression $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\text{SCE}_x}{n}$.

Cela ne veut pas dire que la racine carrée de la variance S_{n-1}^2 est un estimateur sans biais de l'écart-type σ de la population.

D'autre part, lorsque la moyenne μ de la population est **connue**, ce qui est rare, alors la quantité $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur sans biais de la variance de la population. Dans ce cas précis, on peut utiliser l'estimation $\hat{\sigma}^2 = s^2$.

1.3.3. Estimateur de variance minimum

La variance d'un estimateur G de θ est donnée par :

$$\text{Var}(G) = E(G - E(G))^2.$$

Un estimateur sans biais mais de variance très grande n'est pas intéressant. Il indique qu'on ne se trompe pas en moyenne, mais on peut se tromper largement pour certains échantillons.

Dans ce cas, on retient que :

- si on compare deux estimateurs sans biais, on préférera celui dont la variance est la plus petite ;
- si on compare deux estimateurs d'égale variance, on préférera celui dont le biais est le plus petit.

Un estimateur G est dit de **variance minimum** si :

$$E(G - E(G))^2 \leq E(G^* - E(G^*))^2 ,$$

où G^* est tout autre estimateur.

Exemple 1.4.

L'exemple 1.2 nous a permis de retenir que le i -ème élément de l'échantillon aléatoire X_i et la moyenne \bar{X} sont deux estimateurs non-biaisés de la moyenne μ de la population. Calculez les variances de ces estimateurs et commentez les résultats obtenus.

Solution 1.4.

a) Méthode : calcul de la variance d'un estimateur.

b) Calcul :

Soit la variable X_i qui représente le i -ème élément de l'échantillon aléatoire. On a :

$$\text{Var}(X_i) = \sigma^2$$

-Soit la variable \bar{X} qui représente la moyenne. Sa variance est donnée par :

$$\begin{aligned} \text{var}(\bar{X}) &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) \\ &= \frac{1}{n^2} n \text{var}(X_i) = \frac{\sigma^2}{n} , \end{aligned}$$

dans laquelle les variables X_i sont indépendantes.

Il ressort que $\text{var}(\bar{X}) \leq \text{Var}(X_i)$, indépendamment de la valeur de μ . Il est donc plus probable d'obtenir une estimation de la moyenne proche de μ en utilisant l'estimateur \bar{X} qu'en utilisant l'estimateur X_i .

Retenez que la précision de \bar{X} augmente avec la taille de l'échantillon n .

1.3.4. Estimateur efficace

Un estimateur efficace est un estimateur sans biais dont la variance est égale à une borne dite borne de Cramer–Rao et qui vérifie un certain nombre de conditions.

Nous ne souhaitons pas aborder ici en détail cette qualité d'estimateur, mais il y a lieu de retenir :

- un estimateur efficace est un estimateur sans biais de variance minimum ;
- un estimateur efficace n'existe pas toujours ;

- un estimateur efficace de θ , s'il existe, est unique ;
- si l'on a deux estimateurs non biaisés G et G' d'un paramètre θ d'une population, on dira que l'estimateur G est efficace relativement à l'estimateur G' lorsque la variance de G est inférieure à celle de G' . Autrement dit, de deux estimateurs non biaisés, le plus efficace est celui qui a la plus petite variance.

1.3.5. Erreur quadratique moyenne d'un estimateur

L'erreur quadratique moyenne (**EQM**) permet de voir comment un estimateur G , biaisé ou non-biaisé, est dispersé autour de sa véritable valeur θ . Elle s'écrit :

$$EQM = E(G - \theta)^2 .$$

Il s'agit d'une sorte de combinaison de biais et de variance, puisqu'elle se décompose comme suit :

$$\begin{aligned} EQM &= Var(G) + (E(G - \theta))^2 \\ &= \text{variance} + (\text{biais})^2 . \end{aligned}$$

Cela veut dire qu'une variance plus petite peut compenser le biais d'un estimateur biaisé. L'erreur quadratique moyenne constitue donc un critère qui peut être utilisé pour choisir entre deux estimateurs quelconques (avec ou sans biais). On choisira l'estimateur possédant une **EQM** petite.

Si l'estimateur est sans biais, **EQM** n'est autre que la variance de la distribution d'échantillonnage.

1.3.6. Estimateur convergent

Un estimateur G du paramètre θ est dit convergent en probabilité, si :

$$\lim_{n \rightarrow \infty} P(|G - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0 .$$

Un estimateur convergent s'écarte donc du paramètre avec une faible probabilité, si la taille de l'échantillon est assez grande. Autrement dit, lorsque l'effectif de l'échantillon devient suffisamment grand, on est assez sûr que l'estimateur G sera compris dans un intervalle assez étroit autour du paramètre inconnu θ .

Il en découle que si :

$$\lim_{n \rightarrow \infty} E(G) = \theta \quad \text{et} \quad \lim_{n \rightarrow \infty} Var(G) = 0$$

alors G est un estimateur convergent de θ . Autrement dit, le biais et la variance tendent l'un et l'autre vers zéro lorsque la taille de l'échantillon tend vers l'infini.

On peut remarquer que \bar{X} est un estimateur sans biais et convergent de la moyenne μ , car :

$$E(\bar{X}) = \mu \quad \text{et} \quad Var(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 .$$

Exemple 1.5 :

Considérons trois estimateurs U , V et W du paramètre inconnu θ , dont les fonctions de densité de probabilité sont représentées par la figure 1.2. Lequel des estimateurs préfère-t-on ?

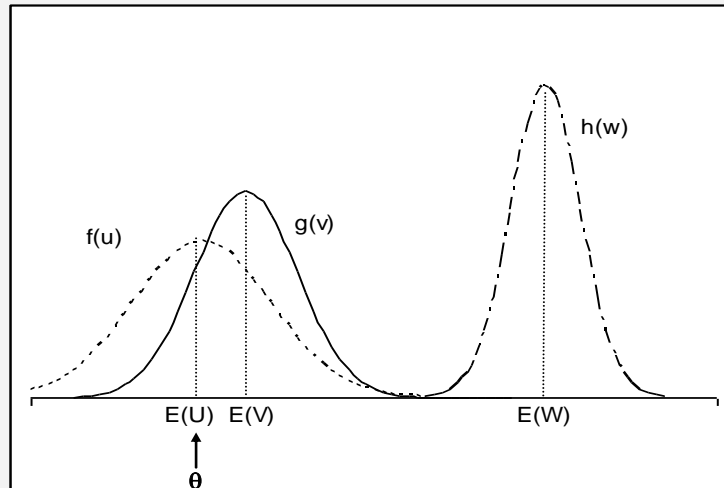


Figure 1.2. Illustration de la comparaison des biais et des variances de trois estimateurs (U , V et W).

Solution 1.5 :

- l'estimateur U est sans biais, mais sa variance est plus grande que celles des estimateurs V et W ;
- l'estimateur V est biaisé, mais moins fortement biaisé que l'estimateur W . Sa variance est par contre plus grande que celle de l'estimateur W ;
- l'estimateur W a la plus petite variance des trois estimateurs, mais il est fortement biaisé.

De ce qui précède, il semble que l'estimateur V assure une meilleure combinaison de biais et variance.

Autre exemple :

Trois tireurs (A, B et C) visent le point bleu situé au centre du cercle. Commenter la qualité des tirs ?



Tireur A



Tireur B



Tireur C

1.3.7. Estimateur asymptotiquement sans biais

Un estimateur est dit asymptotiquement sans biais quand le biais tend vers zéro lorsque la taille de l'échantillon croît vers l'infini.

1.3.8. Meilleur estimateur linéaire non biaisé

On dit que G est le meilleur estimateur linéaire non biaisé du paramètre θ si :

a) G est un estimateur sans biais : $E(G) = \theta$

b) G est une fonction linéaire des observations de l'échantillon : $G = \sum_{i=1}^n u_i x_i$

c) $Var(G) \leq Var(G')$ où G' est n'importe quel autre estimateur non biaisé de θ et qui satisfait la condition (b).

1.4. MÉTHODES D'ESTIMATION

Plusieurs méthodes permettant d'obtenir des estimations des paramètres inconnus existent. Parmi celles-ci, on peut citer la méthode du maximum de vraisemblance, la méthode des moments et la méthode des moindres carrés. Les deux premières méthodes seront présentées dans les paragraphes 1.4.1 et 1.4.2 et nous reviendrons à la troisième méthode au chapitre réservé à la régression.

1.4.1. Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est la plus populaire des méthodes d'estimation. Soient x_1, x_2, \dots, x_n les valeurs observées d'un échantillon aléatoire X_1, X_2, \dots, X_n de taille n de la variable aléatoire X dont la densité de probabilité $f(x)$ dépend d'un paramètre inconnu θ . La méthode du maximum de vraisemblance a pour principe de choisir comme estimations du paramètre inconnu θ , les valeurs qui donnent à l'événement observé (x_1, x_2, \dots, x_n) la plus grande probabilité/densité possible.

Exemple 1.6 :

Soient deux fabricants qui produisent des boulons. Le fabricant A avec un taux de défectueux moyen de 2% et le fabricant B avec un taux de défectueux moyen de 5%. Si l'on contrôle un échantillon de 200 boulons et que l'on trouve 4 défectueux, de quel fabricant provient l'échantillon ?

Solution 1.6 :

a) **Méthode :** Méthode du maximum de vraisemblance.

b) **Calcul :**

Soit X le nombre de boulons défectueux. X est une variable aléatoire binomiale de paramètres $n=200$ et p , où p est égale à 0.02 pour le fabricant A et 0.05 pour le fabricant B.

La loi de probabilité s'écrit :

$$P(X = x) = C_{200}^x p^x (1 - p)^{200-x}$$

ce qui donne $P(X=4)=0.197$ si les boulons provenaient du fabricant A et $P(X=4)=0.017$ si les boulons provenaient du fabricant B.

En raisonnant en termes de vraisemblance, il est donc plus vraisemblable que les boulons proviennent du fabricant A.

2) Pour obtenir des estimations au sens du maximum de vraisemblance, on définit une fonction de vraisemblance qui apparaît comme la loi de probabilité dans le cas d'une variable aléatoire discrète ou comme la densité de la probabilité dans le cas d'une variable aléatoire continue. Cette fonction s'écrit :

$$L(x_1, \dots, x_n; \theta) = P(x_1, \dots, x_n; \theta) = P(x_1; \theta) \dots P(x_n; \theta) \quad (\text{cas d'une variable discrète})$$

ou :

$$L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta) = f(x_1; \theta) \dots f(x_n; \theta) \quad (\text{cas d'une variable continue}).$$

On appelle estimateur du maximum de vraisemblance de θ toute statistique θ^* , fonction de X_1, X_2, \dots, X_n indépendamment et identiquement distribués, telle que :

$$L(X_1, \dots, X_n; \theta^*) \geq L(X_1, \dots, X_n; \theta), \quad \forall \theta.$$

On est donc amené à chercher le maximum de la fonction de vraisemblance $L(X_1, \dots, X_n; \theta)$.

En disposant des valeurs observées d'un échantillon aléatoire et simple de n observations, le maximum est obtenu en écrivant la quantité suivante :

$$\text{Log}_e L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \text{Log}_e P(x_i; \theta)$$

ou

$$\text{Log}_e L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \text{Log}_e f(x_i; \theta)$$

et en annulant sa dérivée par rapport à θ , lorsque cette dérivée est définie :

$$\frac{\partial \text{Log}_e L(x_1, \dots, x_n; \theta)}{\partial \theta} = 0.$$

Le maximum correspond à toute solution de l'équation précédente qui vérifie :

$$\frac{\partial^2 \text{Log}_e L(x_1, \dots, x_n; \theta)}{\partial \theta^2} < 0,$$

pour autant que cette dérivée existe.

En présence de plus d'un paramètre à estimer, l'équation doit être remplacée par des équations simultanées de vraisemblance. Dans le cas de deux paramètres inconnus θ_1 et θ_2 , comme les paramètres μ et σ de la loi normale, on commence par annuler les deux dérivées partielles par rapport à θ_1 et θ_2 (c'est-à-dire par rapport μ et σ dans le cas de la loi normale) suivantes :

$$\frac{\partial \text{Log}_e L(x_1, \dots, x_n; \theta_1, \theta_2)}{\partial \theta_1} \quad \text{et} \quad \frac{\partial \text{Log}_e L(x_1, \dots, x_n; \theta_1, \theta_2)}{\partial \theta_2}.$$

La procédure du maximum de vraisemblance possède plusieurs propriétés intéressantes.

Exemple 1.7 :

Soit X une variable aléatoire de Poisson de paramètre θ . Chercher l'estimation de θ au sens du maximum de vraisemblance en utilisant l'échantillon suivant de 10 observations ?

5 3 1 2 2 0 1 2 3 6

Solution 1.7 :

a) Méthode : estimation du paramètre de la loi de Poisson par la méthode du maximum de vraisemblance.

b) Calcul :

La loi de probabilité est donnée par : $P(X = x) = \frac{e^{-\theta} \theta^x}{x!} \quad x \geq 0$

La fonction de vraisemblance est : $L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n P(x_i; \theta) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!}$.

On cherche le maximum de la quantité :

$$\text{Log}_e L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n (-\theta + x_i \text{Log}_e \theta - \text{Log}_e x_i!).$$

La dérivée par rapport au paramètre θ donne :

$$\frac{\partial \text{Log}_e L(x_1, \dots, x_n; \theta)}{\partial \theta} = \sum_{i=1}^n (-1 + x_i / \theta) = -n + \frac{\sum_{i=1}^n x_i}{\theta}.$$

En annulant la dérivée, on obtient l'estimation $\hat{\theta}$ au sens du maximum de vraisemblance :

$$\frac{\partial \text{Log}_e L(x_1, \dots, x_n; \theta)}{\partial \theta} = -n + \frac{\sum_{i=1}^n x_i}{\hat{\theta}} = 0$$

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

On peut vérifier que la dérivée seconde est toujours négative ($-\theta^{-2} \sum_{i=1}^n x_i$). La valeur $\hat{\theta}$ est bien un maximum. L'estimateur du maximum de vraisemblance de θ est $\hat{G} = \bar{X}$.

En utilisant l'échantillon de données, on obtient l'estimation : $\hat{\theta} = \bar{x} = 2.5$.

Exemple 1.8 :

Soit \mathbf{X} une variable aléatoire qui modélise le temps d'attente jusqu'au prochain appel au standard téléphonique d'une société. La fonction de densité de \mathbf{X} s'écrit :

$$f(\mathbf{x}) = \lambda e^{-\lambda x} \quad \text{pour } x \geq 0.$$

Déterminez l'estimation de λ au sens du maximum de vraisemblance en utilisant l'échantillon de données ci-dessous ?

5.62 1.27 4.89 5.02 0.43 1.88 4.59 3.16 1.74 0.82

Solution 1.8 :

a) **Méthode** : estimation du paramètre de la loi exponentielle par la méthode du maximum de vraisemblance.

b) **Calcul** :

La fonction de vraisemblance est : $L(x_1, \dots, x_n; \lambda) = \prod_{i=1}^n f(x_i; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i}$

et son logarithme : $\text{Log}_e L(x_1, \dots, x_n; \lambda) = \sum_{i=1}^n (\text{Log}_e \lambda - \lambda x_i)$.

La dérivée par rapport au paramètre λ :

$$\frac{\partial \text{Log}_e L(x_1, \dots, x_n; \lambda)}{\partial \lambda} = \sum_{i=1}^n (1/\lambda - x_i) = n/\lambda - \sum_{i=1}^n x_i.$$

En annulant la dérivée, on obtient l'estimation $\hat{\lambda}$ au sens du maximum de vraisemblance :

$$n/\hat{\lambda} - \sum_{i=1}^n x_i = 0$$

$$\hat{\lambda} = 1/\bar{x}.$$

On peut vérifier que la dérivée seconde est toujours négative ($-n/\lambda^2$). La valeur $\hat{\lambda}$ est bien un maximum. L'estimateur du maximum de vraisemblance de λ est $\mathbf{G} = 1/\bar{\mathbf{X}}$.

En utilisant l'échantillon de données, on obtient l'estimation :

$$\hat{\lambda} = 1/\bar{x} = 0.340 \text{ (min}^{-1}\text{)}.$$

1.4.2. Méthode des moments

Soit X une variable aléatoire ayant une distribution donnée, continue ou discrète, dépendant de k paramètres inconnus $\theta_1, \theta_2, \dots, \theta_k$, et soit (X_1, X_2, \dots, X_n) un échantillon aléatoire de taille n et (x_1, x_2, \dots, x_n) les valeurs observées correspondantes.

La méthode des moments consiste à estimer les paramètres inconnus $\theta_1, \theta_2, \dots, \theta_k$, en posant l'égalité entre les k premiers moments observés (connus) de l'échantillon et les k premiers moments (inconnus) correspondants de la population. Cette égalité se justifie par la loi des

grands nombres qui implique que plus l'échantillon est grand, plus on peut accorder de confiance en estimant les moments théoriques par les moments observés.

Exemple 1.9 :

Soit X une variable aléatoire dont la fonction de densité de probabilité dépend du paramètre θ :

$$f(x) = \begin{cases} 1/\theta & 0 < x \leq \theta \\ 0 & \text{sinon} \end{cases}.$$

Déterminez l'estimation de θ par la méthode des moments sachant que les 12 observations suivantes proviennent de la population concernée : 15 18 15 12 11 17 22 11 14 23 20 14 16 ?

Solution 1.9 :

a) Méthode : estimation du paramètre inconnu θ par la méthode des moments.

b) Calcul :

$$E(X) = \int_0^\theta xf(x)dx = \frac{1}{\theta} \int_0^\theta xdx = \frac{\theta}{2}$$

$$\hat{\theta}/2 = \bar{x} \Rightarrow \hat{\theta} = 2\bar{x}.$$

En utilisant les données de l'échantillon, on obtient : $\hat{\theta} = 32$.

1.5. ESTIMATION PAR INTERVALLE

1.5.1. Introduction

Dans les paragraphes précédents, nous avons vu comment obtenir des estimations **ponctuelles** qui puissent être utilisées à la place du paramètre inconnu θ de la population. Mais, dans la pratique, les chercheurs souhaitent aussi obtenir une estimation de la **précision** de ce paramètre, car l'estimation ponctuelle peut être plus ou moins éloignée de la valeur de θ . Cette précision peut être exprimée par un intervalle, appelé **intervalle de confiance**, qui contiendrait le paramètre inconnu θ avec une très forte probabilité. Les limites G_1 et G_2 de cet intervalle s'appellent les **limites de confiance**.

D'une manière générale, on appelle intervalle de confiance de θ de niveau de confiance ou de sécurité $1-\alpha$, un intervalle tel que :

$$P(G_1 \leq \theta \leq G_2) = 1 - \alpha.$$

où le paramètre α est, comme nous le verrons au chapitre 2, un réel positif inférieur à 1 appelé niveau de signification.

Dans ce cas, on dit qu'il y a $(1-\alpha)$ % de chances que les limites G_1 et G_2 de l'intervalle encadrent le vrai paramètre inconnu θ . Il sera incorrect de dire qu'il y a $(1-\alpha)$ % de chances

que le vrai paramètre se trouve encadré par les limites de confiance G_1 et G_2 , car le paramètre inconnu est fixe et ce sont les limites de confiances qui sont des variables aléatoires. Une estimation par intervalle de confiance sera d'autant meilleure que l'intervalle de confiance sera petit.

Outre les données d'un échantillon de la population, le point de départ de la détermination de l'intervalle de confiance sera la connaissance de la loi de probabilité de l'estimateur G du paramètre à estimer θ .

Dans ce qui suit, nous aborderons l'exemple de l'estimation de l'intervalle de confiance de la moyenne μ d'une population lorsqu'on connaît la variance σ^2 (paragraphe 1.5.2). Les intervalles de confiance de bien d'autres paramètres statistiques (variance, coefficient de régression, etc.) seront examinés dans les prochains chapitres. Ensuite, nous verrons comment calculer la taille minimum de l'échantillon de telle sorte que les limites inférieure et supérieure de l'intervalle de confiance ne s'écartent pas de plus d'une valeur donnée (paragraphe 1.5.3).

1.5.2. Cas de la moyenne arithmétique

1.5.2.1. Intervalle de confiance de la moyenne

Soit X une variable aléatoire normale de moyenne inconnue μ et de variance connue σ^2 et soit \bar{X} un estimateur du paramètre inconnu μ . L'objectif est de déterminer les limites de confiance \bar{X}_1 et \bar{X}_2 de l'intervalle qui a une probabilité importante de contenir le paramètre μ .

Pour obtenir cet intervalle, on se fixe un degré de confiance $1-\alpha$ assez proche de 1, généralement de 0.90, 0.95 ou 0.99. On peut donc écrire :

$$P(\bar{X}_1 \leq \mu \leq \bar{X}_2) = 1 - \alpha.$$

L'intervalle de confiance $[\bar{X}_1, \bar{X}_2]$ peut s'écrire sous la forme $[\bar{X} - \varepsilon_1, \bar{X} + \varepsilon_2]$.

En général, on répartit le risque d'erreur α en deux parties égales, c'est-à-dire :

$$P(\mu < \bar{X} - \varepsilon_1) = P(\mu > \bar{X} + \varepsilon_2) = \alpha / 2$$

et on obtient $P(\varepsilon_1 < \bar{X} - \mu) = P(\varepsilon_2 < \mu - \bar{X}) = \alpha / 2$.

D'autre part, on sait que la variable X suit une loi normale de moyenne μ et de variance σ^2 , alors la variable \bar{X} suit elle aussi une loi normale de moyenne μ et de variance σ^2/n . Il en résulte :

$$P\left(\frac{\varepsilon_1}{\sigma / \sqrt{n}} < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right) = P\left(\frac{\varepsilon_2}{\sigma / \sqrt{n}} < \frac{\mu - \bar{X}}{\sigma / \sqrt{n}}\right) = \alpha / 2$$

$$P\left(\frac{\varepsilon_1}{\sigma / \sqrt{n}} < U_1\right) = P\left(\frac{\varepsilon_2}{\sigma / \sqrt{n}} < U_2\right) = \alpha / 2$$

$$P\left(U_1 < \frac{\varepsilon_1}{\sigma / \sqrt{n}}\right) = P\left(U_2 < \frac{\varepsilon_2}{\sigma / \sqrt{n}}\right) = 1 - \alpha / 2$$

$$\frac{\varepsilon_1}{\sigma / \sqrt{n}} = U_{1-\alpha/2} \quad \text{et} \quad \frac{\varepsilon_2}{\sigma / \sqrt{n}} = U_{1-\alpha/2}$$

On obtient :

$$\varepsilon_1 = \varepsilon_2 = u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

ce qui donne les limites de confiance suivantes :

$$\boxed{\bar{x}_1 = \bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}} \quad \text{et} \quad \boxed{\bar{x}_2 = \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

L'intervalle de confiance de niveau de confiance $1-\alpha$ pour μ , lorsque σ^2 est connue, s'écrit donc :

$$\left[\bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Il ressort que la longueur de l'intervalle de confiance de la moyenne :

- augmente avec l'augmentation de l'écart-type σ (c'est-à-dire avec la variabilité dans la population),
- augmente avec le degré de confiance ($1-\alpha$), et
- diminue avec l'augmentation de l'effectif de l'échantillon (n).

On peut aussi énoncer l'intervalle de confiance de la façon suivante : si on prélève un nombre très élevé d'échantillons d'effectif n dans la population et qu'on calcule l'intervalle de confiance de la moyenne pour chacun d'eux par la méthode précédente avec un degré de confiance de **0.95** ($\alpha=0.05$), on trouve **95%** des intervalles construits qui contiennent la vraie moyenne inconnue de la population μ .

La quantité $u_{1-\alpha/2} \sigma / \sqrt{n}$ est appelée **marge d'erreur** ou demi-longueur de l'intervalle de confiance.

Exemple 1.10 :

On s'intéresse au poids des œufs des poules pondeuses d'une race donnée. Sur un échantillon aléatoire et simple de 12 œufs appartenant à des poules différentes, on a relevé les poids suivants (en grammes) :

64.6 66.5 61.4 65.5 62.3 66.6 62.4 64.8 64.7 63.2 63.9 61.3

Déterminez l'estimation du poids moyen des œufs des poules de la race en question et son intervalle de confiance au niveau de confiance 0.95, en supposant que la variance des poids dans la population est de **3 g²** ?

Solution 1.10 :

a) Méthode : détermination de l'intervalle de confiance de la moyenne lorsque l'écart-type est connu (attention l'écart-type de la population est rarement connu dans la pratique (voir chapitre 3)).

b) Conditions d'application :

- l'échantillon des 12 œufs est prélevé de manière aléatoire et simple
- la distribution des poids des œufs de cette race de poules est normale

c) Calcul :

- la moyenne de la population est estimée par la moyenne de l'échantillon (estimation sans biais) :

$$\hat{\mu} = \bar{x} = 63.93g$$

- les limites de confiance sont obtenues par :

$$\bar{x} \pm u_{1-\alpha/2} \sigma / \sqrt{n} = 63.93 \pm 1.96 \sqrt{3} / \sqrt{12}$$

On retient qu'il y a 95% de chances que les valeurs **62.95** et **64.91 g** encadrent le vrai poids moyen des œufs de la race étudiée.

1.5.2.2. Taille de l'échantillon

Parfois, l'expérimentateur s'intéresse à la taille minimum de l'échantillon **n** tel que l'intervalle de confiance possède, pour un degré de confiance de **1-α**, des limites inférieure et supérieure ne s'écartant pas de plus de **ε**. Dans ces conditions, on pose :

$$u_{1-\alpha/2} \sigma / \sqrt{n} \leq \varepsilon$$

et on tire l'effectif **n**, en utilisant la relation suivante :

$$n \geq \frac{u_{1-\alpha/2}^2 \sigma^2}{\varepsilon^2} .$$

Souvent, la valeur de **ε** s'exprime en pour-cent de la moyenne. Autrement dit, si on parle de **ε** de 10% de la moyenne, cela veut dire que **ε = 0.10μ**.

Exemple 1.11 :

Reprenez l'exemple 1.10 où l'on s'intéresse à l'estimation du poids des œufs des poules pondeuses d'une race donnée. Quelle est la taille minimale de l'échantillon d'œufs qu'il faut observer si l'on veut estimer le poids moyen avec un intervalle de confiance dont la longueur est inférieure à 1.5 g ? (prendre $\alpha=0.05$)

Solution 1.11 :

a) Méthode : détermination de l'effectif minimum de l'échantillon pour obtenir un intervalle de confiance d'une longueur donnée.

b) Conditions d'application

- la population des poids des œufs est normale
- l'échantillonnage est aléatoire et simple

c) Calcul :

- la longueur de l'intervalle de confiance est : $L = 2u_{1-\alpha/2}\sigma / \sqrt{n}$



$$2u_{1-\alpha/2}\sigma / \sqrt{n} \leq 1.5 \quad \text{c'est-à-dire} \quad \sqrt{n} > (2)(1.96)(\sqrt{3})/(1.5) .$$

On est amené à observer un échantillon dont l'effectif minimum est de **21** œufs.

1.6. CONCLUSION

Nous avons vu que les problèmes d'estimation ont pour but d'obtenir des estimations des paramètres inconnus d'une population. Nous avons souligné l'existence de différentes méthodes d'estimation tout en illustrant les méthodes du maximum de vraisemblance et des moments par des exemples. La méthode des moindres carrés sera utilisée pour l'estimation de paramètres pour des problèmes de régression (chapitre 9).

D'autre part, on est assez souvent confronté au choix du bon estimateur. A ce propos, il faut signaler qu'on a généralement tendance à préférer l'estimateur non biaisé par rapport à l'estimateur biaisé, alors qu'il existe, comme nous l'avons vu, plusieurs autres critères pour caractériser un estimateur, notamment l'efficacité et la convergence.

Enfin, nous avons vu comment calculer, à partir d'un échantillon, les bornes d'un intervalle, dit intervalle de confiance, dont on a une forte chance qu'elles encadrent le paramètre inconnu de la population. Cette notion a été illustrée par l'intervalle de confiance de la moyenne lorsque la variance de la population est connue. Les intervalles de confiance de bien d'autres paramètres seront présentés en détail au fur et à mesure que nous avançons dans ce cours.

CHAPITRE 2

LES TESTS D'HYPOTHESES

2.1. INTRODUCTION

Un test d'hypothèse est une démarche de l'inférence statistique dont le but est de contrôler la validité d'une hypothèse relative à une ou plusieurs populations, dite hypothèse nulle, considérée *a priori* comme vraie, et à admettre, lorsque les faits observés l'infirmement, une autre hypothèse, dite hypothèse alternative. L'hypothèse porte sur les valeurs vraies mais inconnues des populations et les valeurs observées portent sur le (les) échantillon(s) servant à la tester.

Les expérimentateurs sont, en effet, souvent amenés à réaliser des tests d'hypothèses, en souhaitant tester, à partir de l'observation d'un ou de plusieurs échantillons, la validité d'une hypothèse relative à une ou plusieurs populations. Dans le cas, par exemple, de la comparaison des teneurs moyennes théoriques en protéines sériques totales du sang chez des vaches adultes indemnes et des vaches adultes malades, ces expérimentateurs peuvent déterminer, avec une probabilité calculée par des méthodes de l'inférence statistique, si les différences de résultats obtenues au niveau des échantillons sont suffisamment grandes pour annoncer que ces échantillons proviennent de deux populations vraisemblablement différentes ou si elles ne sont dues qu'au hasard. Ils ont avec précision la probabilité d'avoir rejeté à tort l'hypothèse nulle au profit de l'hypothèse alternative.

Les tests d'hypothèses sont généralement réalisés pour rejeter une hypothèse donnée que pour la démontrer, mais cela ne veut pas dire que l'expérimentateur souhaite toujours la rejeter. L'application de ces tests fait généralement appel à un certain nombre de conditions d'application concernant la nature des populations dont proviennent les échantillons étudiés et la méthode de prélèvement de ces échantillons. Selon le degré de non respect des conditions d'application, la validité des résultats se trouve plus ou moins affectée. On définit alors ce qu'on appelle la robustesse d'un test pour exprimer sa tolérance vis à vis du respect de ces conditions d'application. Parmi les tests que nous présenterons dans les chapitres suivants, plusieurs exigent la normalité des populations et le caractère aléatoire et simple des échantillons prélevés.

Dans ce chapitre, nous commencerons par définir le principe de réalisation d'un test d'hypothèse. Nous verrons que cette réalisation se déroule pratiquement toujours selon les mêmes étapes (paragraphe 2.2). Nous parlerons ensuite des principaux types de tests d'hypothèses rencontrés dans la pratique par les chercheurs. Ceux-ci varient principalement selon le problème posé, la nature des variables et les conditions de leur application (paragraphe 2.3). Enfin, pour illustrer la démarche d'un test d'hypothèse, nous considérerons comme exemple la comparaison des moyennes de deux populations (paragraphe 2.4), sachant que nous reviendrons plus en détail sur ce test fort intéressant au chapitre 3.

2.2. RÉALISATION DES TESTS

2.2.1. Principales étapes à suivre

Il existe une suite logique d'étapes afin de réaliser un test d'hypothèse. Certaines de ces étapes doivent être précisées avant même la collecte et la manipulation des données expérimentales (étapes 1 à 3), les autres ne peuvent être réalisées qu'après le recueil des données. Avant d'entrer dans les détails et l'explication des nouveaux termes dans les paragraphes qui suivent, voici, de manière résumée, les différentes étapes qu'il est conseillé de suivre :

- 1- Formulez correctement, en relation avec la question posée, l'hypothèse nulle, notée H_0 , que vous souhaitez tester. Précisez à ce niveau si le test est unilatéral ou bilatéral, en définissant l'hypothèse alternative, notée H_1 ;
- 2- Définissez la statistique S à appliquer pour tester l'hypothèse nulle, en précisant sa distribution lorsque l'hypothèse nulle est vraie et en vérifiant les conditions de son application, c'est-à-dire les conditions de validité de cette statistique ;
- 3- Etablissez la (les) valeur(s) critique(s) S_{critique} calculée(s) d'après la loi de probabilité de la statistique S et le niveau de signification accepté du test (α). Cela suppose que vous avez choisi un risque α . Vous définissez ainsi la ou les région(s) critique(s) du test au seuil α ;
- 4- Calculez, en utilisant les valeurs des échantillons, la valeur S_{obs} prise par la statistique S . Cette quantité est appelée la valeur observée de S ;
- 5- Rejetez ou non l'hypothèse nulle après avoir comparé la valeur observée S_{obs} à la valeur critique S_{critique} de la statistique S . Si l'hypothèse nulle est rejetée, vous déterminez ce qu'on appelle le degré de signification ou la probabilité d'erreur ;
- 6- Faites une interprétation des résultats en relation avec le problème que vous avez posé initialement (interprétations biologique, socio-économique, psychologique, etc.).

Les paragraphes suivants expliquent les nouveaux termes utilisés dans ces six étapes.

2.2.2. Formulation des hypothèses nulle et alternative

2.2.2.1. Choix de l'hypothèse à tester

Un test d'hypothèse commence par une assertion ou une théorie qu'on souhaite tester. Cela s'exprime par la formulation d'une hypothèse nulle et d'une hypothèse alternative. La formulation de ces hypothèses nécessite une compréhension et une simplification du problème posé.

Les données nous permettront de prendre une décision sur l'hypothèse nulle par référence à l'hypothèse alternative. On dit qu'on teste l'hypothèse nulle contre l'hypothèse alternative. Ces deux hypothèses jouent des rôles dissymétriques :

- L'hypothèse nulle, notée H_0 , est celle qu'on veut tester. Elle joue un rôle privilégié dans le test. Elle est choisie par l'expérimentateur et sa formulation dépend du problème posé ;
- L'hypothèse alternative, notée H_1 , est celle qui est admise lorsque l'hypothèse nulle est rejetée. Elle change avec l'objet du test.

La formulation de l'hypothèse nulle est en effet souvent une interprétation de la question posée. De plus, on est amené à formuler H_0 de telle sorte que son rejet lorsqu'elle est correcte soit plus grave que son acceptation lorsqu'elle est fausse.

Si l'on considère les deux hypothèses suivantes :

$$H_0 : \text{Innocence d'un accusé} \quad \text{et} \quad H_0' : \text{Culpabilité d'un accusé}$$

Quelle est l'hypothèse nulle qu'il serait plus grave de rejeter à tort. Autrement dit, est-il plus grave de condamner un innocent ou d'acquitter un coupable ? Vous direz qu'il serait plus grave de condamner un innocent.

De même, si l'on considère le risque de l'étudiant et le risque de l'enseignant, on peut se poser la question : est-il plus grave de voir redoubler un étudiant qui mérite passer à la classe supérieure ou de voir passer un étudiant qui doit refaire l'année ?

Exemple 2.1 :

On souhaite étudier le lien entre l'état de santé de vaches adultes (vaches saines et vaches atteintes d'une maladie donnée) et la teneur en protéines sériques totales du sang. Comment formuler l'hypothèse nulle ?

Solution 2.1 :

Dans cet exercice, on souhaite tester l'effet de l'état de santé des vaches sur la teneur en protéines sériques totales du sang.

On a deux populations, l'une des vaches indemnes et l'autre des vaches atteintes. La variable est la teneur en protéines sériques totales du sang (en g/l).

On pose l'hypothèse nulle :

$$H_0 : \mu_s = \mu_M \quad \text{contre l'hypothèse alternative } H_1 : \mu_s \neq \mu_M,$$

où μ_s et μ_M sont les teneurs moyennes théoriques en protéines sériques totales du sang respectivement chez les vaches saines et les vaches malades. Il s'agit d'un test d'égalité de deux moyennes théoriques.

L'hypothèse nulle consiste ici à dire qu'il n'existe pas de différence entre les moyennes théoriques des deux populations (μ_s et μ_M).

2.2.2.2. Test unilatéral et test bilatéral

Nous avons vu que l'hypothèse alternative H_1 est une nouvelle hypothèse pour décrire la réalité si l'hypothèse nulle H_0 est fausse. La nature du problème posé détermine la façon de formuler H_1 et, par conséquent, la nature unilatérale ou bilatérale du test.

a) Test bilatéral

Un test est dit **bilatéral** lorsqu'on ne peut spécifier de direction particulière pour l'hypothèse alternative. Dans ce cas, l'hypothèse nulle correspond le plus souvent à une égalité et l'hypothèse alternative à une inégalité. On ne se préoccupe pas du signe ou du sens de la différence.

Si l'on souhaite savoir si l'on peut affirmer par exemple que la production laitière moyenne d'une race bovine est égale à 7500 kg par vache et par an, on peut écrire les deux hypothèses :

Hypothèse nulle	$H_0 : \mu_A = 7500$
Hypothèse alternative	$H_1 : \mu_A \neq 7500$

Ce test tel qu'il est formulé est dit bilatéral, car le plus important est que la production moyenne diffère de la valeur 7500 kg, tout en étant plus grande ou plus petite.

Nous verrons plus loin que la zone de rejet de l'hypothèse nulle se fait de part et d'autre de la distribution de probabilité de référence.

b) Test unilatéral

L'hypothèse alternative peut changer avec l'objet du test et elle peut ne pas correspondre à une inégalité. On peut en effet spécifier une direction particulière pour l'hypothèse alternative et on parle dans ce cas de test **unilatéral**.

Supposons que l'on s'intéresse toujours à la comparaison de la production laitière moyenne d'une race bovine à la valeur de 7500 kg par vache et par an. On se pose la question si l'on peut affirmer que la moyenne de la production laitière est supérieure à 7500 kg. Dans ce cas, les deux hypothèses, nulle et alternative, s'écrivent :

Hypothèse nulle,	$H_0 : \mu_A = 7500 ;$
Hypothèse alternative,	$H_1 : \mu_A > 7500.$

Ce test est unilatéral, car le rejet de l'hypothèse nulle permet de conclure que la production laitière moyenne de la race est supérieure à 7500 kg.

Nous verrons plus loin que la zone de rejet de l'hypothèse nulle est située d'un seul côté de la distribution de probabilité de référence.

On choisit un test unilatéral lorsqu'on est sûr que le contraire est impossible.

2.2.3. Risques de première et seconde espèce et puissance d'un test

2.2.3.1. Risques de première et seconde espèce

Nous avons vu qu'un test d'hypothèse consiste à « trancher », au vu des résultats du ou des échantillon(s), entre l'hypothèse nulle H_0 , considérée comme vraie *a priori*, et l'hypothèse alternative H_1 . Le tableau 2.1 montre qu'il y a quatre situations à envisager :

- l'acceptation de l'hypothèse nulle alors qu'elle est vraie,
- le rejet de l'hypothèse nulle alors qu'elle est vraie,
- l'acceptation de l'hypothèse nulle alors qu'elle est fausse, et
- le rejet de l'hypothèse nulle alors qu'elle est fausse.

On constate alors que le hasard de l'échantillonnage peut fausser les conclusions dans deux de ces situations :

- la première consiste à rejeter l'hypothèse nulle alors qu'elle est vraie ; **c'est l'erreur de première espèce ou l'erreur I** ;
- la seconde est l'acceptation de l'hypothèse nulle alors que celle-ci est fausse ; **c'est l'erreur de deuxième espèce ou l'erreur II**.

La probabilité qui correspond à l'erreur de première espèce est **le risque de première espèce**. On la symbolise par α (alpha) :

$$\alpha = P(\text{rejeter } H_0 / H_0 \text{ vraie}) ,$$

et la probabilité qui correspond à l'erreur de deuxième espèce est **le risque de deuxième espèce**. On la symbolise par β (bêta) :

$$\beta = P(\text{accepter } H_0 / H_0 \text{ fausse}) .$$

Tableau 2.1. Événements et probabilités correspondantes lors de la réalisation d'un test d'hypothèse.

Décision	Vérité	
	H_0 vraie	H_0 fausse
Non-rejet de H_0	Correct ($1-\alpha$)	Erreur II (β)
Rejet de H_0	Erreur I (α)	Correct ($1-\beta$)

On retient donc que les tests statistiques répondent aux lois de probabilités et ils sont entachés de risques d'erreur (prendre une décision, c'est "accepter les risques"). Si ces risques sont connus, ils permettront d'apprécier la validité des conclusions.

a) La valeur de α est fixée *a priori* **par l'expérimentateur** avant même la collecte des données. Elle mesure l'importance de la preuve, plus α est petit plus la preuve est forte et inversement. Si on suppose que l'on effectue l'expérience un très grand nombre de fois, en admettant α comme risque de première espèce, on conclut à tort dans ($\alpha \cdot 100$) % des cas (c'est-à-dire dans 5% des cas si on considère un risque α de 0.05).

Lorsque l'expérimentateur fixe la valeur de α , il localise la région du rejet de l'hypothèse nulle. La principale crainte de l'expérimentateur est en effet de conclure à tort à la validité de l'hypothèse alternative. Le niveau de risque α constitue le seuil de signification du test d'hypothèse et permet de définir la condition de rejet de l'hypothèse nulle.

La valeur arbitraire de 5% ($\alpha=0,05$) est la plus communément admise par les expérimentateurs comme seuil de signification pour tester l'hypothèse nulle. D'autres valeurs de α sont également d'usage courant : $\alpha=0,10$ (ou 10%), $\alpha=0,01$ (ou 1%) et $\alpha=0,001$ (ou 0,1%). On doit garder à l'esprit que plus les conséquences du rejet à tort de l'hypothèse nulle sont graves, plus la valeur de α doit être petite. Mais, il ne faut pas perdre de vue que la diminution de la valeur de α entraîne l'augmentation de β . Nous reviendrons sur ce sujet plus bas dans ce paragraphe.

Lorsqu'on rejette l'hypothèse nulle au seuil α , on dit que le test est significatif à $(\alpha \cdot 100) \%$ (à 5% par exemple). Dans ce cas, il faut essayer de voir si on peut toujours rejeter l'hypothèse nulle en prenant des valeurs de α plus petites (comme $\alpha = 0,01$ ou $\alpha = 0,001$).

b) Nous devons, autant que possible, essayer de déterminer le risque d'accepter l'hypothèse nulle lorsqu'elle est fausse, c'est-à-dire le risque de deuxième espèce (β). La valeur de β nous permettra de calculer, comme nous le verrons au paragraphe suivant, la valeur de la puissance.

Généralement, le risque d'accepter l'hypothèse nulle lorsqu'elle est fausse (β) est plus grand que le risque de la rejeter lorsqu'elle est en fait vraie (α).

c) Pour un échantillon d'effectif donné, la valeur de α est inversement reliée avec la valeur de β . Il n'est donc pas possible de vouloir déterminer les régions de rejet et d'acceptation pour minimiser simultanément α et β . Plus petit sera le risque de commettre une erreur de première espèce, plus grand sera le risque de commettre une erreur de seconde espèce. Généralement, les expérimentateurs choisissent de maîtriser le risque α , quitte à ignorer le risque β .

Pour réduire simultanément les deux risques d'erreur, on augmentera les effectifs des échantillons. Nous reviendrons sur ce point précis dans les chapitres qui suivent.

2.2.3.2. Puissance d'un test

Dans la plupart des expériences biologiques, l'expérimentateur désire mettre en évidence l'hypothèse alternative H_1 . C'est pourquoi il s'intéresse à la quantité $1 - \beta$ qui représente la probabilité d'obtenir le résultat qu'il souhaite démontrer.

Cette probabilité, qui reflète par exemple la capacité de détecter une différence réelle, s'appelle la **puissance du test**. Il s'agit de la probabilité de rejeter H_0 en ayant raison :

$$1 - \beta = P(\text{rejeter } H_0 / H_0 \text{ fausse})$$

On retient aussi que la puissance d'un test augmente avec la taille de l'échantillon.

Le calcul de la puissance d'un test est souvent une opération complexe, ce qui fait que l'utilisateur se trouve fréquemment dans une situation de ne pas pouvoir calculer cette puissance. La difficulté résulte essentiellement de l'hypothèse alternative qui est vague.

Pour comprendre les deux sortes de risques d'erreur et la puissance, considérons les hypothèses nulle et alternative suivantes :

H_0 : Innocence d'un accusé et **H_1** : Culpabilité d'un accusé

- | | |
|---|--------------------------------------|
| – L'erreur de première espèce : | on condamne un innocent |
| – L'erreur de deuxième espèce : | on acquitte un coupable |
| – Le risque de première espèce α : | risque de condamner un innocent |
| – Le risque de deuxième espèce β : | risque d'acquitter un coupable |
| – La puissance : | probabilité de condamner un coupable |

2.2.3.3. Exemple

Considérons l'exemple de la comparaison des teneurs moyennes théoriques en protéines sériques totales du sang chez des vaches malades (μ_M) et des vache saines (μ_S).

a) Illustration du risque de première espèce

1) Dans le cas d'un test bilatéral, les hypothèses nulle et alternative s'écrivent :

$$H_0 : \mu_M = \mu_S \text{ contre } H_1 : \mu_M \neq \mu_S.$$

Supposons que la teneur en protéines chez des vaches malades suit une distribution normale de moyenne (μ_M) et que la teneur en protéines chez les vaches saines suit une distribution normale de moyenne (μ_S) et que les deux distributions sont indépendantes et ont la même variance (σ^2). Supposons aussi que les teneurs en protéines moyennes chez les vaches malades et saines **sont égales**, c'est-à-dire que l'hypothèse nulle **H_0 est vraie**.

Dans ce cas, si l'on réalise une première expérience en prélevant un échantillon aléatoire et simple de vaches malades et un autre échantillon aléatoire et simple de vaches saines, on peut calculer la différence d_1 entre les moyennes observées \bar{x}_M et \bar{x}_S des deux échantillons. Si on réalise cette expérience un très grand nombre de fois, les différences observées d_1, d_2, \dots vont se distribuer selon une loi normale de moyenne nulle. La distribution normale peut être réduite en divisant les différences par leur écart-type.

On peut constater, d'après la figure 2.1, que certaines différences d_i peuvent être grandes en valeur absolue. La probabilité d'apparition de ces différences est très faible puisqu'on a supposé que les teneurs moyennes en protéines chez les vaches saines et les vaches malades **sont égales**. Dans ces cas extrêmes, on rejette à tort l'hypothèse nulle. Le seuil de probabilité α , dit aussi risque de première espèce, est le risque de conclure faussement que les teneurs moyennes en protéines chez les vaches saines et les vaches malades sont différentes, alors que l'existence de cette différence n'est que le fait du hasard.

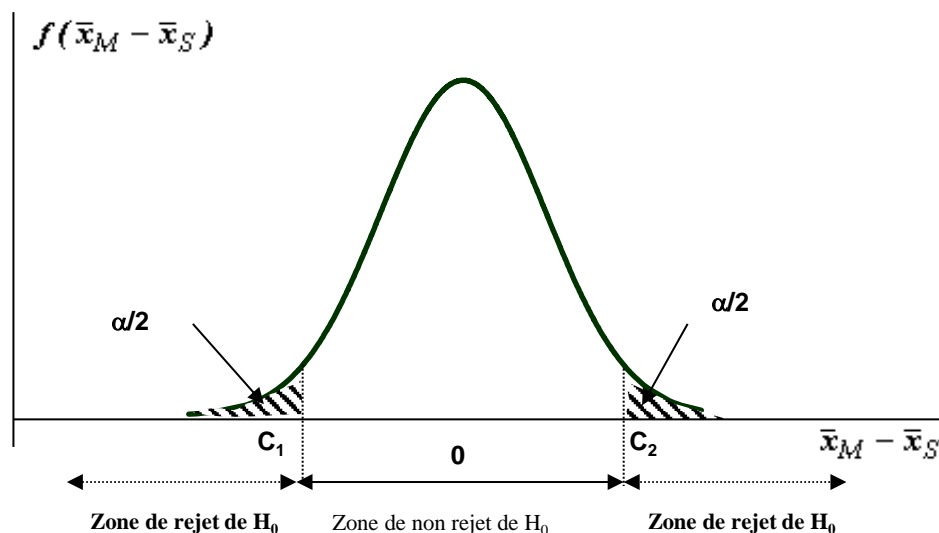


Figure 2.1. Distribution des différences des moyennes dans le cas de la comparaison de deux moyennes et surface de rejet de H_0 (en hachuré) : cas du test bilatéral.

Le test étant bilatéral, il y a eu donc deux surfaces de rejet situées, comme le montre la figure 2.1, aux deux extrémités de la distribution et chacune a une aire de $\alpha/2$. C_1 et C_2 représentent les valeurs critiques des différences qui spécifient les limites des zones de rejet de l'hypothèse nulle.

Si l'on fixe α à 5% et si l'hypothèse nulle est vraie, on retient qu'une différence d_i entre les deux moyennes se situe dans la zone critique (zone de rejet) dans 5% des cas.

2) Si l'on prévoit que la teneur moyenne en protéines sériques est plus élevée chez les vaches malades, les hypothèses nulle et alternative s'écrivent :

$$H_0 : \mu_M = \mu_S \text{ contre } H_1 : \mu_M > \mu_S.$$

Dans ce cas, le test est unilatéral et le risque α ne se trouve qu'à une seule extrémité de la distribution. Il y a une seule surface de rejet dont l'aire est égale à α et représentée en hachuré à la figure 2.2.

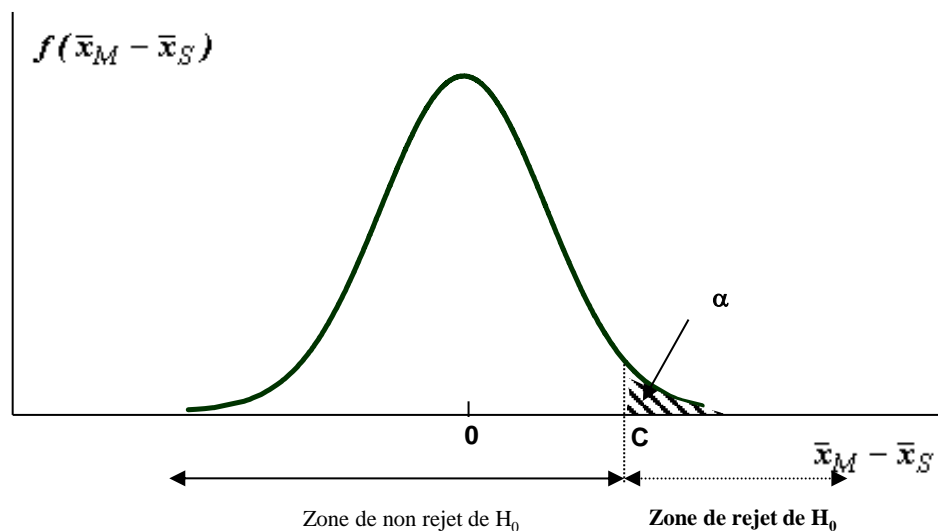


Figure 2.2. Distribution des différences des moyennes dans le cas de la comparaison de deux moyennes et surface de rejet de H_0 (en hachuré) : cas du test unilatéral.

3) Dans les deux cas, si la différence observée entre les deux moyennes n'appartient pas à la zone du rejet, c'est-à-dire sa probabilité de réalisation est supérieure à α , on ne rejette pas l'hypothèse nulle et on dit que le test n'est pas significatif. Dans le cas contraire, c'est-à-dire si la différence observée entre les deux moyennes appartient à la zone du rejet, on dit que le test est significatif.

b) Illustration du risque de deuxième espèce

Supposons maintenant que l'hypothèse nulle est fausse, c'est-à-dire que la différence théorique $\mu_M - \mu_S$ est égale à la valeur δ qui est **différente de zéro**.

Dans les conditions de la normalité et de l'indépendance des deux populations et si l'on répète l'expérience un nombre élevé de fois, les différences obtenues des différentes expériences vont se distribuer cette fois selon une loi normale de moyenne δ . La distribution normale peut être réduite en divisant les différences par leur écart-type. Cette situation est illustrée à la figure 2.3 par la distribution H_1 , tout en gardant la distribution sous H_0 de la figure 2.2 où l'on a considéré que le test est unilatéral.

Nous ne connaissons pas la distribution sous H_1 , car nous ne connaissons pas δ . Nous l'avons placée à droite de la distribution sous H_0 , car nous avons estimé que les vaches malades ont une teneur plus élevée en protéines sériques.

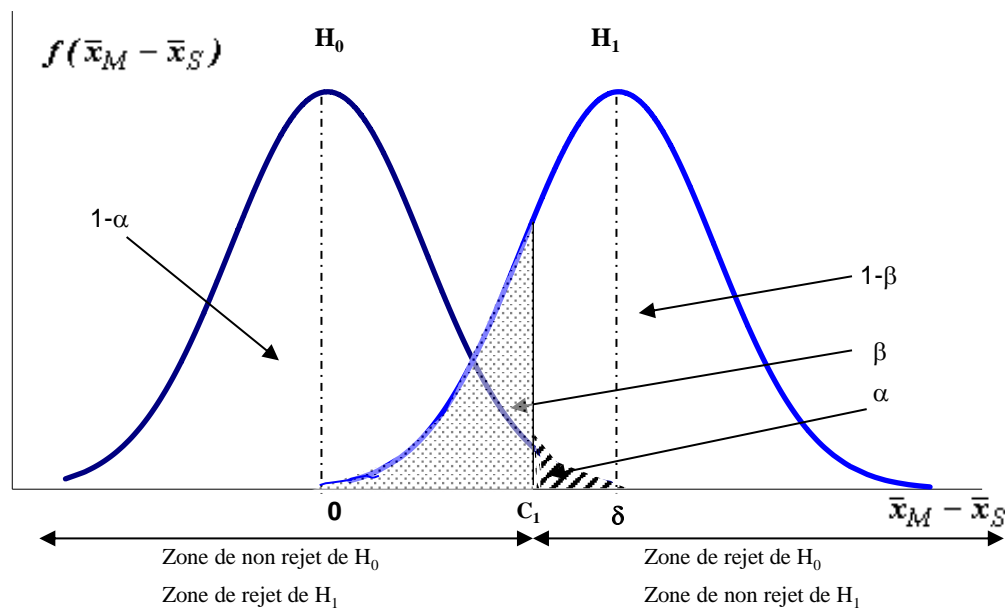


Figure 2.3. Distributions des différences des moyennes dans le cas où les deux moyennes sont égales (H_0) et dans le cas où elles sont différentes (H_1).

On voit qu'il y a un risque d'accepter l'hypothèse nulle alors qu'elle est fausse (aire en pointillée) : c'est le risque β de deuxième espèce. Il s'agit de la probabilité de ne pas déceler une différence qui, en réalité, existe. C'est le cas lorsque les deux moyennes sont différentes mais la valeur de la différence d_i n'était pas suffisamment grande pour rejeter l'hypothèse nulle.

On peut aussi voir sur la même figure qu'une diminution du risque de première espèce α s'accompagne d'une augmentation du risque de seconde espèce β et donc d'une diminution de la puissance du test.

2.2.4. Choix de la statistique du test

Pour réaliser un test statistique, on utilise une statistique S pour contrôler l'hypothèse nulle. C'est une fonction des variables aléatoires représentant l'échantillon dont on connaît la loi de probabilité quand l'hypothèse nulle est vraie et qui permettra de prendre une décision à propos de l'hypothèse nulle. La statistique S peut être une variable aléatoire normale, de Student (t), Khi carrée (χ^2), F de Snedecor, etc.

Le choix de la statistique S dépend, entre autres, du type de l'hypothèse à tester, de la nature des données (quantitatives, semi-quantitatives, qualitatives) et parfois aussi de l'importance des effectifs des échantillons et de la vérification ou non de certaines conditions relatives aux populations étudiées (normalité, égalités des variances, etc.). Nous reviendrons longuement sur ce sujet dans les prochains chapitres de ce document.

2.2.5. Détermination de la région critique

L'ensemble des valeurs observées de la statistique S provoquant le rejet de l'hypothèse nulle constitue la zone de rejet ou la région critique du test statistique. Le reste des valeurs forme la zone d'acceptation de l'hypothèse nulle. Cela suppose que le seuil de signification α , la distribution de probabilité de la variable S et le type de test (unilatéral ou bilatéral) aient été déterminés auparavant.

Toutes choses égales par ailleurs, la région critique diminue lorsque α diminue, c'est-à-dire que l'hypothèse nulle est moins fréquemment rejetée. Comme nous l'avons annoncé auparavant, lorsqu'on rejette l'hypothèse nulle avec un niveau de signification α donné ($\alpha=0,05$ par exemple), on essaiera de voir si on rejette cette hypothèse avec un niveau de signification α plus petit ($\alpha=0,01$ par exemple).

2.2.6. Règle de décision

Les observations obtenues des échantillons apportent ou non la preuve que l'hypothèse nulle doit être rejetée. On rejette ou non l'hypothèse nulle après avoir comparé les valeurs observée (S_{obs}) et critique (S_{critique}) de la statistique S :

- Si S_{obs} appartient à la région critique du test au niveau α , l'hypothèse nulle est rejetée ;
- Si S_{obs} appartient à la région d'acceptation du test au niveau α , dites que vous ne pouvez pas conclure ou que vous ne rejetez pas l'hypothèse nulle.

En effet, l'acceptation de l'hypothèse nulle permet de dire que les observations recueillies ne montrent pas que H_0 est fausse. Autrement dit, rien ne s'est produit qui permette de la mettre en doute. Il ne faut pas en déduire automatiquement que H_0 est vraie, sans avoir pris connaissance du risque de deuxième espèce (β) ou de la puissance ($1-\beta$). Si la valeur de β est égale à 0,52 par exemple, cela veut dire qu'on a 52% de chance de se tromper en déclarant que l'hypothèse nulle est vraie, la puissance n'étant que de 48% dans le cas de cet exemple.

2.2.7. Degré de signification

Le test se réalise avec un risque α . Si l'hypothèse nulle est rejetée, on évalue le degré de signification ou la probabilité d'erreur du test qu'on note p ou *p-value* en anglais. Plus p est petit, plus confortable est la conclusion de rejeter l'hypothèse nulle.

Il s'agit d'une probabilité *a posteriori*. Lorsqu'on rejette l'hypothèse nulle, elle correspond au risque de première espèce qui aurait permis avec les mêmes données de rejeter l'hypothèse nulle.

La plupart des logiciels statistiques donnent cette probabilité à la fin de chaque test. De même, certains tableurs (Excel, par exemple) permettent d'obtenir cette probabilité.

2.2.8. Interprétation des résultats

On termine les étapes par l'interprétation des résultats et la formulation de conclusions pratiques en des termes compatibles avec la nature du problème initialement posé. Nous apprendrons à le faire au travers des exemples des chapitres qui vont suivre.

2.3. TYPES DE TESTS

Selon le problème posé, on peut distinguer différents types de tests. On peut citer les tests de conformité, les tests d'égalité, les tests d'ajustement, les tests d'indépendance, etc. L'objectif de ce paragraphe est de donner une idée générale des plus connus d'entre eux, sachant qu'ils seront vus en détail dans les autres chapitres.

2.3.1. Tests de conformité

Le but de ce test est de vérifier si une population possède une valeur spécifiée d'un paramètre statistique. Ce paramètre peut être la moyenne, la variance, le coefficient de régression, etc.

Exemple :

- | | | |
|---|--------|--------------------------------|
| a) $H_0 : \mu = 7500 \text{ kg}$ | contre | $H_1 : \mu \neq 7500 .$ |
| b) $H_0 : \beta_{y,x} = 4.5 \text{ mg/h}$ | contre | $H_1 : \beta_{y,x} \neq 4.5 .$ |

L'hypothèse nulle H_0 signifie par exemple que la production laitière théorique de la population est de 7500 kg par vache et par an et l'hypothèse nulle H_0' signifie par exemple que la concentration en un élément chimique donné augmente théoriquement de 4.5 milligrammes par heure.

2.3.2. Tests d'égalité

Le but est de comparer deux ou plusieurs populations vis-à-vis d'un paramètre statistique. Ce paramètre peut être la moyenne, la variance, le coefficient de régression, etc.

Exemple :

- | | | |
|--|--------|---|
| a) $H_0 : \mu_1 = \mu_2 = \mu_3$ | contre | $H_1 : \text{au moins une moyenne est différente des autres}$ |
| b) $H_0' : \beta_{y,x} = \beta_{y,x}'$ | contre | $H_1' : \beta_{y,x} \neq \beta_{y,x}'$ |

L'hypothèse nulle H_0 signifie que les moyennes théoriques des trois populations sont égales (exemple : productions laitières de trois races bovines) et l'hypothèse nulle H_0' signifie que les coefficients de régressions théorique des deux populations sont égaux (le pourcentage de germination par jour d'une espèce végétale augmente de la même manière dans deux localités différentes).

2.3.3. Tests d'ajustement

Le but est de vérifier si un échantillon peut être considéré comme extrait d'une population ayant une distribution donnée.

Exemple :

- | | | |
|---|--------|--|
| a) H_0 : la population est normale | contre | H_1 : la population n'est pas normale |
| b) H_0' : la population est binomiale | contre | H_1' : la population n'est pas binomiale |

Dans le cas de H_0 , on peut considérer l'exemple d'une machine qui remplit les caisses de farines en se posant la question si le remplissage se fasse selon une loi de Gausse de moyenne 50 kg.

2.3.4. Tests d'indépendance

Le but est de vérifier l'indépendance de deux ou plusieurs critères de classification, c'est-à-dire de deux ou plusieurs caractères, généralement qualitatifs.

Exemple :

- a) H_0 : il y a indépendance entre le froid hivernal et la pourriture des fruits de poires
 H_1 : il n'y a pas d'indépendance entre ces deux critères.
- b) H_0 : il y a indépendance entre la filière de formation suivie par les étudiants et la catégorie socioprofessionnelle de leurs pères
 H_1 : il n'y a pas d'indépendance entre ces deux critères.

2.4. EXEMPLE D'APPLICATION

2.4.1. Enoncé

Les teneurs du sang en bêta globulines de 10 bovins non gestantes et de 10 bovins au huitième mois de gestation sont données dans le tableau 2.2 (en g/l). Les deux échantillons sont indépendants et prélevés de manière aléatoire et simple parmi les bovins adultes de la même race. Toutes les données concernent des vaches différentes. Peut-on affirmer, au seuil $\alpha=0,05$, que la gestation influence les teneurs en bêta globulines ?

Tableau 2.1. Teneurs en bêta globulines obtenues pour l'échantillon de vaches gestantes et l'échantillon de vaches non gestantes.

Vaches non gestantes	7.6	7.7	6.7	9.5	8.4	9.8	7.9	7.3	9.4	7.8
Vaches gestantes	10.2	8.6	8.8	8.0	7.4	9.7	9.8	8.7	9.1	6.7

Pour simplifier, nous considérons que les variances des deux populations sont **égales et connues** ($\sigma^2=1$ (g/l)²), ce qui est rarement le cas dans la pratique. Au chapitre suivant, nous traiterons le même type de problèmes sans cette hypothèse contraignante.

2.4.2. Solution

Dans cet exercice, on souhaite savoir si l'on peut affirmer que les teneurs du sang en bêta globulines sont les mêmes chez les vaches non gestantes et les vaches au huitième mois de gestation. Pour ce faire, on va utiliser des données de deux échantillons de vaches pour tester l'hypothèse sur les deux populations. On a utilisé deux échantillons de dix observations chacun.

- a) **Méthode** : il s'agit de la comparaison des moyennes de deux populations de même variance (cette variance est supposée connue).

- Première population : population de vaches non gestantes (la moyenne est μ_n) ;
- Deuxième population : population de vaches gestantes (la moyenne est μ_g) ;
- Variable mesurée : teneur du sang en bêta globulines.

- b) **Conditions d'application** :

- Pour chaque population, on suppose que la teneur du sang en bêta globulines suit une distribution normale ;

- Chaque échantillon de 10 vaches est prélevé de manière aléatoire et simple dans chaque population ;
- Les deux échantillons sont indépendants (pas de relation entre vaches gestantes et vaches non gestantes) ;
- Les deux populations ont la même variance σ^2 .

c) Hypothèses nulle et alternatives :

$$H_0 : \mu_n = \mu_g \quad \text{contre} \quad H_1 : \mu_n \neq \mu_g$$

d) Raisonnement :

Soient :

\bar{x}_n : la moyenne de l'échantillon des 10 vaches non gestantes

\bar{x}_g : la moyenne de l'échantillon des 10 vaches gestantes

On pense que si μ_n et μ_g sont égales il y a une forte probabilité que \bar{x}_n et \bar{x}_g soient proches.

D'autre part :

- \bar{x}_n est associée à la variable \bar{X}_n , cette variable suit une distribution normale de moyenne μ_n et d'écart-type σ / \sqrt{n} ;
- \bar{x}_g est associée à la variable \bar{X}_g , cette variable suit une distribution normale de moyenne μ_g et d'écart-type σ / \sqrt{n} .

Il en découle, en raison des conditions de la normalité et de l'indépendance des échantillons, que la variable $(\bar{X}_n - \bar{X}_g)$ suit une distribution normale de moyenne $(\mu_n - \mu_g)$ et d'écart-type $\sqrt{\sigma^2/n + \sigma^2/n} = \sqrt{2\sigma^2/n}$.

Si l'hypothèse nulle est vraie alors :

$$\frac{\bar{X}_n - \bar{X}_g}{\sqrt{2\sigma^2/n}}$$

suit une distribution normale centrée et réduite.

La probabilité d'observer une différence qui est au moins égale à $|\bar{x}_n - \bar{x}_g|$ est :

$$P\left[|\bar{X}_n - \bar{X}_g| \geq |\bar{x}_n - \bar{x}_g|\right].$$

On rejette l'hypothèse nulle, comme peu vraisemblable, si cette probabilité est inférieure ou égale au risque α :

$$P\left[|\bar{X}_n - \bar{X}_g| \geq |\bar{x}_n - \bar{x}_g|\right] \leq \alpha$$

Or, on a :

$$\begin{aligned}
P\left[|\bar{X}_n - \bar{X}_g| \geq |\bar{x}_n - \bar{x}_g|\right] &= P\left[|U| \geq \frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}}\right] \\
&= 2P\left[U \geq \frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}}\right] \\
&= 2\left[1 - \Phi\left(\frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}}\right)\right]
\end{aligned}$$

ce qui donne :

$$\Phi\left(\frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}}\right) \geq 1 - \alpha/2$$

ou encore :

$$\frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}} \geq U_{1-\alpha/2}.$$

La quantité :

$$U_{\text{obs}} = \frac{|\bar{x}_n - \bar{x}_g|}{\sigma\sqrt{2/n}}$$

est appelée la **valeur observée de la statistique U**. On vient donc de démontrer que l'hypothèse nulle est rejetée, au niveau de signification α , lorsque :

$$U_{\text{obs}} \geq U_{1-\alpha/2}.$$

Reprenons l'exemple précédent sachant que l'écart-type vaut 1 g/l. En comparant les valeurs observée et théorique de la statistique U, on a, pour $\alpha=0.05$:

$$U_{\text{obs}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma\sqrt{2/n}} = \frac{|8.700 - 8.210|}{\sqrt{2/10}} = 1.096$$

et $U_{0.975} = 1.96$. On est donc amené à ne pas rejeter l'hypothèse nulle, car la valeur observée U_{obs} est inférieure à la valeur théorique $U_{1-\alpha/2}$.

On peut passer par le calcul de la probabilité d'observer une différence qui est au moins égale à $|\bar{x}_n - \bar{x}_g| = |8.700 - 8.210| = 0.49$ g/l. On a :

$$\begin{aligned}
P\left[|\bar{X}_n - \bar{X}_g| \geq |0.49|\right] &= P\left[|U| \geq \frac{|0.49|}{\sqrt{2/10}}\right] = 2[1 - \Phi(1.096)] \\
&= 2(1 - 0.8634) = 0.273 \text{ ou } 27\%.
\end{aligned}$$

Il ressort qu'on a 27 chances sur 100 d'obtenir, par le simple fait du hasard, une différence au moins aussi importante entre les deux moyennes observées. On n'a pas pu mettre en évidence

des différences significatives, mais n'oublions pas que l'effectif est petit. Il se peut qu'il ne soit pas suffisant pour montrer que l'hypothèse nulle est fausse.

2.4.3. Remarque

a) Supposons maintenant que l'on souhaite tester (test unilatéral) :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 > \mu_2 .$$

Nous rejetons l'hypothèse nulle lorsque :

$$P[\bar{X}_1 - \bar{X}_2 \geq \bar{x}_1 - \bar{x}_2] \leq \alpha .$$

On a :

$$P[\bar{X}_1 - \bar{X}_2 \geq \bar{x}_1 - \bar{x}_2] = P\left[U \geq \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}}\right] = 1 - \Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}}\right),$$

ce qui donne :

$$\Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}}\right) \geq 1 - \alpha$$

ou encore :

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}} \geq U_{1-\alpha} .$$

On vient de démontrer que l'hypothèse nulle est rejetée lorsque $\boxed{U_{\text{obs}} \geq U_{1-\alpha}}$ où :

$$U_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}} . \quad (1)$$

b) Supposons maintenant que l'on souhaite tester :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 < \mu_2 .$$

En suivant un raisonnement semblable que le précédent, on rejette l'hypothèse nulle lorsque $U_{\text{obs}} \leq -U_{1-\alpha}$, où la quantité U_{obs} est donnée par l'équation (1).

Résumé du test statistique

Pour tester l'hypothèse d'égalité de deux moyennes lorsque l'écart-type est connu au seuil de signification α :	
$H_0 : \mu_1 = \mu_2$	
Calculer la valeur observée :	$U_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}}$
a) Si $H_1 : \mu_1 \neq \mu_2$, rejeter H_0 lorsque :	$ U_{\text{obs}} \geq U_{1-\alpha/2}$
b) Si $H_1 : \mu_1 > \mu_2$, rejeter H_0 lorsque :	$U_{\text{obs}} \geq U_{1-\alpha}$
c) Si $H_1 : \mu_1 < \mu_2$, rejeter H_0 lorsque :	$U_{\text{obs}} \leq -U_{1-\alpha}$

Attention : en plus des conditions d'application classiques, ce test est appliqué lorsque la variance σ^2 est connue. Sinon, les effectifs doivent être suffisamment grands.

Remarque : le chapitre 3 traitera le cas où la variance σ^2 n'est pas connue.

2.5. CONCLUSION

Nous avons vu qu'un test statistique est un procédé de l'inférence statistique dont le but est de tester si une hypothèse est vraie ou fausse afin d'appréhender la réalité que l'on ne connaît pas. La formulation de l'hypothèse nulle est souvent une interprétation de la question posée. Cette hypothèse est privilégiée par rapport à l'hypothèse alternative. Le test se réalise, comme nous l'avons vu, en suivant une succession d'étapes bien définies et il exige parfois la vérification au préalable d'un certain nombre de conditions d'application. A ce propos, nous avons mentionné la robustesse d'un test en disant qu'un test est plus ou moins robuste en fonction de sa tolérance vis à vis du respect de ces conditions

Nous avons aussi souligné que le raisonnement est statistique, on ne peut donc jamais être certain que notre décision d'accepter ou de rejeter l'hypothèse nulle reflète correctement la réalité. Ceci est dû à deux types d'erreurs, l'un dit de première espèce et l'autre de seconde espèce. Nous avons vu au travers de l'exemple de la comparaison de deux moyennes que le premier consiste à conclure faussement à l'existence d'une différence et le second à ne pas déceler une différence existante. Nous avons vu que les deux types de risques sont liés. Il y a donc lieu de tenir compte de l'importance relative de chaque type d'erreur pour l'étude menée.

Dans la pratique cependant, bon nombre de chercheurs s'attachent à contrôler le risque de première espèce en le fixant le plus souvent à 5% ou 1% et ont donc souvent tendance à ignorer le risque, pourtant important, de commettre une erreur de seconde espèce. Cette attitude est compréhensible pour les problèmes relativement compliqués, mais, pour d'autres problèmes, des approches et des logiciels statistiques sont disponibles dans la littérature statistique et le commerce pour le calcul du risque de seconde espèce.

Enfin nous avons défini la puissance comme étant la probabilité de rejeter l'hypothèse nulle lorsqu'elle est fausse. Nous avons souligné que le non rejet de l'hypothèse nulle peut être due au fait que cette hypothèse est vraie ou que la puissance du test n'est pas suffisante. Ainsi, il est souvent question de vouloir calculer un effectif nécessaire pour obtenir une puissance donnée.

CHAPITRE 3

INFÉRENCES RELATIVES À UNE ET DEUX MOYENNES

3.1. INTRODUCTION

Les méthodes relatives aux moyennes se comptent parmi les plus simples et les plus utilisées par les expérimentateurs pour analyser des données quantitatives. La plupart du temps, il s'agit d'obtenir des intervalles de confiance et de réaliser des tests d'hypothèses sur les moyennes de populations.

Nous avons déjà entamé cette inférence lorsqu'on a calculé l'intervalle de confiance et comparé deux moyennes en considérant que les écarts-types des populations sont connus (*cf.* paragraphes 1.5 et 2.4). Mais, souvent dans la pratique, l'expérimentateur ne connaît pas les écarts-types des populations et il est amené à les estimer à partir des données expérimentales.

Ce chapitre est d'abord consacré à l'étude de l'inférence relative à une moyenne lorsque l'écart-type est inconnu. Nous verrons comment obtenir une estimation de la moyenne théorique de la population et l'entourer de limites de confiances. C'est le cas par exemple lorsqu'on souhaite estimer la teneur en protéines sériques chez des vaches gestantes de huit mois en observant un échantillon aléatoire et simple de vaches sélectionnées dans la population étudiée et chercher les limites de confiance dont on a de fortes chances de croire qu'elles encadrent la vraie teneur moyenne. Nous verrons aussi comment tester si une moyenne théorique est inférieure, égale ou supérieure à une valeur de référence donnée. C'est le cas par exemple lorsqu'on désire tester si le gain de poids de taurillons recevant une alimentation spécifiée est au moins égal à une valeur de référence donnée (paragraphe 3.2).

Il s'intéresse ensuite à évaluer les différences entre les effets de deux traitements (groupes). Ce problème est très fréquent en recherches scientifiques, notamment en sciences agronomiques, en médecines humaine et vétérinaire et en industrie agro-alimentaire. Les deux groupes peuvent être formés naturellement ou expérimentalement. Nous verrons qu'il est primordial de distinguer entre deux classes de tests selon que les deux échantillons sont indépendants ou associés par paires. C'est le cas par exemple lorsqu'on désire tester si deux méthodes de dosage de la teneur en albumine sérique de bovins donnent les mêmes résultats, en dosant un certain nombre de prélèvements par les deux méthodes. L'objectif étant de voir si la différence entre les moyennes observées est due à une vraie différence des effets des deux méthodes. Ce genre de problèmes sera étudié au paragraphe 3.3.

Ce chapitre se termine par la présentation des principales méthodes non paramétriques qui peuvent être utilisées lorsque certaines conditions d'application des tests paramétriques sont loin d'être vérifiées (paragraphe 3.4).

3.2. INFÉRENCE CONCERNANT UNE MOYENNE

3.2.1. Exemple introductif

Intéressons-nous à une alimentation standard de dindons avec un additif A et considérons une expérience où l'on a affecté cette alimentation à un échantillon aléatoire et simple de 12 dindons à peu près de même poids initial et élevés dans les mêmes conditions. Après une période standard, les gains de poids ont été enregistrés pour les douze dindons (tableau 3.1).

Tableau 3.1. Gains de poids (en grammes) d'un échantillon de 12 dindons recevant une alimentation standard avec un additif A.

Numéro	1	2	3	4	5	6	7	8	9	10	11	12
Poids	2740	3110	2770	3250	3160	3020	2990	3170	3150	2990	2760	3150

A partir des résultats de cet échantillon, on souhaite :

- obtenir une estimation du gain du poids moyen théorique de dindons nourris par cette alimentation et l'entourer de limites de confiance dont on a de fortes chances $(1-\alpha)$ de croire qu'elles entourent cette moyenne (on prendra $\alpha=5\%$) ;
- vérifier si l'on peut affirmer, au seuil $\alpha=0.05$, que le gain moyen de poids théorique des dindons recevant cette alimentation après une période standard est supérieur à 2.9 kilogrammes.

L'objet de ce paragraphe est d'apprendre à répondre à ce genre de questions.

3.2.2. Estimation et intervalle de confiance d'une moyenne

Nous savons qu'on peut utiliser la variable \bar{X} pour avoir de l'information sur la moyenne inconnue μ de la population et nous savons aussi que lorsqu'un échantillon de taille n est prélevé de manière aléatoire et simple dans une population distribuée selon une loi normale de moyenne μ et de variance σ^2 , la statistique \bar{X} est distribuée selon une loi normale de moyenne μ et de variance σ^2/n .

La variable \bar{X} a une plus grande chance d'être proche de la moyenne de la population μ qu'il ne l'est une observation quelconque de l'échantillon, et plus la taille de l'échantillon est grande plus on a tendance à cerner la moyenne μ .

Il est donc de coutume d'utiliser la moyenne arithmétique \bar{x} d'un l'échantillon aléatoire et simple pour obtenir une estimation non biaisée de la moyenne inconnue μ de la population :

$$\hat{\mu} = \bar{x} \quad (3.1)$$

et de l'entourer de limites de confiance.

Pour obtenir ces limites \bar{X}_1 et \bar{X}_2 , il faut se rappeler que dans les conditions de la normalité de la population et de l'indépendance des observations, la quantité :

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (3.2)$$

est une variable aléatoire normale centrée et réduite, c'est-à-dire de moyenne 0 et de variance 1.

Toutefois, cette expression que nous avons utilisée au paragraphe 2.4 suppose que l'on connaisse la variance de la population, ce qui est assez rare dans la pratique. C'est pourquoi, on cherche plutôt à estimer la variance de la population en utilisant la variance de l'échantillon s^2 selon la relation suivante :

$$\hat{\sigma}^2 = \frac{ns^2}{n-1} = \frac{SCE}{n-1}. \quad (3.3)$$

On s'attend à ce que l'intervalle de confiance de μ soit plus large par rapport au cas où la variance de la population est connue et, ce, à cause de l'erreur qui peut être déduite de l'estimation de σ^2 . Dans ce cas, si la population est normale, un échantillon aléatoire et simple permet de formuler la quantité :

$$t = \frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}} \quad (3.4)$$

qui suit une distribution t de Student à $k=n-1$ degrés de liberté. Les limites de confiance peuvent être déterminées, en se fixant un degré de confiance $1-\alpha$ assez proche de 1 :

$$P(\bar{X}_1 \leq \mu \leq \bar{X}_2) = 1 - \alpha.$$

On peut donc utiliser la table statistique pour calculer la valeur $t_{1-\alpha/2}$ de telle sorte qu'il y ait une probabilité $1-\alpha$ qu'une valeur t tirée au hasard se situe entre $-t_{1-\alpha/2}$ et $+t_{1-\alpha/2}$. Autrement dit, il y a $(1-\alpha)$ de chance que :

$$-t_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}} \leq t_{1-\alpha/2}.$$

Si l'on dispose d'un échantillon aléatoire et simple de n valeurs, on obtient par un simple calcul :

$$\mu \geq \bar{x} - t_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \quad \text{et} \quad \mu \leq \bar{x} + t_{1-\alpha/2} \hat{\sigma} / \sqrt{n}$$

ce qui donne, au degré de confiance $(1-\alpha)$, les valeurs :

$$\boxed{\begin{aligned} \bar{x}_1 &= \bar{x} - t_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \\ \bar{x}_2 &= \bar{x} + t_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \end{aligned}} \quad (3.5)$$

où la variable t de Student possède $k=n-1$ degrés de liberté.

Lorsque l'effectif de l'échantillon est grand, on peut utiliser la même expression 3.5 en remplaçant la distribution t de Student par la distribution normale réduite. Cette approximation est généralement utilisée lorsque l'effectif de l'échantillon atteint une trentaine d'observations ($n \geq 30$).

On constate que la largeur de l'intervalle de confiance augmente avec la variabilité des observations dans la population (σ^2) et avec le degré de confiance ($1-\alpha$) et diminue avec l'effectif de l'échantillon (n).

La longueur l de l'intervalle de confiance peut être calculée par l'expression :

$$l = 2t_{1-\alpha/2} \hat{\sigma} / \sqrt{n} . \quad (3.6)$$

La demi-longueur de l'intervalle de confiance, c'est-à-dire $l/2$, est appelée **marge d'erreur**.

Exemple 3.1

Reprenez les données du tableau 3.1 et cherchez une estimation du gain moyen de poids de la population des dindons recevant l'alimentation en question et son intervalle de confiance (prendre un degré de certitude de 95%) ?

Solution 3.1

a) Méthode : estimation et intervalle de confiance d'une moyenne, la variance est inconnue

b) Conditions d'application :

- L'échantillon des douze dindons est prélevé de manière aléatoire et simple dans la population des dindons ayant les mêmes caractéristiques ;
- La distribution de la population des gains de poids est normale : Nous verrons le test de la normalité sur un petit échantillon dans le chapitre « Tests d'ajustement ».

c) Calcul

La moyenne de la population est estimée par :

$$\hat{\mu} = (2740 + 3110 + \dots + 3150) / 12 = 3021.7 \text{ g ou } 3022 \text{ g}$$

et son intervalle de confiance, au degré de confiance 95% ($\alpha=0.05$), est :

$$\begin{aligned} \bar{x}_1 &= 3021.67 - 2.201(177.65) / \sqrt{12} = 2908.8 \\ \bar{x}_2 &= 3021.67 + 2.201(177.65) / \sqrt{12} = 3134.5 \end{aligned}$$

Il y a 95% de chance que les valeurs **2909 et 3135 g** encadrent le gain de poids moyen théorique μ .

Remarque

Souvent, on est amené à estimer la taille de l'échantillon minimum n pour estimer la moyenne de la population μ avec une précision donnée. Cette précision peut être exprimée, entre autres, en termes d'intervalle de confiance. On peut par exemple chercher à estimer la moyenne avec une marge d'erreur d fixée par l'expérimentateur.

On peut alors écrire :

$$\bar{x} \pm d .$$

Puisque $d = t_{1-\alpha/2} \hat{\sigma} / \sqrt{n}$, on obtient :

$$n \geq \frac{t_{1-\alpha/2}^2 \hat{\sigma}^2}{d^2} \quad (3.7)$$

où l'estimation de la variabilité dans la population $\hat{\sigma}^2$ peut être tirée d'une étude préliminaire sur une population similaire ou d'un échantillonnage pilote, même d'un effectif très réduit. Souvent le chercheur prend la valeur 4 pour $t_{1-\alpha/2}^2$.

L'expression (3.7) peut s'exprimer aussi sous la forme :

$$n \geq \frac{t_{1-\alpha/2}^2 cv^2}{d_r^2} \quad (3.8)$$

où cv désigne le coefficient de variation ($cv = (\hat{\sigma} / \bar{x}) (100)$). L'expérimentateur a souvent une idée sur ce coefficient pour la variable qu'il étudie. La marge d'erreur relative d_r , c'est-à-dire l'erreur maximum d exprimée en % de la moyenne ($d_r = (d / \bar{x}) (100)$) est fixée par l'expérimentateur pour pouvoir déterminer l'effectif.

Exemple 3.2

On désire obtenir une estimation du gain de poids moyen de la population de dindons nourris par une alimentation standard avec l'additif A (cf. paragraphe 3.2), en lui donnant un intervalle de confiance. On se demande quelle est la taille minimum de l'échantillon si l'on souhaite que la demi-longueur de cet intervalle soit ≤ 70 g.

Solution 3.2

a) Méthode : détermination de l'effectif de l'échantillon pour estimer la moyenne avec une précision donnée.

b) Conditions d'application :

- la distribution de la population des gains de poids est normale
- l'échantillon prévu est aléatoire et simple

c) Calcul

Pour pouvoir estimer l'effectif minimum, on doit disposer d'une estimation de la variance ou de l'écart-type des gains de poids. On peut utiliser l'estimation de l'écart-type obtenue à partir de l'échantillon de données du tableau 3.1, soit $\hat{\sigma} = 178$ g.

On a $t_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \leq 70$, ce qui donne $n \geq \frac{t_{1-\alpha/2}^2 \hat{\sigma}^2}{70^2}$.

On peut prendre $t_{1-\alpha/2}^2 = 4$, sinon on commence par donner une valeur arbitraire à n , puis on calcule n minimum par la formule précédente. Si on donne à n la valeur 20, on a $t_{1-\alpha/2} = 2.093$ et $n \geq \frac{(2.093)^2 (178)^2}{(70)^2}$ ou $n \geq 29$ après avoir arrondi le résultat. On reprend le calcul en considérant $n=29$,

on obtient : $n \geq (2.048)^2 (178)^2 / (70)^2$, c'est-à-dire $n \geq 28$. Si on continue le calcul de la même manière, on remarquera que la valeur se stabilisera à $n \geq 28$.

3.2.3. Test de conformité d'une moyenne

L'objectif de ce test est de vérifier si la moyenne théorique d'une population μ est égale à une moyenne hypothétique μ_0 (une constante). L'hypothèse nulle s'écrit :

$$H_0 : \mu = \mu_0 \quad (3.9)$$

contre l'une ou l'autre des trois hypothèses alternatives usuelles :

a) $H_1 : \mu > \mu_0$

b) $H_1' : \mu < \mu_0$

c) $H_1'' : \mu \neq \mu_0$.

En disposant d'un échantillon aléatoire et simple tiré dans la population, le test consiste, sous la condition de la normalité, à calculer la valeur observée de la variable t de Student :

$$t_{obs} = \frac{\bar{x} - \mu_0}{\hat{\sigma} / \sqrt{n}} \quad (3.10)$$

et, pour un niveau de signification α , on est amené à :

a) rejeter H_0 en faveur de H_1 si $t_{obs} > t_{1-\alpha}$

b) rejeter H_0 en faveur de H_1' si $t_{obs} < -t_{1-\alpha}$

c) rejeter H_0 en faveur de H_1'' si $|t_{obs}| > t_{1-\alpha/2}$

où la variable t de Student possède $k=n-1$ degrés de liberté.

On constate que la décision du test dépend de l'hypothèse alternative posée. Si on opte pour un test unilatéral, il ne sera pas possible d'identifier de différences significatives dans le sens opposé à celui qui a été prédit.

Exemple 3.3

Reprenez l'exemple de l'élevage des dindons. Dans le passé, les dindons recevant une alimentation standard montraient pour une même période et dans les mêmes conditions un gain moyen de poids de 2900 grammes. En utilisant les données du tableau 3.1, peut-on affirmer que le gain moyen de poids théorique des dindons recevant cette alimentation avec l'additif A est supérieur à 2900 grammes.

Solution 3.3

a) **Méthode** : test de conformité d'une moyenne

b) **Conditions d'application** :

- la distribution de la population des gains de poids des dindons est normale
- l'échantillon des 12 dindons est prélevé de manière aléatoire et simple

c) **Hypothèses**

$$H_0 : \mu = 2900 \text{ g} \quad \text{contre} \quad H_1' : \mu > 2900 \text{ g}$$

d) Calcul

$$t_{obs} = \frac{3021.67 - 2900}{177.65 / \sqrt{12}} = 2.372 \text{ ou } 2.37.$$

Pour un risque de 5%, cette valeur est supérieure à la valeur de t de Student à 11 degrés de liberté ($t_{1-\alpha} = 1.796$), ce qui conduit au **rejet de l'hypothèse nulle**. Voyons si on peut rejeter H_0 au niveau de signification 1% : on a $t_{1-\alpha} = 2.718$, ce qui conduit au non rejet de H_0 . On conclut donc que le test est significatif.

Remarque : les logiciels statistiques nous fournissent le degré de signification désigné souvent par ***p-value*** (cf. paragraphe 2.2.7) Nous avons signalé que plus ***p-value*** est petit, plus confortable est la conclusion de **rejeter l'hypothèse nulle**. La valeur de ***p-value***, qui est ici égale à **0.019**, est inférieure à $\alpha=0.05$, mais supérieure à 0.01. H_0 est donc rejetée au niveau 0.05 et pas au niveau 0.1.

Cette probabilité ***p-value*** peut être aussi obtenue par le tableur Excel en utilisant la fonction ***Loi.Student*** (2.372 ; 11 ; 1), où 11 représente le nombre de degrés de liberté et 1 indique que le test est unilatéral (on peut aussi utiliser ***Loi.Student.Droite*** (2.372 ; 11)).

Conclusion : En prenant un risque de 5%, on peut conclure que le gain moyen de poids des dindons recevant la nourriture en question avec l'additif A est supérieur à 2.9 kg.

3.3. INFÉRENCE CONCERNANT DEUX MOYENNES

3.3.1. Généralités

On s'intéresse dans ce paragraphe aux moyennes de deux populations, ce qui est souvent plus intéressant puisque l'expérimentateur cherche généralement à évaluer les différences entre les effets de deux traitements. L'hypothèse à tester est que la différence entre les moyennes des deux traitements est égale à une constante. Dans la plupart des applications, cette constante est égale à zéro, ce qui permet de tester si les deux moyennes théoriques sont égales.

Pour tester l'hypothèse d'égalité de deux moyennes, nous distinguerons deux situations que nous exposerons en détail dans les paragraphes 3.3.2 et 3.3.3 :

1. La première situation concerne deux échantillons prélevés indépendamment l'un de l'autre, chacun étant prélevé dans une population. Autrement dit, une observation quelconque relative à un échantillon n'est en aucune manière liée à une autre observation de l'autre échantillon (paragraphe 3.3.2) ;
2. La deuxième situation concerne deux échantillons dépendants (dits aussi associés par paires ou par couples). Dans ce cas, chaque observation du premier échantillon est liée à une observation du second échantillon. Il s'agit le plus souvent d'observations obtenues sur les mêmes unités, mais nous verrons que ce n'est toujours pas le cas (paragraphe 3.3.3).

La confusion entre ces deux situations est une erreur, car le test statistique à utiliser dans le cas de chacune d'elles est différent de l'autre. Chacun de ces deux tests exige la normalité.

Considérons deux populations dont les moyennes sont μ_1 et μ_2 , et plaçons-nous dans les conditions de la normalité et de l'échantillonnage aléatoire et simple. La variable $\bar{X}_1 - \bar{X}_2$ suit, en vertu des propriétés de la distribution normale, une distribution normale de moyenne $\mu_1 - \mu_2$ et de variance $\sigma_{\bar{X}_1 - \bar{X}_2}^2$.

Le plus souvent, on ne connaît pas la variance de la différence entre les deux moyennes $\sigma_{\bar{x}_1 - \bar{x}_2}^2$. On est en effet amené à utiliser une estimation $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}^2$.

Pour tester l'hypothèse d'égalité des moyennes des deux populations, on utilise un critère de la forme :

$$\mathbf{T} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \quad (3.11)$$

Le calcul de $\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$, et donc la méthode de l'inférence statistique, se fait différemment selon que les deux échantillons sont associés par paires ou indépendants, et dans ce dernier cas selon que les variances des deux populations sont égales ou différentes.

3.3.2. Échantillons indépendants

3.3.2.1. Variances égales

3.3.2.1.1. Exemple introductif

Reprenons l'exemple des gains de poids de dindons du paragraphe 3.2. L'expérience avait pour objet, en effet, de comparer deux alimentations : une alimentation standard en y ajoutant un additif A et la même alimentation standard en y ajoutant un additif B.

Le tableau 3.2 reprend les données du tableau 3.1 en présentant également les valeurs obtenues pour un échantillon de 10 dindons nourris avec l'alimentation standard à laquelle on a ajouté l'additif B. Les 22 dindons sont élevés dans les mêmes autres conditions et ils avaient, au début de l'expérience, à peu près le même poids.

Ici, on souhaite savoir si l'on peut dire qu'il existe une relation entre la croissance des dindons et les deux alimentations (deux traitements/deux groupes). Si cette relation existe, cela veut dire qu'une alimentation entraîne une croissance plus importante que l'autre ou encore qu'il y a un effet alimentation.

Tableau 3.2. Gains de poids (en grammes) de deux échantillons de dindons recevant deux alimentations différentes (Alim + additif A et Alim + additif B).

Alim + additif A	2740	3110	2770	3250	3160	3020	2990	3170	3150	2990	2760	3150
Alim + additif B	2710	2910	2640	2900	2960	2990	2840	2920	2610	2790		

3.3.2.1.2. Variance de la distribution des différences

Considérons deux populations normales dont les moyennes sont μ_1 et μ_2 et les variances sont σ_1^2 et σ_2^2 . De chacune de ces populations, on suppose disposer d'un échantillon aléatoire et simple, les effectifs des deux échantillons étant n_1 et n_2 .

Lorsque les variances des deux populations sont égales ($\sigma_1^2 = \sigma_2^2$), on peut obtenir différentes estimations de la variance commune des deux populations. La plus utilisée de ces estimations est l'estimation de la variance combinée¹ ou conjointe qui est donnée par la relation :

¹ En anglais : *pooled variance*.

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}. \quad (3.12)$$

Cette variance est une moyenne pondérée des variances intra-groupes.

De ce qui précède, on peut obtenir une estimation de la variance de la distribution d'échantillonnage des différences entre les deux moyennes.

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}^2 = \hat{\sigma}_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right] = \frac{SCE_1 + SCE_2}{n_1 + n_2 - 2} \left[\frac{1}{n_1} + \frac{1}{n_2} \right]. \quad (3.13)$$

Dans ces conditions, la quantité **T** de l'expression (3.11) suit, lorsque l'hypothèse d'égalité des deux moyennes théoriques est vraie, une distribution **t** de Student à $n_1 + n_2 - 2$ degrés de liberté.

3.3.2.1.3. Intervalle de confiance de la différence des moyennes

L'intervalle de confiance de la différence de deux moyennes dans le cas d'échantillons indépendants peut être obtenu, lorsque les deux populations sont normales et de variances inconnues mais égales, à partir de l'expression suivante :

$$\left(\bar{x}_1 - \bar{x}_2 \right) \pm t_{1-\alpha/2} \sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2} \left[\frac{1}{n_1} + \frac{1}{n_2} \right]} \quad (3.14)$$

où la variable **t** de Student possède $n_1 + n_2 - 2$ degrés de liberté. La distribution **t** de Student peut être remplacée par la distribution normale réduite lorsque les effectifs sont assez grands.

Si les effectifs des deux échantillons sont égaux ($n_1 = n_2 = n$), l'expression (3.14) se simplifie pour donner :

$$\left(\bar{x}_1 - \bar{x}_2 \right) \pm t_{1-\alpha/2} \sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}} \quad (3.15)$$

3.3.2.1.4. Comparaison des moyennes

L'objectif est, d'une manière générale, de tester si la différence entre les moyennes de deux traitements est égale à une certaine constante. L'hypothèse nulle s'écrit :

$$H_0 : \mu_1 - \mu_2 = d_0 \quad (3.16)$$

où la constante **d₀** est le plus souvent égale à zéro. L'hypothèse nulle est testée contre l'une ou l'autre des trois hypothèses alternatives usuelles suivantes :

- a) $H_1 : \mu_1 - \mu_2 > d_0$
- b) $H_1' : \mu_1 - \mu_2 < d_0$
- c) $H_1'' : \mu_1 - \mu_2 \neq d_0$

Lorsque l'hypothèse nulle est vraie, le critère T de l'expression 3.11 suit une distribution t de Student pour des échantillons prélevés de populations normales de variances inconnues mais égales. Le test consiste à calculer la valeur observée de la variable t de Student :

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{\text{SCE}_1 + \text{SCE}_2}{n_1 + n_2 - 2} \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}} \quad (3.17)$$

et à rejeter l'hypothèse nulle, au niveau de signification α , selon l'hypothèse alternative utilisée :

- a) rejeter H_0 en faveur de H_1 si $t_{\text{obs}} > t_{1-\alpha}$
- b) rejeter H_0 en faveur de H_1' si $t_{\text{obs}} < -t_{1-\alpha}$
- c) rejeter H_0 en faveur de H_1'' si $|t_{\text{obs}}| > t_{1-\alpha/2}$

où la variable t de Student possède $n_1 + n_2 - 2$ degrés de liberté.

Ce test est appelé test t de Student ou de Student-Fisher¹. On constate que la décision du test dépend de l'hypothèse alternative. Si on opte pour un test unilatéral, il ne sera pas possible d'identifier de différence significative dans le sens opposé à celui qui a été prédit.

Exemple 3.4

Reprenez les données des deux échantillons de dindons (tableau 3.2) et testez si l'on peut admettre qu'il existe un effet alimentation significatif en prenant un risque de 5%.

Solution 3.4

a) Méthode : comparaison de deux moyennes théoriques, cas de deux échantillons indépendants

b) Conditions d'application

- Les deux échantillons de dindons sont aléatoires, simples et indépendants (aucun lien entre les dindons des deux échantillons) ;
- Les distributions des deux populations sont normales et possèdent la même variance σ^2 (nous verrons comment tester l'égalité des variances au chapitre 4).

c) Hypothèses

$$\begin{array}{ll} H_0 : \mu_1 - \mu_2 = 0 & \text{(absence d'effet alimentation)} \\ \text{contre } H_1'' : \mu_1 \neq \mu_2 & \text{(il y a effet alimentation)} \end{array}$$

d) Calcul

$$t_{\text{obs}} = \frac{3021.667 - 2827.00}{\sqrt{\frac{34716667 + 16241000}{12 + 10 - 2} \left[\frac{1}{12} + \frac{1}{10} \right]}} = 2.85$$

Puisque la valeur absolue de cette valeur observée dépasse la valeur critique à 20 degrés de liberté ($t_{0.975} = 2.086$), l'hypothèse nulle doit être rejetée. La p-value qui peut être obtenue par un logiciel statistique (**p=0.010**) est inférieure à $\alpha=0.05$.

Conclusion : En prenant un risque de 5%, on conclut à des différences significatives entre les deux gains moyens de poids. Le gain de poids avec l'alimentation standard en y ajoutant l'additif A est plus important.

¹ En anglais : two- sample t-test

Lorsque les effectifs des deux échantillons sont égaux, l'expression (3.17) se simplifie :

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}}} \quad (3.18)$$

3.3.2.2. Variances différentes

3.3.2.2.1. Généralités

Si les deux populations ont des variances différentes ($\sigma_1^2 \neq \sigma_2^2$), le critère T de l'expression 3.11, lorsque l'hypothèse nulle est vraie, n'est pas distribué comme une variable t de Student et les résultats du test t de Student ordinaire du paragraphe précédent peuvent être incorrects ou trompeurs.

L'estimation conjointe des variances n'est en effet pas valide et les degrés de liberté sont trop importants, ce qui a pour conséquences de donner des résultats « trop » significatifs. Le test t de Student est d'autant plus vulnérable à l'hétéroscédasticité (inégalité des variances) que les effectifs des deux échantillons sont différents et que l'effectif le plus petit est associé à la variance la plus grande.

Dans la pratique, si les deux effectifs sont égaux et si la variance d'une population n'est pas plus de trois fois la variance de l'autre population, le test t de Student est assez robuste et les résultats restent relativement fiables [Ott, 1988]. Si ces conditions ne sont pas rencontrées, on peut :

- 1) procéder à une transformation de variables, destinée à stabiliser les variances (cf. chapitre 8), et utiliser ensuite le test t de Student ;
- 2) utiliser l'approximation de Welch-Satterthwaite ou l'approximation de Cochran-Cox.

3.3.2.2.2. Approximations de Welch-Satterthwaite et de Cochran-Cox

L'approximation de Satterthwaite et celle de Cochran-Cox consistent d'abord à obtenir une estimation de l'écart-type des différences à partir des estimations séparées de variances :

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad (3.19)$$

et calculer ensuite une statistique ajustée t' en utilisant l'expression :

$$t'_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{SCE_1}{n_1(n_1-1)} + \frac{SCE_2}{n_2(n_2-1)}}} \quad (3.20)$$

La statistique t' n'est pas à proprement parlé distribuée selon une loi t de Student. C'est pourquoi on utilise l'approximation de Satterthwaite pour corriger les nombres de degrés de liberté ou celle de Cochran-Cox pour obtenir les valeurs critiques.

1) Selon l'approximation de Satterthwaite, on calcule la valeur observée t'_{obs} selon l'expression (3.20) et, pour un niveau de signification donné α , on est amené à :

- a) rejeter H_0 en faveur de H_1 si $t'_{obs} > t_{1-\alpha}$
- b) rejeter H_0 en faveur de H_1 si $t'_{obs} < -t_{1-\alpha}$
- c) rejeter H_0 en faveur de H_1 si $|t'_{obs}| > t_{1-\alpha/2}$

où la variable t de Student possède k degrés de liberté. Ce nombre k est calculé selon la formule donnée par Satterthwaite [1946] :

$$k = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}}. \quad (3.21)$$

avec $w_1 = \frac{SCE_1}{n_1(n_1 - 1)}$ et $w_2 = \frac{SCE_2}{n_2(n_2 - 1)}$.

Si le nombre k n'est pas entier, on l'arrondit à l'entier le plus proche. Toutefois, certains auteurs proposent la prudente décision de tronquer la partie décimale, tandis que certains logiciels statistiques réalisent les calculs sur la loi de Student dont le nombre de degrés de liberté n'est pas entier. Le degré de liberté k est inférieure ou égale à $n_1 + n_2 - 2$ et il d'autant plus petit que les deux variances sont différentes. Cette diminution du nombre de degrés de liberté entraîne une augmentation des valeurs critiques auxquelles on compare la valeur observée de la statistique du test, ce qui fait que le test utilisé est plus conservatif.

Le test de Satterthwaite a à peu près la même propriété de robustesse que le test t ordinaire lorsque les variances sont égales.

2) Selon l'approximation de Cochran-Cox [1957], on calcule la statistique t'_{obs} selon l'expression (3.20) et, pour un niveau de signification donné α , on est amené à :

- a) rejeter H_0 en faveur de H_1 si $t'_{obs} > t'_{1-\alpha}$
- b) rejeter H_0 en faveur de H_1 si $t'_{obs} < -t'_{1-\alpha}$
- c) rejeter H_0 en faveur de H_1 si $t'_{obs} < -t'_{1-\alpha/2}$ ou $t'_{obs} > t'_{1-\alpha/2}$

où $t'_{1-\alpha/2} = \frac{w_1 t_{1-\alpha/2, n_1-1} + w_2 t_{1-\alpha/2, n_2-1}}{w_1 + w_2}$ et $t'_{1-\alpha} = \frac{w_1 t_{1-\alpha, n_1-1} + w_2 t_{1-\alpha, n_2-1}}{w_1 + w_2}$.

L'approche utilisant l'approximation de Cochran-Cox tend, en général, à être plus conservatrice [Lee et Gurland, 1975].

Bien d'autres approximations permettant de réaliser le test d'égalité de deux moyennes lorsque les variances des deux populations sont différentes existent. On peut citer en particulier celle de Howe [1974].

Exemple 3.5

La longueur de l'aile droite d'une espèce d'insectes est mesurée sur deux échantillons dans deux régions. Les résultats ont permis d'obtenir les paramètres statistiques du tableau 3.3. Peut-on conclure à des différences significatives entre les longueurs des ailes dans les deux régions ($\alpha=5\%$).

Tableau 3.3. Paramètres statistiques obtenus pour deux échantillons d'insectes prélevés dans deux régions.

Échantillon	Effectif	Moyenne (mm)	Ecart-type estimé (mm)
Région 1	10	9.1845	0.0895
Région 2	15	9.2930	0.1970

Solution 3.5

a) Méthode : comparaison de deux moyennes, cas de deux échantillons indépendants

b) Conditions d'application

- les deux échantillons sont indépendants et sont prélevés de manière aléatoire et simple
- les deux populations sont normales

c) Hypothèses

$H_0 : \mu_1 = \mu_2$ (les longueurs moyennes sont les mêmes dans les deux régions)

contre $H_1 : \mu_1 \neq \mu_2$ (il y a différences entre les longueurs des deux régions)

d) Calcul

Puisque les deux variances ne sont pas égales (le test d'égalité de variances sera étudié au chapitre 4), on calcule la statistique t'_{obs} approximative :

$$t'_{obs} = \frac{9.1845 - 9.293}{\sqrt{\frac{0.0721686}{10(10-1)} + \frac{0.545886}{15(15-1)}}} = -1.86.$$

Selon l'approximation de Satterthwaite, cette valeur doit être comparée à la valeur de t à $k=20.9$ degrés de liberté, ou à $k=21$ degrés de liberté après l'arrondi.

Puisque la valeur absolue de la valeur observée est inférieure à la valeur critique $t_{0.975}=2.080$, l'hypothèse nulle ne peut être rejetée (si on calcule la probabilité de signification, on trouve $p\text{-value}=0.077$, qui est supérieure à la valeur nominale 0.05 , cette probabilité est donnée par les logiciels statistiques ou par certains tableurs comme Excel).

Conclusion : en prenant un risque de 5%, il n'y a pas suffisamment de preuve pour conclure à des différences significatives entre les longueurs d'ailles de l'insecte en question dans les deux régions.

Remarque : Dans les mêmes conditions, si on avait opté pour l'approximation de Cochran-Cox, on aurait trouvé :

$$t'_{0.975} = \frac{(0.00080103)(2.262) + (0.002587)(2.145)}{0.00080103 + 0.002587} = 2.17.$$

En tenant compte des règles de décisions, on aboutit à la même conclusion que précédemment.

3.3.3. Échantillons associés par paires

3.3.3.1. Généralités

Il s'agit ici d'un appariement des unités. Cet appariement peut être réalisé de deux façons différentes :

1) Soit que les unités sont regroupées en couples de telle sorte que les deux unités de chaque couple soient similaires. L'une des deux unités du couple reçoit alors un traitement et l'autre unité reçoit l'autre traitement. Dans ce cas, on peut citer les deux exemples suivants :

- il peut s'agir de la comparaison de deux traitements antalgiques. Un échantillon de n couples de malades peut être constitué de telle sorte que les deux malades d'un même couple soient du même âge, du même sexe et du même état général de santé. De cette façon, s'il y a différence entre les deux moyennes, celle-ci sera due aux différences entre les deux traitements et non aux différences entre les malades ;
- il peut s'agir aussi de la comparaison de deux enseignants du point de vue sévérité dans la notation des copies d'examens. On peut constituer un échantillon de n couples de copies de telle sorte que les deux copies d'un même couple appartiennent à deux étudiants jugés de niveaux similaires.

2) Soit qu'il s'agit du même individu qui subit les deux traitements (auto-appariement). Dans ce cas, on peut citer les deux exemples suivants :

- il peut s'agir de la comparaison des poids moyens avant et après un régime amaigrissant chez des hommes. Dans ce cas, on peut utiliser un échantillon de n individus, en mesurant le poids de chacun d'eux avant puis après régime ;
- il peut s'agir aussi de la comparaison des effets de deux traitements (la benzédrine et un placebo) sur la fréquence cardiaque des chiens. Pour chacun des chiens de l'expérience, on peut administrer l'un des traitements et enregistrer le nombre de battements par minutes après deux heures. Après deux semaines, on commute les régimes, les chiens sous la benzédrine reçoivent le placebo et inversement.

Dans les deux cas de figure, on remarque que chaque unité de l'une des populations est mise en relation avec une unité de l'autre population. C'est pourquoi on parle d'échantillons associés par paires ou par couples ou encore d'échantillons appariés.

Pour chaque couple, la différence entre les mesures faites sur les deux observations est une estimation de la différence entre les effets des deux traitements. Pour tenir compte des variabilités des unités, on doit disposer de mesures sur plusieurs couples.

L'objectif de l'appariement est d'augmenter la précision de la comparaison entre les deux traitements, ce gain de précision se manifeste lorsque les deux échantillons sont fortement liés (coefficient de corrélation positif). Le critère d'appariement doit être lié aux performances de l'individu quant aux réactions des traitements à comparer. Si l'on souhaite par exemple comparer deux méthodes d'enseignement à des étudiants, il y a intérêt à appairer les étudiants en fonction de leurs niveaux.

Pour comparer les effets moyens des deux traitements, supposons que les n couples d'individus sont tirés de manière aléatoire et simple et que les différences D_i ($i=1, \dots, n$) entre les valeurs des couples sont distribuées autour de la moyenne μ_D qui représente la différence moyenne des effets des deux traitements.

Lorsque l'hypothèse nulle d'égalité des deux moyennes est vraie et que les différences D_i sont indépendantes et distribuées selon une loi normale de moyenne μ_D et de variance σ_D^2 , la différence moyenne \bar{D} est alors distribuée selon une loi normale de moyenne μ_D et de variance σ_D^2 / n . Il en découle que la quantité :

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

où la variable aléatoire S_D est l'écart-type des différences, suit une distribution t de Student à $n-1$ degrés de liberté.

La quantité σ_D^2 , qui désigne la variance théorique des différences entre les valeurs des couples, n'est généralement pas connue, elle est estimée à partir des résultats de l'expérience.

Les données observées des n couples ainsi que les différences observées d_i peuvent être consignées dans un tableau semblable au tableau 3.4. On peut ainsi obtenir une estimation de σ_D^2 par la relation :

$$\hat{\sigma}_D^2 = \frac{SCE_d}{n-1} = \frac{\sum_{i=1}^n d_i^2 - \left[\sum_{i=1}^n d_i \right]^2 / n}{n-1}.$$

Tableau 3.4. Forme de présentation des données de deux échantillons associés par paires.

Couple	Observation 1 du couple	Observation 2 du couple	Différence (d_i)
1	X_{11}	X_{21}	$X_{11}-X_{21}$
2	X_{12}	X_{22}	$X_{12}-X_{22}$
3	X_{13}	X_{23}	$X_{13}-X_{23}$
.	.	.	.
.	.	.	.
n	X_{1n}	X_{2n}	$X_{1n}-X_{2n}$
Moyenne	\bar{x}_1	\bar{x}_2	$\bar{d} = \bar{x}_1 - \bar{x}_2$

3.3.3.2. Intervalle de confiance de la différence de deux moyennes

L'intervalle de confiance à $(1-\alpha)$ de la différence moyenne μ_D des effets de deux traitements est, dans le cas d'échantillons associés par paires, donné par la relation :

$$\bar{d} \pm t_{1-\alpha/2} \hat{\sigma}_{\bar{D}}$$

ou encore :

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2} \sqrt{\frac{SCE_d}{n(n-1)}}$$

où, comme nous l'avons vu, la variable t de Student possède $n-1$ degrés de liberté.

3.3.3.3. Comparaison des moyennes

L'hypothèse nulle dans le cas du test de comparaison de deux moyennes lorsque les échantillons sont appariés¹ s'écrit :

$$H_0 : \mu_D = d_0$$

où $\mu_D = \mu_1 - \mu_2$ et d_0 est une constante qui est le plus souvent égale à zéro.

Cette l'hypothèse nulle est testée contre l'une ou l'autre des trois hypothèses alternatives usuelles suivantes :

- a) $H_1 : \mu_D > d_0$
- b) $H_1' : \mu_D < d_0$
- c) $H_1'' : \mu_D \neq d_0$.

D'après les relations des paragraphes précédents, le test statistique consiste à calculer la valeur observée de la variable t de Student :

$$t_{\text{obs}} = \frac{\bar{d} - d_0}{\hat{\sigma}_D / \sqrt{n}} = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{SCE_d}{n(n-1)}}}$$

et à prendre, pour un niveau de signification α et selon l'hypothèse alternative, l'une des décisions suivantes :

- a) sous $H_1 : \mu_1 > \mu_2$, on rejette H_0 si $t_{\text{obs}} > t_{1-\alpha}$
- b) sous $H_1' : \mu_1 < \mu_2$, on rejette H_0 si $t_{\text{obs}} < -t_{1-\alpha}$
- c) sous $H_1'' : \mu_1 \neq \mu_2$, on rejette H_0 si $|t_{\text{obs}}| > t_{1-\alpha/2}$

où la variable t de Student possède $n-1$ degrés de liberté.

¹ En anglais : *paired t-test*.

Exemple 3.7

On souhaite tester si deux méthodes A et B de dosage d'une protéine sérique de bovins donnent des résultats différents. Pour ce faire, on a prélevé du sang sur 10 vaches adultes saines. Pour chaque prélèvement, la teneur en cette protéine est déterminée par les deux méthodes. En fonction des résultats obtenus (tableau 3.5), quelle conclusion peut-on faire ($\alpha=0.05$) ?

Tableau 3.5. Taux d'une protéine sérique obtenus par deux méthodes sur 10 vaches.

N° de la vache	1	2	3	4	5	6	7	8	9	10
Méthode A	4.02	3.95	4.11	4.24	3.41	4.87	3.87	2.89	4.04	4.17
Méthode B	3.80	4.04	3.78	4.01	3.88	4.38	3.46	3.02	3.72	3.95

Solution 3.7

a) **Méthode** : comparaison de deux moyennes, échantillons associés par paires

b) **Conditions d'application** :

- l'échantillon des dix bovins est prélevé de manière aléatoire et simple dans la population,
- la population des différences D_i est distribuée selon une loi normale.

c) **Hypothèses**

$$H_0 : \mu_1 - \mu_2 = 0 \quad (\text{les deux méthodes donnent en moyenne les mêmes teneurs})$$

contre $H_1 : \mu_1 - \mu_2 \neq 0$ (les deux méthodes ne donnent pas en moyenne les mêmes teneurs)

d) **Calcul**

La valeur observée de t de Student est :

$$t_{\text{obs}} = \frac{0.1530}{\sqrt{\frac{0.7810}{10(10-1)}}} = 1.64$$

Puisque la valeur absolue de cette valeur observée de t est inférieure à la valeur critique $t_{0,975} = 2.262$, l'hypothèse nulle ne peut être rejetée (on peut vérifier par un logiciel statistique que $p\text{-value}$ est supérieure à α ($p\text{-valeur} = 0.135$ et $\alpha = 0.05$)).

Conclusion : En prenant un risque de 5%, il n'y a pas suffisamment de preuve pour conclure à des différences significatives entre les résultats obtenus par les deux méthodes.

3.4. CONCLUSION

Dans le cas où l'inférence concerne une seule moyenne, nous avons vu d'une part comment obtenir une estimation non biaisée de la moyenne de la population et l'entourer d'un intervalle de confiance en fixant un risque de première espèce et, d'autre part, comment tester l'égalité de la moyenne de la population, bien sûr inconnue, à une valeur de référence donnée.

Dans le cas où l'inférence concerne deux moyennes, nous avons vu qu'on s'intéresse généralement à tester leur égalité. Nous avons souligné à ce propos qu'il est primordial de distinguer entre la situation où les échantillons sont indépendants de la situation où ceux-ci sont associés par paires. Une confusion entre les deux tests est une erreur.

Nous avons également discuté les conditions d'application du test t de Student. Le non respect de ces conditions peut, dans certains cas, entraîner des résultats incorrects ou trompeurs.

Dans le cas d'échantillons indépendants, nous avons insisté sur l'examen des deux variances. Si les effectifs des deux échantillons sont égaux, le test **t** ordinaire reste applicable jusqu'à même si la variance d'une population est trois fois la variance de l'autre population, surtout si l'effectif est grand. Par contre, si les effectifs sont différents, plus particulièrement lorsqu'ils sont petits et la plus grande variance est associée au plus petit échantillon, un test d'égalité des variances s'impose. Nous avons, à cette occasion, présenté les approximations de Satterthwaite et de Cochran-Cox comme alternatives.

Lorsque l'hypothèse de la normalité n'est pas respectée ou lorsqu'il y a présence de valeurs anormales, le test **t** ordinaire risque de ne pas être le test correct. Le test **t** est en effet assez robuste, mais si on s'éloigne trop de la condition de la normalité, il est préférable d'utiliser des tests non paramétriques. Deux de ces tests ont été présentés, le test de Mann-Whitney pour les échantillons indépendants et le test de Wilcoxon pour les échantillons associés par paires (cette partie a été supprimée suite à la réduction du volume horaire).

Si les conditions de l'égalité des variances et de la normalité ne sont pas vérifiées en même temps, nous pensons qu'il est préférable de commencer par essayer des transformations de variables pour essayer de stabiliser les variances et de s'approcher de la condition de la normalité des populations. Nous reviendrons sur ce sujet en détail au chapitre 8.

CHAPITRE 4

INFÉRENCES CONCERNANT LES VARIANCES

4.1. INTRODUCTION

Dans le chapitre précédent, nous avons vu comment réaliser le test de comparaison de deux moyennes théoriques à partir d'échantillons indépendants lorsque les variances ne sont pas connues et nous avons souligné l'intérêt de tester au préalable l'égalité des deux variances théoriques surtout lorsque le test sur les moyennes ne se révèle pas robuste.

Ce chapitre est d'autant plus important que nous reviendrons aux chapitres 5 et 6 sur les tests de comparaison de plusieurs moyennes et nous serons amené de nouveau à poser l'égalité des variances des populations comme condition d'application.

Il est vrai aussi que l'expérimentateur est souvent intéressé par les moyennes, mais il peut aussi être, dans plusieurs cas, préoccupé par la variabilité du matériel étudié. Il peut ainsi être intéressé, par exemple, par la comparaison des variances de la taille des pièces produites par deux ou plusieurs machines et, d'une manière générale, savoir si la variabilité d'un paramètre présente la même dispersion dans plusieurs populations différentes.

Plusieurs méthodes relatives aux variances seront brièvement passées en revue dans ce chapitre. Dans un premier temps, nous présenterons un tableau de données qui nous servira à illustrer les calculs des différentes méthodes (paragraphe 4.2). Nous verrons ensuite comment obtenir l'intervalle de confiance d'une variance (paragraphe 4.3). Nous aborderons enfin les principaux autres problèmes de tests d'hypothèses, à savoir le test de conformité (paragraphe 4.4), le test d'égalité de deux variances (paragraphe 4.5) et le test d'égalité de plusieurs variances (paragraphe 4.6).

4.2. EXEMPLE INTRODUCTIF

Le tableau 4.1 présente les poids de fruits de trois variétés de dattes marocaines : *Ahardane*, *Bouijjou* et *Jihel*. Chaque valeur est une moyenne calculée à partir de vingt fruits. Pour chaque variété, les mesures ont été obtenues pour huit prélèvements [Harrak *et al.*, 2003].

Ces données nous serviront d'exemple pour illustrer les méthodes qui seront exposées le long de ce chapitre.

Tableau 4.1. Poids moyens, en grammes, des fruits de trois variétés de dattes obtenues pour huit prélèvements.

Variétés de dattes		
Ahardane	Bouijjou	Jihel
8.07	9.47	8.10
9.10	11.14	7.36
7.68	9.34	8.57
6.94	10.56	10.12
7.78	8.52	8.91
9.10	9.69	8.55
8.21	8.71	7.57
6.37	9.88	9.79

4.3. INTERVALLE DE CONFIANCE D'UNE VARIANCE

Rappelons d'abord qu'une estimation ponctuelle sans biais de la variance σ^2 de la population est donnée par :

$$\hat{\sigma}^2 = \frac{ns^2}{n-1} = \frac{SCE}{n-1},$$

où s^2 est la variance observée d'un échantillon aléatoire et simple de n observations prélevé dans cette population et **SCE** est la somme des carrés des écarts.

De même, sous l'hypothèse de la normalité, on démontre que la quantité :

$$\frac{nS^2}{\sigma^2}$$

suit une distribution χ^2 à $n-1$ degrés de liberté où l'on considère que $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$.

Cette propriété nous permet de déterminer les limites S_1^2 et S_2^2 de l'intervalle qui encadrent la vraie valeur de la variance σ^2 de la population avec une forte probabilité $(1-\alpha)$.

On a :

$$P(S_1^2 < \sigma^2 < S_2^2) = 1 - \alpha.$$

Pour obtenir un intervalle de confiance symétrique en probabilité, on peut écrire :

$$P(\sigma^2 < S_1^2) = \alpha/2 \quad \text{et} \quad P(\sigma^2 > S_2^2) = \alpha/2.$$

En se référant à la figure 4.1, il ressort que :

$$P\left(\frac{nS^2}{\sigma^2} < \chi_{\alpha/2}^2\right) = \alpha/2 \quad \text{et} \quad P\left(\frac{nS^2}{\sigma^2} > \chi_{1-\alpha/2}^2\right) = \alpha/2$$

ou

$$P(\chi_{\alpha/2}^2 < \frac{nS^2}{\sigma^2} < \chi_{1-\alpha/2}^2) = 1 - \alpha.$$

Les quantités $\chi_{\alpha/2}^2$ et $\chi_{1-\alpha/2}^2$ sont respectivement les $\alpha/2$ et $1-\alpha/2$ quantiles de la loi khi-deux à $(n-1)$ degrés de liberté.

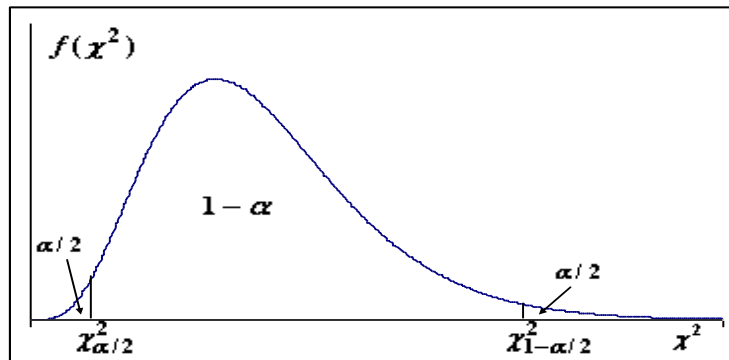


Figure 4.1. Valeurs critiques de la distribution χ^2 .

De manière équivalente, on peut écrire :

$$P\left(\frac{nS^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{nS^2}{\chi^2_{\alpha/2}}\right) = 1 - \alpha,$$

et on peut en déduire que :

$$S_1^2 = \frac{nS^2}{\chi^2_{1-\alpha/2}} \quad \text{et} \quad S_2^2 = \frac{nS^2}{\chi^2_{\alpha/2}}.$$

Il ressort que, pour une valeur donnée s^2 de S^2 d'un échantillon aléatoire et simple, on obtient l'intervalle de confiance de la variance σ^2 de la population au niveau de confiance $(1-\alpha)$ par la relation suivante :

$$\frac{ns^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{ns^2}{\chi^2_{\alpha/2}} \quad \text{ou encore par :} \quad \boxed{\frac{SCE}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{SCE}{\chi^2_{\alpha/2}}},$$

où la variable χ^2 possède $k=n-1$ degrés de liberté.

Ces limites ne sont pas symétriques par rapport à la valeur estimée $\hat{\sigma}^2$ et elles ne peuvent être calculées que sous l'hypothèse de la normalité de la distribution sous-jacente. Si celle-ci reste proche de la normale, la procédure est suffisamment robuste et donne généralement de bons résultats.

L'intervalle de confiance de l'écart-type au niveau de confiance $1-\alpha$ peut être obtenu en calculant la racine carrée des limites de confiance de la variance :

$$\sqrt{\frac{SCE}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{SCE}{\chi^2_{\alpha/2}}}.$$

Exercice 4.1

Reprenez les données du tableau 4.1 et cherchez l'intervalle de confiance de l'écart-type des poids des dattes pour la variété *Ahardane* en utilisant un degré de confiance de 95% ?

Solution 4.1

a) Méthode : intervalle de confiance d'une variance

b) Conditions d'application :

- le poids des dattes pour la variété *Ahardane* suit une loi normale
- l'échantillon prélevé est aléatoire et simple

b) Calcul :

- l'écart-type σ de la population est estimé par : $\hat{\sigma} = \sqrt{\frac{SCE}{n-1}} = \sqrt{\frac{6.33}{7}} = 0.95 \text{ g.}$
- l'intervalle de confiance de l'écart-type est obtenu par :

$$\sqrt{\frac{6.33}{16.01}} < \sigma < \sqrt{\frac{6.33}{1.69}}.$$

c) Conclusion : les limites **0.63** et **1.94 grammes** encadrent la vraie valeur de l'écart-type σ avec un degré de confiance de 95%.

4.4. TEST DE CONFORMITÉ D'UNE VARIANCE

L'objet est de tester si la variance théorique σ^2 de la population est égale à une valeur spécifiée σ_0^2 . L'hypothèse nulle s'écrit :

$$H_0 : \sigma_1^2 = \sigma_0^2$$

contre l'une ou l'autre des trois hypothèses alternatives suivantes :

$$(a) H_1 : \sigma^2 > \sigma_0^2$$

$$(b) H_1' : \sigma^2 < \sigma_0^2$$

$$(c) H_1'' : \sigma^2 \neq \sigma_0^2$$

Sachant que, sous l'hypothèse de la normalité, la variable $Z = nS^2 / \sigma^2$ suit une distribution χ^2 à $n-1$ degrés de liberté, on peut raisonner de la manière suivante :

1) Sous l'hypothèse alternative (a), la variance S^2 a tendance à prendre des valeurs plus élevées que lorsque H_0 est vraie. Autrement dit, les valeurs élevées de Z poussent à rejeter H_0 . On construit donc un test unilatéral dont le domaine de rejet est situé à droite de la distribution χ^2 ;

2) Sous l'hypothèse alternative (b), la variance S^2 a tendance à prendre des valeurs plus faibles que lorsque H_0 est vraie. Autrement dit, les valeurs faibles de Z conduisent au rejet de H_0 . On construit donc un test unilatéral dont le domaine de rejet est situé à gauche de la distribution χ^2 ;

3) Sous l'hypothèse alternative (c), on construit un test bilatéral dont les domaines de rejet sont situés à gauche et à droite de la distribution χ^2 , car, dans ce cas la variance S^2 a tendance à prendre des valeurs faibles ou élevées.

En utilisant un échantillon aléatoire et simple de données et en désignant par s^2 la valeur observée de S^2 , on calcule la valeur observée de χ^2 :

$$\chi_{\text{obs}}^2 = \frac{ns^2}{\sigma_0^2} = \frac{\text{SCE}}{\sigma_0^2}$$

et on prend, pour un niveau de signification α , les décisions suivantes :

$$(a) \text{ sous } H_1 : \sigma^2 > \sigma_0^2, \text{ on rejette } H_0 \text{ si } \chi_{\text{obs}}^2 > \chi_{1-\alpha}^2$$

$$(b) \text{ sous } H_1' : \sigma^2 < \sigma_0^2, \text{ on rejette } H_0 \text{ si } \chi_{\text{obs}}^2 < \chi_{\alpha}^2$$

$$(c) \text{ sous } H_1'' : \sigma^2 \neq \sigma_0^2, \text{ on rejette } H_0 \text{ si } \chi_{\text{obs}}^2 < \chi_{\alpha/2}^2 \text{ ou } \chi_{\text{obs}}^2 > \chi_{1-\alpha/2}^2$$

où la variable χ^2 possède $k=n-1$ degrés de liberté.

Remarques :

- Dans le cas du test bilatéral (c), on peut remarquer que l'hypothèse nulle est rejetée lorsque σ_0^2 n'appartient pas à l'intervalle de confiance de σ^2 ;
- Quand le nombre de degrés de liberté est assez grand ($k > 30$), on peut utiliser un test basé sur la normale réduite. Dans le cas du test bilatéral, par exemple, on calcule :

$$u_{obs} = \left| \sqrt{2SCE / \sigma_0^2} - \sqrt{2n - 3} \right|$$

et on rejette l'hypothèse nulle lorsque $u_{obs} \geq u_{1-\alpha/2}$.

Exercice 4.2

Un marchand de dattes de la variété Jihel assure que la variabilité des poids de ces dattes est au plus de 0.5 g². A partir des données de l'échantillon du tableau 4.1, dites si l'on peut accepter ce qu'avance le marchand au niveau de probabilité $\alpha=0.05$?

Solution 4.2

a) **Méthode** : test de conformité d'une variance (test khi-deux)

b) **Conditions d'application** :

- la distribution des poids des dattes de la variété Jihel est normale
- l'échantillon est prélevé de manière aléatoire et simple

c) **Hypothèse nulle** :

$$H_0 : \sigma^2 = 0.5$$

contre

$$H_1 : \sigma^2 > 0.5 \quad (\text{la variabilité garantie est dépassée})$$

d) **Calcul** : La valeur observée de χ^2 :

$$\chi_{obs}^2 = \frac{6.671}{0.5} = 13.34$$

est inférieure à la valeur critique $\chi_{1-\alpha}^2 = 14.07$ à $k=7$ degrés de liberté. La p-value étant de **0.064**.

On ne peut donc pas rejeter **H₀**.

e) **Conclusion** : il n'y a suffisamment pas de preuves pour rejeter ce qu'annonce le marchand. Cependant, il serait plus raisonnable d'augmenter l'effectif de l'échantillon et de refaire le test, surtout que la p-value est comprise entre 0.05 et 0.10.

4.5. TEST D'ÉGALITÉ DE DEUX VARIANCES

Le but est de tester si les variances σ_1^2 et σ_2^2 de deux populations ayant des distributions normales sont identiques. On suppose que l'on dispose des valeurs de deux échantillons d'effectifs n_1 et n_2 prélevés de manière aléatoire et simple et indépendamment l'un de l'autre dans les deux populations.

L'hypothèse nulle que l'on cherche à tester s'écrit :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

contre l'une ou l'autre des hypothèses alternatives suivantes :

- (a) $H_1 : \sigma_1^2 > \sigma_2^2$
- (b) $H_1' : \sigma_1^2 < \sigma_2^2$.
- (c) $H_1'' : \sigma_1^2 \neq \sigma_2^2$

Dans les conditions précitées (normalité et indépendance), on peut montrer que la quantité :

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

suit une distribution F de Snedecor à $n_1 - 1$ et $n_2 - 1$ degrés de liberté. Les indices 1 et 2 se réfèrent respectivement à la première et à la deuxième population. Il s'agit en effet du rapport de deux variables khi-deux (χ^2), chacune étant divisée par son nombre de degrés de liberté.

Lorsque l'hypothèse nulle est vraie, c'est-à-dire $\sigma_1^2 = \sigma_2^2$, l'expression précédente se simplifie pour donner la valeur observée de la variable **F** de Snedecor :

$$F'_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}.$$

D'après cette relation, il ressort que :

1) Sous l'hypothèse alternative (a), les valeurs relatives à la population **1** auront tendance à être plus dispersées que sous **H₀**. Autrement dit, on est amené à rejeter **H₀** lorsque F'_{obs} a tendance à prendre des valeurs élevées. Le domaine de rejet est donc situé à droite de la distribution **F**, c'est-à-dire lorsque $F'_{obs} \geq F_{1-\alpha}$ où la variable **F** possède $n_1 - 1$ et $n_2 - 1$ degrés de liberté ;

2) Sous l'hypothèse alternative (b), les valeurs relatives à la population **2** auront tendance à être plus dispersées que sous **H₀**. Autrement dit, on est amené à rejeter **H₀** lorsque F'_{obs} a tendance à prendre des valeurs faibles. Le domaine de rejet est donc situé à gauche de la distribution **F**, c'est-à-dire lorsque $F'_{obs} \leq F_{\alpha}$ où la variable **F** possède toujours $n_1 - 1$ et $n_2 - 1$ degrés de liberté ;

3) Sous l'hypothèse alternative (c), il ressort que l'on rejette l'hypothèse nulle lorsque la valeur de F'_{obs} est assez différente de l'unité (de 1), c'est-à-dire lorsque $F'_{obs} \leq F_{\alpha/2}$ ou $F'_{obs} \geq F_{1-\alpha/2}$ où la variable F possède $n_1 - 1$ et $n_2 - 1$ degrés de liberté (figure 4.2).

Dans le cas du test bilatéral, la règle de décision peut être simplifiée en calculant la quantité observée de F en mettant au numérateur la plus grande des deux variances estimées :

$$F_{obs} = \frac{\hat{\sigma}_{\max}^2}{\hat{\sigma}_{\min}^2}$$

et en rejetant H_0 lorsque $F_{obs} \geq F_{1-\alpha/2}$ où la variable F possède k_{\max} et k_{\min} degrés de liberté, où k_{\max} est le nombre de degrés de liberté qui correspond à $\hat{\sigma}_{\max}^2$ et k_{\min} est le nombre de degrés de liberté qui correspond à $\hat{\sigma}_{\min}^2$.

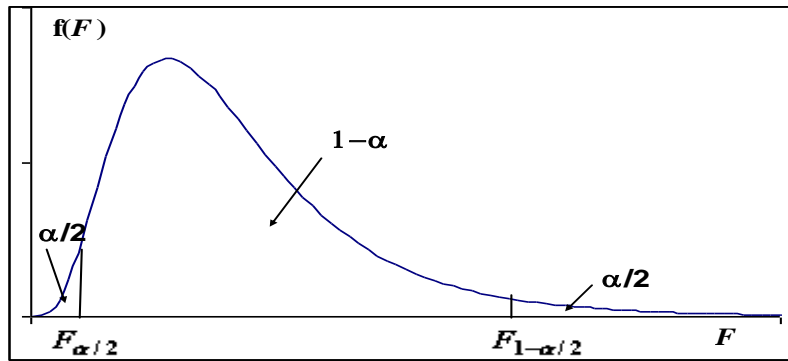


Figure 4.2. Valeurs critiques de la distribution F dans le cas du test bilatéral.

On retient que le test de Fisher, pour un niveau de signification α , permet de prendre les décisions suivantes :

- (a) sous $H_1 : \sigma_1^2 > \sigma_2^2$, on calcule $F'_{obs} = \hat{\sigma}_1^2 / \hat{\sigma}_2^2$ et on rejette H_0 si $F'_{obs} \geq F_{1-\alpha}$
où la variable F possède $k_1 = n_1 - 1$ et $k_2 = n_2 - 1$ degrés de liberté
- (b) sous $H'_1 : \sigma_1^2 < \sigma_2^2$, on calcule $F''_{obs} = \hat{\sigma}_2^2 / \hat{\sigma}_1^2$ et on rejette H_0 si $F''_{obs} \geq F_{1-\alpha}$
où la variable F possède $k_1 = n_2 - 1$ et $k_2 = n_1 - 1$ degrés de liberté
- (c) sous $H''_1 : \sigma_1^2 \neq \sigma_2^2$, on calcule $F_{obs} = \hat{\sigma}_{\max}^2 / \hat{\sigma}_{\min}^2$ et rejette H_0 si $F_{obs} \geq F_{1-\alpha/2}$
où la variable F possède k_{\max} et k_{\min} degrés de liberté, k_{\max} est le nombre de degrés de liberté qui correspond à $\hat{\sigma}_{\max}^2$ et k_{\min} est le nombre de degrés de liberté qui correspond à $\hat{\sigma}_{\min}^2$.

Exercice 4.3

Reprenez les données du tableau 4.1 et comparez au niveau de probabilité $\alpha=5\%$, les variances des poids des dattes des variétés Ahardane et Bouijjou ?

Solution 4.3

a) **Méthode** : comparaison de deux variances par le test de Fischer

b) **Conditions d'application** :

- les deux distributions des poids des dattes sont normales
- les deux échantillons prélevés sont aléatoires, simples et indépendants

c) **Hypothèse nulle** :

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ contre l'alternative bilatérale } H_1'' : \sigma_1^2 \neq \sigma_2^2$$

d) **Calcul** : La valeur observée de F :

$$F_{obs} = \frac{\hat{\sigma}_{\max}^2}{\hat{\sigma}_{\min}^2} = \frac{0.904}{0.770} \text{ ou } 1.174$$

est inférieure à la valeur critique $F_{1-\alpha/2} = 4.99$ à $k_1=7$ et $k_2=7$ degrés de liberté. On ne peut donc pas rejeter H_0 . La probabilité de signification est de **0.838**.

e) **Conclusion** : il n'y a suffisamment pas de preuve pour rejeter l'hypothèse d'égalité des variances des poids des dattes des variétés Ahardane et Bouijjou.

Remarques :

1) Si les deux échantillons ne sont pas indépendants, on utilise des tests qui tiennent compte de la corrélation, tels que le test de Pitman [1939]. On peut citer l'exemple de dosage de la β -globuline chez des bovins. Pour chacune des n vaches de l'expérience, un échantillon de sang est dosé par deux méthodes. On souhaite savoir si les valeurs obtenues par une méthode sont plus dispersées que celles qu'on obtient par l'autre méthode.

2) On peut s'intéresser aussi à l'intervalle de confiance du rapport de deux variances σ_1^2 / σ_2^2 . Lorsque les distributions des populations sont normales. On a :

$$P\left(a_1 < \frac{n_1 S_1^2 / ((n_1 - 1) \sigma_1^2)}{n_2 S_2^2 / ((n_2 - 1) \sigma_2^2)} < a_2\right) = 1 - \alpha$$

où $a_1 = F_{\alpha/2}$ et $a_2 = F_{1-\alpha/2}$ sont les valeurs d'une variable F possédant $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

En disposant des valeurs de deux échantillons, on obtient l'intervalle au niveau de confiance $1-\alpha$ suivant :

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \frac{1}{F_{1-\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} F'_{1-\alpha/2}$$

Les quantités $F_{1-\alpha/2}$ et $F'_{1-\alpha/2}$ sont les valeurs de la variable F de Snedecor à $n_1 - 1$ et $n_2 - 1$ degrés de liberté pour la première et $n_2 - 1$ et $n_1 - 1$ degrés de liberté pour la seconde.

Exercice 4.4

Supposons que l'on ne dispose pas de la première observation de la variété de dattes Ahardane (tableau 4.1). Cherchez l'intervalle de confiance du rapport des variances des poids des fruits des variétés Ahardane et Bouijjou au niveau de probabilité $\alpha=5\%$?

Solution 4.4

a) Méthode : intervalle de confiance du rapport de deux variances

b) Conditions d'application :

- les deux distributions des poids des dattes sont normales
- les deux échantillons prélevés sont aléatoires, simples et indépendants

c) Calcul :

$$\frac{1.0499}{0.7700} \left(\frac{1}{5.119} \right) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1.0499}{0.7700} (5.695)$$

d) Conclusion : l'intervalle [0.27, 7.77] encadre le vrai rapport des variances σ_1^2 / σ_2^2 avec un degré de confiance de 95%.

Le test de Fisher de la comparaison de deux variances n'est pas robuste, il est sensible à la non normalité des populations.

4.6. TESTS D'ÉGALITÉ DE PLUSIEURS VARIANCES

4.6.1. Tests disponibles

Plusieurs procédures ont été développées pour tester l'hypothèse d'égalité de plusieurs variances. Nous présentons dans ce paragraphe ce que nous estimons les plus connues d'entre elles, à savoir les procédures de Bartlett (1937), de Hartley (1940,1950), de Levene (1960), de Brown-forsythe (1974) et d'O'Brien (1979).

L'hypothèse nulle à tester concerne l'égalité des variances de **p** populations :

$$\begin{array}{ll} H_0 : & \sigma_1^2 = \sigma_2^2 = \dots = \sigma_p^2 \\ \text{contre} & H_1 : \quad \sigma_k^2 \neq \sigma_l^2 \text{ pour au moins un couple } (k, l). \end{array}$$

L'hypothèse **H₁** signifie que deux au moins des variances sont différentes. Dans tous les cas, nous supposons que l'on dispose pour toute population *i* d'un échantillon aléatoire et simple d'effectif **n_i** et que les **p** échantillons sont indépendants.

4.6.2. Test de Bartlett

Ce test, proposé par Bartlett (1937), est le plus communément utilisé pour tester l'égalité de plusieurs variances.

Soit $\hat{\sigma}_i^2$ la variance estimée de la population i . Si l'hypothèse nulle est vraie, la quantité :

$$\hat{\sigma}^2 = \frac{SCE}{n. - p}$$

où $n. = \sum_{i=1}^p n_i$ et $SCE = \sum_{i=1}^p SCE_i$ désignent l'effectif total et la somme des carrés des écarts globale, constitue une estimation non biaisée de la variance commune σ^2 . Le test consiste à calculer la quantité :

$$\chi_{obs}^2 = \frac{(n. - p) \ln \hat{\sigma}^2 - \sum_{i=1}^p \left((n_i - 1) \ln \hat{\sigma}_i^2 \right)}{1 + \frac{1}{3(p-1)} \left(\sum_{i=1}^p \frac{1}{n_i - 1} - \frac{1}{n. - p} \right)}$$

et à rejeter l'hypothèse nulle au niveau de signification α lorsque $\chi_{obs}^2 \geq \chi_{1-\alpha}^2$ où la distribution χ^2 possède $p-1$ degrés de liberté.

Lorsque les effectifs des échantillons sont égaux ($n_1 = n_2 = \dots = n_p = n$), on peut obtenir la valeur observée de la distribution χ^2 par la relation simplifiée suivante :

$$\chi_{obs}^2 = \frac{(n-1) \left(p \ln \frac{SCE}{p} - \sum_{i=1}^p \ln SCE_i \right)}{1 + \frac{p+1}{3p(n-1)}}$$

Lorsque la condition de la normalité des populations est vérifiée, le test de Bartlett contrôle le risque de première espèce et il est puissant. Il est très sensible à la non-normalité des populations (Box, 1953, Zar, 1999), et il ne peut donc être recommandé lorsque cette condition n'est pas vérifiée. Des études ont montré que ce test est à éviter lorsque les effectifs des échantillons sont petits (inférieurs à 4 ou 5) et lorsque le nombre de populations p est élevé par rapport aux effectifs des échantillons.

Exercice 4.5

Reprenez les données du tableau 4.1 et comparez, au niveau de probabilité $\alpha=5\%$, les variances des poids des fruits des trois variétés de dattes par la méthode de Bartlett ?

Solution 4.5

a) **Méthode** : comparaison de plusieurs variances par la procédure de Bartlett

b) **Conditions d'application** :

- les trois distributions des poids des dattes sont normales
- les trois échantillons prélevés sont aléatoires, simples et indépendants
- les effectifs des échantillons ne sont pas très limitants ($n > 4$) et le nombre de populations n'est pas élevé

c) **Hypothèse nulle** :

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

H_1 : deux au moins des variances sont différentes

d) Calcul : La valeur observée de χ^2 :

$$\chi_{obs}^2 = \frac{(8-1) \left[(3) \ln \frac{18.391}{3} - 5.4276 \right]}{1 + (4)/63} = 0.080$$

est inférieure à la valeur critique $\chi_{0.95}^2 = 5.99$ à $k=2$ degrés de liberté. On ne peut donc pas rejeter H_0 . La p-value est de **0.961**.

e) Conclusion : il n'y a suffisamment pas de preuve pour considérer les variances des poids des fruits des trois variétés comme significativement différentes.

4.6.3. Test de HARTLEY

Le test de Hartley est une extension du test de Fisher que nous avons utilisé pour comparer deux variances au cas de plusieurs variances (Hartley, 1940, 1950). Il nécessite l'égalité des effectifs des échantillons ($n_1 = n_2 = \dots = n_p = n$), mais peut être utilisé comme un test approché pour autant que les effectifs ne soient pas trop différents.

Le test consiste à calculer le rapport de la plus grande variance estimée ($\hat{\sigma}_{max}^2$) sur la plus petite variance estimée ($\hat{\sigma}_{min}^2$) parmi les variances estimées des p populations :

$$H_{obs} = \frac{\hat{\sigma}_{max}^2}{\hat{\sigma}_{min}^2}$$

et à rejeter l'hypothèse nulle lorsque $H_{obs} \geq H_{1-\alpha}$, où les valeurs critiques $H_{1-\alpha}$ sont consignées dans des tables statistiques en fonction du niveau de signification α , du nombre de populations p et du nombre de degrés de liberté $k=n-1$ (n est l'effectif de chaque échantillon, ou la moyenne des effectifs des échantillons si ceux-ci sont légèrement différents).

Le test de Hartley est très peu utilisé par les praticiens. Il ne tient compte que des deux variances estimées extrêmes ($\hat{\sigma}_{max}^2$ et $\hat{\sigma}_{min}^2$) et il est très sensible à la non normalité des populations.

Exercice 4.6

Reprenez les données du tableau 4.1 et comparez, au niveau de probabilité $\alpha=5\%$, les variances des poids des fruits des trois variétés de dattes par la méthode de Hartley ?

Solution 4.6

a) Méthode : comparaison de plusieurs variances par la procédure de Hartley

b) Conditions d'application :

- les trois distributions des poids des fruits sont normales
- les trois échantillons prélevés sont aléatoires, simples et indépendants
- les effectifs des échantillons sont égaux

c) Hypothèse nulle :

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

H_1 : deux au moins des variances sont différentes

d) Calcul : la valeur observée de Hartley :

$$H_{\text{obs}} = \frac{0.953}{0.770} = 1.24$$

est inférieure à la valeur critique $H_{0.95} = 6.94$. On ne peut donc pas rejeter H_0 .

e) Conclusion : il n'y a suffisamment pas de preuve pour considérer les variances des poids des fruits des trois variétés comme significativement différentes.

4.6.4. Test de Levene

Lorsque les populations ne sont pas normales ou elles ne sont pas au moins très proches de populations normales, le test de Levene (Levene, 1960) et ses variantes constituent une alternative robuste au test de Bartlett.

La réalisation de ces tests consiste à transformer la variable étudiée \mathbf{X} en une nouvelle variable \mathbf{Z} indiquant une dispersion et à calculer la quantité :

$$F_{\text{obs}} = \frac{\left(\sum_{i=1}^p n_i - p \right) \sum_{i=1}^p n_i (\bar{z}_{i.} - \bar{z}_{..})^2}{(p-1) \sum_{i=1}^p \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2}.$$

L'hypothèse d'égalité des variances est rejetée, au niveau de signification α , lorsque :

$$F_{\text{obs}} \geq F_{1-\alpha}$$

où $F_{1-\alpha}$ est la valeur de la variable F de Snedecor à $k_1 = p - 1$ et $k_2 = n - p$ degrés de liberté.

Dans la relation précédente :

z_{ij} est la valeur de la variable Z correspondant à l'observation j de l'échantillon i ;

$\bar{z}_{i.}$ est la moyenne des valeurs transformées de l'échantillon i ; et

$\bar{z}_{..}$ est la moyenne de toutes les valeurs transformées.

La variable transformée \mathbf{Z} peut s'exprimer sous différentes formes :

a) Levene (1960) obtient les valeurs de la variable transformée par le calcul de la distance, en valeur absolue, de chaque observation x_{ij} à la moyenne de l'échantillon \bar{x}_i :

$$z_{ij} = |x_{ij} - \bar{x}_i|$$

ou encore par le calcul du carré de cette distance :

$$z'_{ij} = (x_{ij} - \bar{x}_i)^2.$$

b) Brown et Forsythe (1974) proposent d'utiliser d'autres variantes de la procédure de Levene. Leurs méthodes consistent à soustraire la médiane (\tilde{x}_i) ou la moyenne tronquée (\bar{x}_i') plutôt que la moyenne :

$$z_{ij} = |x_{ij} - \tilde{x}_i|$$

$$z_{ij} = |x_{ij} - \bar{x}_i'|.$$

La moyenne tronquée est la moyenne des données après avoir retiré par exemple 5% des valeurs les plus petites et 5% des valeurs les plus élevées. L'objectif est de retirer les valeurs extrêmes.

Le test utilisant la médiane au lieu de la moyenne est plus efficace pour des échantillons de petites tailles et la procédure est indépendante de la distribution au niveau asymptotique.

c) O'Brien (1979) propose de calculer les valeurs z_{ij} par la formule suivante :

$$z_{ij}'' = \frac{(w + n_i - 2)n_i(x_{ij} - \bar{x}_i)^2 - w(n_i - 1)\hat{\sigma}_i^2}{(n_i - 1)(n_i - 2)}$$

dans laquelle la valeur de w permet de tenir compte de la forme de la distribution. La valeur de w est souvent prise égale à 0.5.

Des études de comparaisons basées sur des simulations ont montré que le test de Levene et ses variantes sont raisonnablement robustes. En particulier, le test de Brown-Forsythe semble l'un des plus conseillé dans la mesure où il assure une puissance pour détecter les différences entre les variances tout en contrôlant le risque de première espèce [Conover *et al.*, 1981; Olejnik et Algina, 1987].

Exercice 4.7

Reprenez les données du tableau 4.1 et comparez au niveau de probabilité $\alpha=5\%$, les variances des poids des fruits des trois variétés de dattes par la méthode de Brown et Forsythe (en utilisant les écarts des observations à la médiane en valeur absolus) ?

Solution 4.7

a) Méthode : comparaison de plusieurs variances par la procédure de Brown-Forsythe (écarts des observations aux médianes des échantillons en valeur absolue) :

b) Conditions d'application :

- les trois échantillons prélevés sont aléatoires, simples et indépendants

c) Hypothèse nulle :

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

$$H_1'': \text{deux au moins des variances sont différentes}$$

d) Calcul : La valeur observée de F :

$$F_{\text{obs}} = \frac{(24-3)(0.0240)}{(3-1)(6.7659)} = 0.037$$

est inférieure à la valeur critique $F_{1-\alpha}=3.47$ à $k_1=2$ et $k_2=21$ degrés de liberté. On ne peut donc pas rejeter H_0 . La probabilité de signification est de **0.964**.

e) Conclusion : voir la conclusion de l'exercice 4.5.

Exercice 4.8

Reprenez l'exercice 4.7 et vérifiez si vous obtenez les mêmes résultats avec les méthodes de Levene et de O'Brien ?

Solution 4.8

a) Procédure de Levene (écarts des observations aux moyennes des échantillons en valeur absolue) :

$$F_{obs} = \frac{(24-3)(0.0307)}{(3-1)(6.5267)} = 0.049 \text{ est inférieure à } F_{1-\alpha} = 3.47 \text{ (p-valeur= } \mathbf{0.952})$$

b) Procédure de Levene (écarts des observations aux moyennes des échantillons élevés au carré):

$$F_{obs} = \frac{(24-3)(0.109987)}{(3-1)(15.004)} = 0.077 \text{ est inférieure à } F_{1-\alpha} = 3.47 \text{ (p-valeur= } \mathbf{0.952})$$

c) Procédure de O'Brien avec $w=0.5$:

$$F_{obs} = \frac{(24-3)(0.1437)}{(3-1)(22.9999)} = 0.066 \text{ est inférieure à } F_{1-\alpha} = 3.47 \text{ (p-valeur= } \mathbf{0.937})$$

Conclusion : pour cet exemple, les différents tests utilisés aboutissent à la même conclusion.

Tous ces résultats peuvent être obtenus par des logiciels statistiques. Pour le logiciel SAS, on peut utiliser la procédure GLM suivante :

```
Proc GLM;
  class Varietes;
  model Poids = Varietes;
  means Varietes / hovtest=levене hovtest=BF hovtest=obrien;
run;
Proc GLM;
  class Varietes;
  model Poids = Varietes;
  means Varietes / hovtest=levене (type=ABS);
run;
```

4.7. CONCLUSION

Nous avons signalé à différentes occasions qu'avant de procéder à la comparaison de deux ou plusieurs moyennes théoriques de populations, on procède préalablement, surtout dans les cas où ce test sur les moyennes ne se révèle pas assez robuste, à un test de comparaison des variances théoriques, et ce pour choisir la procédure la mieux adaptée. Toutefois, dans d'autres cas, la comparaison de la variabilité dans les populations peut être une finalité principale de l'étude du chercheur.

Dans la littérature, on trouve plusieurs procédures qui ont été proposées pour tester l'égalité des variances de plus de deux populations. Plusieurs études ont essayé de les comparer pour opérer le bon choix.

D'une manière générale, si l'on a de bonnes raisons de croire que les populations sont normales ou au moins approximativement normales, le test de Bartlett a une plus grande performance. Dans ces conditions, il est l'un des plus utilisés, pour autant que les effectifs des échantillons

ne soient pas trop petits (<5) et que le nombre de populations ne soit pas très élevé par rapport aux effectifs des échantillons.

Si l'hypothèse de la normalité des distributions n'est pas vérifiée, le test de Levene et ses variantes offrent une alternative robuste à la procédure de Bartlett. Ils n'ont pas tendance à rejeter l'hypothèse d'égalité des variances seulement parce que les distributions des populations ne sont pas normales. Le test de Brown-Forsythe, qui utilise la médiane, semble être le plus recommandé car il a une bonne robustesse contre différentes formes de non-normalité des distributions tout en gardant une bonne puissance. Mais, si l'on connaît les formes des distributions des données, il est possible que l'une des autres variantes du test de Levene soit plus intéressante.

D'autre part, il faut dire que les tests cités dans ce chapitre peuvent être réalisés par différents logiciels statistiques. Il faut tout de même faire attention à l'appellation de certaines variantes du test de Levene. Pour certains logiciels, ce qui est appelé test de Levene est en effet sa variante de Brown-Forsythe. Un coup d'œil rapide sur l'aide du logiciel que vous utilisez permet d'éviter cette confusion.

ANALYSE DE LA VARIANCE

Utilisée pour la première fois par Sir R. A. Fisher [1925, 1935], pour analyser des données issues d'expériences agricoles, l'analyse de la variance ou ANOVA¹ reste l'une des méthodes statistiques les plus connues et les plus employées de nos jours par de nombreux chercheurs dans différentes disciplines : agriculture, médecines humaine et vétérinaire, industrie agro-alimentaire, etc.

Le but de l'ajustement d'un modèle d'ANOVA à des mesures subdivisées en groupes formés expérimentalement ou naturellement et dépendant de plusieurs effets qui interviennent simultanément, est de connaître les sources de variation les plus importantes et d'obtenir les meilleures estimations des moyennes, si les effets sont fixes, ou des variances, si les effets sont aléatoires. Il s'agit d'un ensemble de techniques de tests d'hypothèses et d'estimation de paramètres.

Il semble que plusieurs lecteurs trouvent l'appellation "*analyse de la variance*" un peu curieuse dans la mesure où l'objectif principal est de comparer des moyennes de populations. Nous verrons, en effet, qu'il faudra effectivement analyser les variances pour tester si les moyennes sont égales.

D'une manière générale, il s'agit d'étudier un modèle dans lequel une variable dépendante quantitative est expliquée par une ou plusieurs variables qualitatives (ou quantitatives, mais traitées comme qualitatives) appelées facteurs. Chaque facteur peut avoir deux ou plusieurs niveaux (ou modalités). Ainsi, l'objectif peut être, par exemple, la comparaison des teneurs en sucres réducteurs de trois variétés de dattes : *Mejhoul*, *Boufegous* et *Jihel*. Dans le cas de cet exemple, les "variétés" constituent le facteur étudié avec trois modalités et la teneur en sucres réducteurs représente la variable dépendante. La technique de l'ANOVA consiste à répartir la variation totale des réponses obtenues entre les différentes sources de variation auxquelles elle peut être attribuée.

On doit distinguer les modèles fixes, où les effets sont considérés comme des paramètres fixes, et les modèles aléatoires, où les effets sont traités comme des variables aléatoires. Lorsque ces deux types d'effets sont présents dans le même modèle, on parlera de modèles mixtes. De même, on parlera d'un modèle d'ANOVA à un ou plusieurs critères de classification selon que le modèle fait intervenir un ou plusieurs facteurs. Dans ce dernier cas, il est conseillé de limiter le nombre de facteurs pour ne pas compliquer l'interprétation des résultats.

Du point de vue pratique, l'ANOVA reste généralement une méthode facile à mettre en œuvre par le biologiste, grâce notamment aux nombreux logiciels statistiques. Cependant, il y a lieu d'être attentif quant à l'écriture du modèle mathématique et à l'interprétation des résultats fournis par le logiciel statistique lorsque le modèle n'a pas une structure équilibrée, c'est-à-dire lorsque les fréquences des combinaisons des niveaux des différents facteurs ne sont pas égales. Dans ce dernier cas, nous pensons qu'il serait peut-être utile de faire appel à l'aide d'un statisticien.

¹ En anglais : *Analysis of variance*.

D'autre part, nous voulons aussi signaler que certains termes spécifiques aux techniques d'échantillonnage et d'expérimentation seront cités dans quelques passages. Le modèle de l'ANOVA est en effet lié au dispositif expérimental mis en place et/ou au mode de prélèvement des données.

Dans ce qui suit, nous présenterons séparément l'analyse de la variance à un critère de classification (chapitre 5) et l'analyse de la variance à deux critères de classification (chapitre 6) selon qu'un ou deux facteurs sont étudiés. Nous exposerons, pour chacune des deux méthodes, les tests de signification et les méthodes d'estimation de paramètres, en s'appuyant sur des exemples. Nous présenterons en détail les formules de calcul bien que nous sachions que l'ordinateur s'en charge.

L'extension à l'analyse de la variance à trois ou plus de trois critères de classification se fait selon les mêmes principes qui régissent l'analyse de la variance à deux critères de classification, mais l'interprétation de ce qu'on appelle des interactions des modèles croisés ou partiellement croisés devient souvent moins évidente.

CHAPITRE 5

ANALYSE DE LA VARIANCE A UN CRITERE DE CLASSIFICATION

5.1. INTRODUCTION

Nous avons vu, au chapitre 3, comment comparer les moyennes de deux populations normales, en sélectionnant indépendamment un échantillon aléatoire et simple dans chacune d'elles. Or, il se peut que l'on soit confronté dans la pratique à la comparaison du comportement moyen de plusieurs populations définies par ce qu'on appelle un critère de classification. Il s'agit ici de "généraliser" le test t de Student que nous avons utilisé pour comparer deux moyennes théoriques.

On parle généralement de traitements pour désigner les différents niveaux du facteur de variabilité dont on souhaite étudier l'influence sur une variable dépendante quantitative. C'est le cas lorsqu'on souhaite comparer, par exemple, les rendements, supposés suivre des distributions normales, de quatre variétés de tomates en observant pour chacune d'elles quelques parcelles. Nous avons ici un seul facteur, appelé "variétés", possédant quatre modalités représentant les quatre variétés.

Dans ce cas, il n'est pas approprié de vouloir réaliser six tests t de Student pour comparer deux à deux les moyennes théoriques des quatre variétés. En effet, en fixant un risque de première espèce α pour chacun des six tests, le risque global de première espèce, c'est-à-dire la probabilité de considérer à tort au moins un test significatif, est bien plus grand que α . Dans ces conditions, l'analyse de la variance à un critère de classification, en abrégé "ANOVA1", est appropriée car c'est une procédure qui nous permettra de vérifier, en un seul test, s'il existe au moins une moyenne significativement différente des autres en fixant le risque de première espèce à α .

Après la présentation du principe de l'ANOVA 1 [paragraphe 5.2] et des conditions d'application de cette méthode [paragraphe 5.3], nous exposerons les deux modèles possibles, à savoir le modèle fixe et le modèle aléatoire [paragraphe 5.4]. Nous entamerons ensuite les tests de signification des effets [paragraphe 5.5] et les problèmes d'estimation de certains paramètres [paragraphe 5.6].

5.2. PRINCIPE DE L'ANOVA 1

5.2.1. Exemple introductif et notations

1° On souhaite tester s'il existe des différences significatives entre les acidités moyennes du jus, exprimée en grammes par litre de jus, de trois classes d'hybrides de mandarines. Pour ce faire, un chercheur à l'IAV Hassan II a déterminé l'acidité dans 10 prélèvements de jus choisis au hasard pour chacune des trois classes. Chaque détermination étant obtenue à partir du jus de 20 fruits (tableau 5.1).

Tableau 5.1. Acidité, en grammes par litre de jus, obtenues pour trois classes d'hybrides de Mandarines.

Classes d'hybrides											Moyennes
Classe 1	8.28	11.92	8.62	11.21	11.24	11.33	11.87	10.59	10.35	10.86	10.627
Classe 2	14.97	14.20	12.14	12.74	13.80	14.04	13.27	14.06	13.34	13.27	13.583
Classe 3	11.31	10.55	12.76	12.94	12.54	11.30	12.96	11.33	12.12	12.46	12.027
Moyenne des trois échantillons											12.079

Dans cet exemple :

- si on mélange les 30 observations, on ne peut les reclasser qu'en fonction de la classe d'hybrides de mandarines. On dit que les données ont un seul critère de classification qui est "**classes d'hybrides de mandarines**" ;
- le facteur "classes d'hybrides de mandarines" a trois variantes "ou niveaux" : la classe 1, la classe 2 et la classe 3. On peut dire aussi que l'on a trois populations ou trois traitements ;
- les 10 observations réalisées dans une même classe d'hybrides peuvent être interverties dans n'importe quel ordre. Ce sont de simples répétitions aléatoires et hiérarchisées au critère de classification sans toutefois constituer un critère ;
- les observations prélevées dans chacune des classes d'hybrides sont indépendantes des observations de chacune des deux autres classes ;
- la variable dépendante étudiée, appelée aussi la réponse, est l'acidité. Les acidités moyennes observées sont données dans la dernière colonne du tableau 5.1. Ces moyennes (\bar{y}_i , $i=1, 2, 3$) sont des estimations non biaisées des acidités moyennes théoriques des trois classes d'hybrides de mandarines (μ_i , $i=1, 2, 3$).

Si l'on est tenté de réaliser, pour cet exemple, trois comparaisons des moyennes deux à deux par le test t de Student au niveau α , c'est-à-dire les comparaisons des moyennes des classes 1 et 2, des moyennes des classes 1 et 3 et des moyennes des classes 2 et 3, on doit avoir à l'esprit ce qui suit :

- pour chaque comparaison de deux moyennes, la probabilité d'accepter l'hypothèse nulle d'égalité des deux moyennes lorsque celle-ci est vraie est de $1 - \alpha$;
- pour les trois comparaisons des moyennes deux à deux, cette probabilité devient $(1 - \alpha)^3$.

La probabilité de rejeter à tort l'hypothèse nulle au moins une fois au cours des trois comparaisons est alors obtenue par $1 - (1 - \alpha)^3$. Autrement dit, pour $\alpha=0.05$, on a 14 chances sur 100 de rejeter à tort l'hypothèse nulle au moins une fois au cours des trois comparaisons. Cette probabilité serait encore plus grande si le nombre de populations était plus important.

Pour p populations, le nombre de comparaisons des moyennes deux à deux est de $k = p(p - 1) / 2$ et la probabilité de rejeter à tort l'hypothèse nulle au moins une fois au cours des k comparaisons devient $1 - (1 - \alpha)^k$.

2° D'une manière générale, on considère p populations (on parle aussi de p traitements ou d'une population divisée en p sous-populations). Dans chacune de celles-ci, on prélève un échantillon d'individus et, pour chacun de ces individus, on doit disposer de la valeur de la variable dépendante. Les notations que seront adoptées sont les suivantes :

y_{ik} : la valeur de la variable dépendante pour le $k^{\text{ème}}$ individu de l'échantillon extrait de la population i (dans le tableau 5.1, $y_{2,5}=13.80$) ;

- \bar{y}_i : la moyenne observée (ou empirique) des valeurs de l'échantillon extrait de la population i (dans le tableau 5.1, $\bar{y}_3 = \mathbf{12.027}$) ;
- \bar{y} : la moyenne générale observée (dans le tableau 5.1, $\bar{y} = \mathbf{12.079}$) ;
- n_i : l'effectif de l'échantillon prélevé de la population i (dans le tableau 5.1, $n_i = \mathbf{10}$) ;
- n : l'effectif total des p échantillons (dans le tableau 5.1, $n = \mathbf{30}$).

5.2.2. Logique sous-jacente à l'analyse de la variance

En reprenant les données des trois classes d'hybrides de mandarines, la question que l'on se pose est la suivante :

"Est-ce que les résultats, bien qu'ils soient obtenus de 10 répétitions seulement, permettent de conclure à des différences significatives entre les acidités moyennes des trois classes en question ?"

On cherche en effet à tester l'hypothèse nulle qui concerne l'égalité des acidités moyennes théoriques μ_1 , μ_2 et μ_3 des trois classes d'hybrides de mandarines. Les valeurs de ces moyennes sont inconnues et on se basera sur les moyennes observées (\bar{y}_1 , \bar{y}_2 et \bar{y}_3) et les variances estimées ($\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ et $\hat{\sigma}_3^2$) pour pouvoir répondre à la question.

1° La logique de l'ANOVA 1 consiste à examiner l'importance de la variation entre les moyennes des échantillons (*between*) par rapport à la variation à l'intérieur des échantillons (*within*).

Dans l'exemple que nous traitons (tableau 5.1), il apparaît, sans faire de calcul, que la variation entre les moyennes des échantillons (\bar{y}_1 , \bar{y}_2 et \bar{y}_3) est assez grande en comparaison avec la variation à l'intérieur des échantillons. En d'autres termes, il est logique de penser, en s'appuyant également sur la figure 5.1 (a), que les moyennes observées des trois échantillons ne constituent pas une bonne estimation d'une moyenne générale théorique μ . On a donc tendance à conclure à des différences entre les acidités moyennes théoriques des trois classes d'hybrides de mandarines. Il reste à prouver ce qu'on vient d'avancer et c'est l'objet de l'ANOVA. Celle-ci permet de démontrer que ces différences sont statistiquement significatives, c'est-à-dire qu'elles ne résultent pas de l'effet du hasard.

2° Supposons maintenant qu'à avec les mêmes moyennes observées (\bar{y}_1 , \bar{y}_2 et \bar{y}_3) des trois classes, on avait obtenu les résultats du tableau 5.2 (données simulées). Il ressort que la variation entre les moyennes des échantillons est moins importante en comparaison avec la variation dans les échantillons. La figure 5.1 (b) montre, en effet, un recouvrement important des valeurs des trois échantillons. Dans ce cas, il est peu vraisemblable de conclure à des différences significatives entre les moyennes des trois populations. Si l'ANOVA confirme l'absence de différences, on pourra combiner les valeurs des trois échantillons pour obtenir une estimation de la moyenne générale théorique μ .

Tableau 5.2. Acidité, en grammes par litre de jus, de trois classes d'hybrides de mandarines (données fictives).

Classes d'hybrides												Moyennes
Classe 1	13.67	12.36	7.58	10.68	9.05	6.22	6.84	15.03	10.72	14.12		10.627
Classe 2	10.89	12.02	14.27	13.34	11.88	16.05	17.51	9.88	15.46	14.53		13.583
Classe 3	16.14	10.81	11.14	8.02	16.82	7.28	9.86	11.88	15.04	13.28		12.027

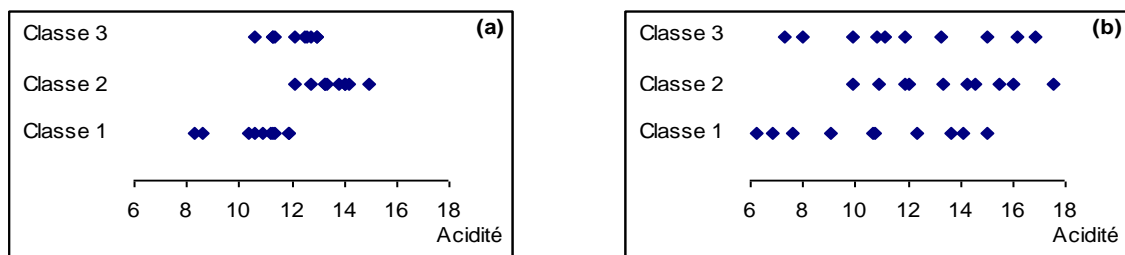


Figure 5.1. Superpositions des valeurs de l'acidité des échantillons des trois classes d'hybrides de mandarines ((a) pour les valeurs du tableau 5.1; (b) pour les valeurs du tableau 5.2).

3° Tout indique donc que l'on doit comparer la variabilité entre les moyennes observées à la variabilité des valeurs autour de ces moyennes.

5.3. CONDITIONS D'APPLICATION

L'analyse de la variance à un critère de classification repose sur plusieurs hypothèses :

- Chaque échantillon de n_i observations est prélevé dans la population i de manière aléatoire et simple, c'est-à-dire que tous les individus de la population ont la même chance d'appartenir à l'échantillon et le prélèvement d'un individu ne dépend pas du prélèvement des autres individus. De même, les p échantillons sont prélevés indépendamment l'un de l'autre. Les observations doivent donc être indépendantes les unes des autres aussi bien à l'intérieur des échantillons qu'entre les échantillons.
- La variable dépendante pour la population i suit une distribution normale (de Gauss) de moyenne μ_i et de variance σ_i^2 . Cette hypothèse, requise pour réaliser les tests de signification et entourer certains paramètres d'intervalles de confiance, peut être vérifiée en utilisant un test d'ajustement à une distribution normale (cf. chapitre 8).
- Les variances théoriques des p populations sont égales, c'est-à-dire que la variabilité des observations autour de la moyenne d'une population est la même pour toutes les populations. Cette hypothèse peut être testée en utilisant un test d'égalité de plusieurs variances (cf. chapitre 4). La variance commune théoriques des populations sera notée σ^2 .

Si certaines de ces conditions d'application sont loin d'être vérifiées, on peut, en premier lieu, essayer des transformations de variables dans le but de s'approcher des distributions normales et/ou de stabiliser les variances. Si les résultats de telles transformations ne sont pas satisfaisants, on peut utiliser des méthodes non paramétriques.

5.4. MODÈLES D'ANOVA 1

5.4.1. Généralités

Le facteur étudié peut être fixe ou aléatoire. On parlera de facteur fixe lorsque ses niveaux sont contrôlés et reproductibles d'une expérience à l'autre. On peut citer, comme exemples, le facteur "sexe" lorsqu'on souhaite étudier son effet sur le pourcentage de carcasse après abattage des bovins d'une même race, ou encore le facteur "aliments" lorsqu'on souhaite comparer l'ingestion de trois aliments spécifiés chez des moutons adultes de même poids.

On parlera de facteur aléatoire lorsque ses niveaux étudiés sont eux-mêmes tirés aléatoirement d'un ensemble plus grand de niveaux. Les niveaux sélectionnés ne sont pas reproductibles d'une expérience à l'autre, puisque, à chaque fois, ce sont des niveaux différents qui sont tirés.

Supposons que l'on souhaite tester si la taille moyenne d'une espèce de poisson ne diffère pas selon les différentes rivières du bassin atlantique. Si l'on choisit aléatoirement trois rivières et que l'on mesure dans chacune d'elles les tailles d'une dizaine de poissons, on dira que le facteur "Rivières" est bien un critère aléatoire.

Selon que le facteur est fixe ou aléatoire, on parlera du modèle fixe ou du modèle aléatoire de l'ANOVA 1. Le modèle fixe s'appelle aussi modèle I et le modèle aléatoire s'appelle aussi modèle II. Nous verrons dans ce paragraphe comment étudier chacun des deux modèles.

5.4.2. Modèle fixe

5.4.2.1. Définition

Dans le cas du modèle fixe, tous les niveaux du facteur étudié sont considérés dans "l'expérience". On considère que leur nombre est p .

5.4.2.2. Hypothèse nulle

L'hypothèse nulle concerne l'égalité des moyennes théoriques des p populations. Elle s'écrit :

$$(5.1) \quad H_0 : \mu_1 = \mu_2 = \dots = \mu_p.$$

contre l'alternative :

$$H_1 : \mu_i \neq \mu_j \quad \text{pour au moins deux populations.}$$

L'hypothèse alternative signifie, dans le cas de notre exemple, qu'il y a un effet des classes des hybrides de mandarines sur l'acidité.

On se basera sur les valeurs obtenues pour les p échantillons pour rejeter ou ne pas rejeter l'hypothèse nulle :


- s'il y a des différences importantes entre les moyennes des populations, on s'attend à ce qu'il en soit de même pour les moyennes des échantillons ;
- le degré de fausseté de H_0 dépend, comme nous l'avons vu, de la grandeur de la variation entre les échantillons, que nous appellerons variation factorielle, par rapport à la variation à l'intérieur des échantillons, que nous appellerons variation résiduelle ;
- l'hypothèse nulle est rejetée si deux au moins des moyennes sont différentes. Pour savoir à quel niveau se situent les différences entre les moyennes lorsqu'on rejette H_0 , on utilisera les méthodes de comparaisons multiples de moyennes (Cf. chapitre 7).

5.4.2.3. Modèles observé et théorique


a) Modèle observé

Il apparaît que l'écart entre toute observation k issue de l'échantillon i et la moyenne des trois échantillons est donné par la relation suivante :

$$(y_{ik} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ik} - \bar{y}_i) \quad (5.2)$$



Ecart de l'échantillon à la moyenne générale
(Ecart factoriel)



Ecart dans l'échantillon
(Ecart résiduel)

Ces écarts apparaissent mieux si l'on représente, sur un axe horizontal, toutes les valeurs individuelles (y_{11}, y_{12}, \dots) ainsi que les moyennes ($\bar{y}_1, \bar{y}_2, \dots, \bar{y}_p$) et la moyenne générale \bar{y} . La figure 5.2 présente ces écarts pour les échantillons des deux premières classes d'hybrides de mandarines (cf. tableau 5.1). On voit comment l'écart entre la première observation y_{11} et la moyenne générale \bar{y} est décomposé en deux écarts.

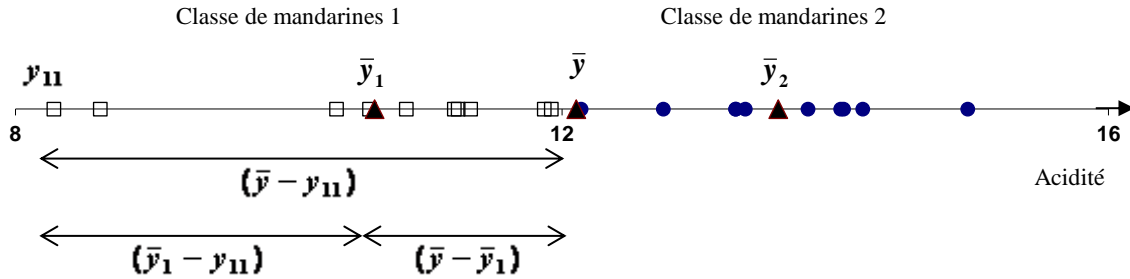


Figure 5.2. Représentation des écarts entre les valeurs individuelles et les moyennes dans le cas des deux premières classes d'hybrides de mandarines.

L'équation fondamentale de l'ANOVA 1 s'obtient en élevant les deux membres de l'équation 5.2 au carré puis en sommant pour toutes les observations. On obtient :

$$\begin{aligned} \sum_{i=1}^p \sum_{k=1}^{n_i} (y_{ik} - \bar{y})^2 &= \sum_{i=1}^p \sum_{k=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2 \\ &= \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2 \end{aligned} \quad (5.3)$$

Cela rappelle les formules des sommes des carrés des écarts. Dans ces conditions, l'expression 5.3 peut s'écrire sous la forme :

$$\boxed{SCE_t = SCE_a + SCE_r} \quad (5.4)$$

qui s'appelle l'équation de l'ANOVA 1. Dans cette équation :

- SCE_t représente la somme des carrés des écarts totale ;
- SCE_a représente la somme des carrés des écarts factorielle ou entre les échantillons ;
- SCE_r représente la somme des carrés des écarts résiduelle ou à l'intérieur des échantillons.

b) Modèle théorique

Pour faire de l'inférence statistique, on doit aussi écrire le modèle théorique. On considère que chaque observation y_{ik} est la valeur d'une variable aléatoire Y_{ik} et que celle-ci suit une distribution normale de moyenne μ_i et de variance σ^2 . Les variables aléatoires Y_{ik} sont supposées être indépendantes et de mêmes variances. Le modèle théorique s'écrit :

$$\boxed{Y_{ik} = \mu_i + \varepsilon_{ik}} \quad (5.5)$$

ou de façon équivalente :

$$Y_{ik} = \mu_i + \varepsilon_{ik} .$$

Toute valeur observée Y_{ik} est en effet la somme de trois effets :

- une moyenne générale ou l'effet moyen général : μ ,
- un effet du facteur étudié associé au niveau i : $a_i = \mu_i - \mu$,
- un terme aléatoire associé à l'unité k soumise au traitement i : $\varepsilon_{ik} = Y_{ik} - \mu_i$.

Pour comprendre l'idée du modèle de l'expression (5.5), supposons que l'acidité moyenne des trois classes d'hybrides de mandarines est de 12.08 g/l et que celle de la classe 1 est inférieure de 1.45 g/l par rapport à la moyenne générale. Dans ce cas, le modèle qui explique l'acidité de la première observation de la classe d'hybride 1 qui est de 8.28 (tableau 5.1) s'écrit :

$$8.28 = 12.08 - 1.45 - 2.35.$$

Les quantités ε_{ik} sont aussi appelées erreurs ou variables résiduelles, et les quantités a_i sont aussi appelées effets principaux. Ces derniers paramètres sont tels que :

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

Il en découle que l'hypothèse nulle signifie l'absence d'effets du facteur étudié, et elle peut donc aussi s'écrire :

$$H_0 : a_1 = a_2 = \dots = a_p = 0 .$$

Les hypothèses sur les variables résiduelles sont équivalentes aux mêmes hypothèses sur la variable Y (cf. paragraphe 5.3). L'interprétation des résultats de l'ANOVA n'est en effet valide que lorsque ces hypothèses sont réunies. On doit veiller à ce que les variables ε_{ik} soient :

- distribuées selon des lois normales de moyenne nulle ;
- indépendantes les unes des autres ;
- de variance constance σ^2 .

6.4.3. Modèle aléatoire

5.4.3.1. Définition

Dans le cas du modèle aléatoire, on s'intéresse à un nombre infini ou presque infini de populations. L'hypothèse nulle concerne l'égalité des moyennes de toutes les populations. Ces dernières ne peuvent être échantillonnées toutes, on procède alors à un échantillonnage à deux degrés :

- on prélève aléatoirement un échantillon de p populations parmi l'infinité des populations (unités du premier degré) ;
- on prélève ensuite, à l'intérieur de chacune de ces p populations tirées, un échantillon aléatoire et simple d'un certain nombre d'individus (unités du second degré).

On parle d'un facteur aléatoire car seul un sous-ensemble de p niveaux fait l'objet des observations alors que le chercheur s'intéresse à tous les niveaux. En pratique, comme facteurs à effets aléatoires, on trouve des animaux, des variétés d'une espèce végétale, des personnes, etc. On peut citer les deux exemples suivants :

- Pour contrôler si le pH du sol est le même dans un champ, on peut choisir au hasard dix emplacements et, dans chacun de ceux-ci, on creuse quelques trous pour lesquels on

détermine le pH. Le facteur "emplacements" est aléatoire et la conclusion concernera tout le champ ;

- Pour savoir si le pourcentage de germination des semences d'une espèce forestière est le même pour des provenances différentes, on peut choisir au hasard quelques provenances et on calcule le pourcentage de germination en utilisant quelques cuves de germination pour les semences de chaque provenance. Le facteur "provenances" est aléatoire.

5.4.3.2. Hypothèse nulle

L'hypothèse nulle concerne l'infinité des niveaux du facteur et non seulement les p niveaux (ou populations) sélectionnés. Or, si les valeurs d'une distribution sont toutes égales entre elles alors l'écart-type est nul. Donc, l'hypothèse nulle peut être exprimée par la nullité de l'écart-type des effets principaux :

$$H_0 : \sigma_A = 0 \quad \text{contre l'hypothèse alternative} \quad H_1 : \sigma_A > 0. \quad (5.6)$$

5.4.3.3. Modèles observé et théorique

Le modèle observé est le même que dans le cas du modèle fixe. Le modèle théorique s'écrit également de la même manière sauf que les moyennes des populations M_i ($i=1, 2, \dots, p$) sont ici des variables aléatoires puisqu'elles résultent d'un tirage aléatoire. La moyenne générale (μ), quant à elle, est fixe :

$$Y_{ik} = \mu + A_i + \varepsilon_{ik} . \quad (5.7)$$

En plus des effets des variables résiduelles qui sont toujours aléatoires, les effets principaux $A_i = M_i - \mu$ sont également aléatoires. Les variables ε_{ik} sont normales et indépendantes de moyenne 0 et de variances σ^2 et les variables A_i sont normales et indépendantes des ε_{ik} de moyenne 0 et de variances σ_A^2 .

5.5. TESTS D'HYPOTHESES

5.5.1. Test statistique

Nous avons vu que l'analyse de la variance est basée sur la partition de la somme des carrés des écarts totale en une somme des carrés des écarts factorielle et une somme des carrés des écarts résiduelle (cf. expression 5.3). Mais, pour mieux apprécier l'information sur la variabilité entre les groupes et dans les groupes, on doit être amené à faire des pondérations, en divisant les sommes des carrés des écarts factorielle et résiduelle par leurs nombres de degrés de liberté. On obtient :

- a) La variance due aux erreurs (ou intra-groupe) qui est donnée par la quantité :

$$CM_r = \frac{SCE_r}{n. - p} . \quad (5.8)$$

Elle exprime une estimation de la variabilité des observations à l'intérieur des p populations. C'est la variance non expliquée par le modèle ;

b) La variance entre les groupes (ou inter-groupe) qui est donnée par la quantité :

$$CM_a = \frac{SCE_a}{p-1} . \quad (5.9)$$

Elle exprime une estimation de la variabilité entre les moyennes échantillonnées. C'est la variance expliquée par le facteur étudié.

On retient aussi que :

- CM_r est une estimation de la variance commune σ^2 ;
- Les quantités CM_r et CM_a sont communément appelées carré moyen résiduel et carré moyen factoriel.

Rappelons-nous que l'objectif est de comparer la variance des moyennes interclasses à la variance intra-classes ou résiduelle. Dans le chapitre 4, nous avons utilisé à cet égard le test F de Snedecor.

Lorsque l'hypothèse nulle est vraie et que les conditions d'application sont supposées vérifiées, le test de l'hypothèse nulle se base sur le calcul de la statistique :

$$F_{obs} = \frac{SCE_a / (p-1)}{SCE_r / (n.-1)} = \frac{CM_a}{CM_r} \quad (5.10)$$

qui une valeur observée de la variable F de Snedecor à $k_1=p-1$ et $k_2=n.-p$ degrés de liberté. Il s'agit en effet du rapport de valeurs observées de deux variables khi-deux divisée chacune par son nombre de degrés de liberté.

On rejette l'hypothèse nulle, donnée par les expressions (5.1) et (5.6) selon le type de modèle, au niveau de signification α , lorsque :

$$F_{obs} \geq F_{1-\alpha} ,$$

c'est-à-dire lorsque le carré moyen CM_a est suffisamment grand par rapport au carré moyen CM_r .

Si l'hypothèse nulle est rejetée, on évalue le degré de signification (appelé aussi probabilité d'erreur et noté *p-value*) qui est, dans ces conditions, inférieur à α , où α est le risque de première espèce (figure 5.3).

Plus la valeur de *p-value* est petite, plus confortable est la conclusion de rejeter l'hypothèse nulle. La plupart des logiciels statistiques donnent cette probabilité à la fin de chaque test :

$$p - \text{valeur} = P(F > F_{obs}) .$$

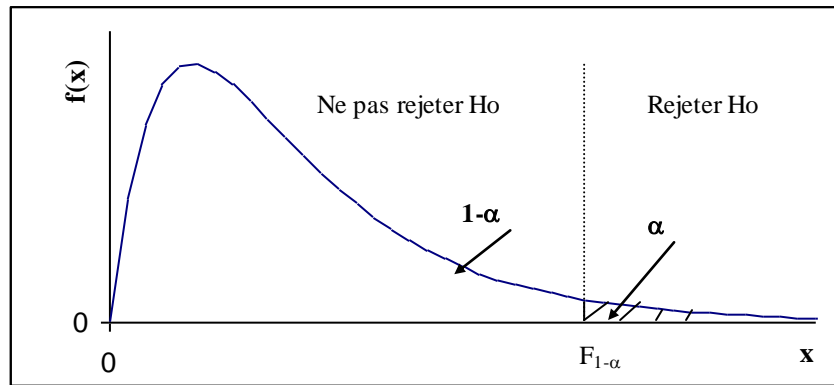


Figure 5.3. Valeur critique de la distribution F de Snedecor et région de rejet de l'hypothèse d'absence de l'effet du facteur étudié.

Remarque :

Lorsque le niveau de signification α n'est pas spécifié, il est de coutume d'utiliser la valeur 0.05 comme valeur par défaut. Si l'hypothèse nulle est rejetée au niveau 0.05, on vérifie si on peut toujours la rejeter au niveau 0.01. De même, si l'hypothèse nulle est rejetée au niveau 0.01, on vérifie si on peut toujours la rejeter au niveau 0.001. On retient :

- lorsque l'hypothèse nulle est rejetée au niveau $\alpha=0.05$, on dit qu'il existe des différences **significatives** entre les moyennes ;
- lorsque l'hypothèse nulle est rejetée au niveau $\alpha=0.01$, on dit qu'il existe des différences **hautement significatives** entre les moyennes ;
- lorsque l'hypothèse nulle est rejetée au niveau $\alpha=0.001$, on dit qu'il existe des différences **très hautement significatives** entre les moyennes.

5.5.2. Formules de calculs

En utilisant les notations des paragraphes précédents, notamment celles du paragraphe 5.2.1, on peut écrire les formules de calculs de l'ANOVA 1 suivantes :

- l'effectif total des observations est : $n_{\cdot} = \sum_{i=1}^p n_i$
- la somme des observations de l'échantillon tiré de la population i est : $y_{i\cdot} = \sum_{k=1}^{n_i} y_{ik}$
- la somme totale de toutes les observations est : $y_{\cdot\cdot} = \sum_{i=1}^p y_{i\cdot}$
- la somme des carrés générale est : $T = \sum_{i=1}^p \sum_{k=1}^{n_i} y_{ik}^2$
- le facteur de correction est : $C = y_{\cdot\cdot}^2 / n_{\cdot}$
- la somme des carrés des écarts des observations de l'échantillon tiré de la population i :

$$SCE_i = \sum_{k=1}^{n_i} y_{ik}^2 - \frac{1}{n_i} \left(\sum_{k=1}^{n_i} y_{ik} \right)^2$$

- la somme des carrés des écarts résiduelle est donnée par la somme des carrés des écarts par échantillon : $SCE_r = \sum_{i=1}^p SCE_i$
- la somme des carrés des écarts totale est : $SCE_t = T - C$
- la somme des carrés des écarts factorielle se calcule par différence :

$$SCE_a = SCE_t - SCE_r$$

ou par la formule d'une somme des carrés des écarts : $SCE_a = \sum_{i=1}^p \frac{y_{i\cdot}^2}{n_i} - C$

5.5.3. Tableau d'analyse de la variance

Quels que soient le modèle de l'analyse de la variance et le nombre de critères de classification (ANOVA 1, ANOVA 2, etc.), les résultats des calculs sont souvent consignés dans un tableau appelé tableau d'analyse de la variance.

Ce tableau est constitué généralement de six colonnes :

- La première présente les différentes sources de variation, c'est-à-dire une décomposition de la variation totale en composantes expérimentales et d'erreurs d'échantillonnage ;
- Les deuxième et troisième colonnes donnent respectivement les degrés de liberté et les sommes des carrés des écarts relatifs à chaque source de variation ;
- Viennent ensuite, dans la quatrième colonne, les carrés moyens qui s'obtiennent en divisant les sommes des carrés des écarts par leurs degrés de liberté ;
- La cinquième colonne présente la valeur observée de la variable F de Snedecor ;
- Le degré de signification ou probabilité d'erreur (*p-values*) apparaît dans la dernière colonne.

Dans le cas de l'ANOVA 1, le tableau d'analyse de la variance se dresse de la même manière aussi bien pour le modèle fixe que pour le modèle aléatoire (tableau 5.3). La fléchette indique la manière de calculer le rapport des carrés moyens pour obtenir la valeur observée de la variable F de Snedecor.

Tableau 5.3. Tableau d'analyse de la variance dans le cas de l'ANOVA 1.

Sources de variation	Degrés de liberté	SCE	Carrés moyens	F_{obs}	Prob
Variation factorielle (entre populations)	p-1	SCE_a	CM_a	F_{obs}	<i>p-value</i>
Variation résiduelle (dans populations)	n.-p	SCE_r	CM_r		
Total	n.-1	SCE_t			

Rappelons qu'on rejette l'hypothèse H_0 lorsque $F_{obs} \geq F_{1-\alpha}$ à (**p-1**) et (**n.-p**) degrés de liberté. La valeur de *p-valeur* est, dans ce cas, inférieure ou égale au seuil de signification α . Si on rejette H_0 on dit que le test est significatif et on conclut qu'il y a au moins deux moyennes qui sont différentes.

Exemple 5.1

Reprenez les données du tableau 5.1 et testez s'il existe des différences significatives entre les acidités des trois classes d'hybrides de mandarines ?

Solution 5.1

a) **Méthode** : ANOVA 1, modèle fixe (ou modèle à effets fixes)

b) **Conditions d'application** :

- les trois échantillons sont aléatoires, simples et indépendants entre eux,
- les distributions des trois populations sont normales,
- les variances des trois populations sont égales : Le test de Bartlett donne $\chi_{obs}^2 = 2.143$ et *p-valeur*=0.343 (>0.05), ce qui conduit au non rejet de l'hypothèse d'égalité des variances.

c) Hypothèse nulle : $H_0 : \mu_1 = \mu_2 = \mu_3$

Hypothèse alternative : $H_0 : \mu_i \neq \mu_j$ pour au moins deux hybrides

d) Calcul :

- Nombre de populations : $p=3$
- Effectifs des échantillons : $n_1=n_2=n_3=10$ observations
- Effectif total : $n.=30$ observations
- Total par échantillon : $Y_1=106.27$; $Y_2=135.83$; $Y_3.=120.27$
- Total général : $Y_{..}=106.27+135.83+120.27=362.37$
- Terme de correction : $C=(362.37/30)=4377.07$
- Somme des carrés : $T=8.28^2 + 11.92^2 + \dots + 12.46^2 = 4447.14$
- Somme des carrés des écarts par échantillon :
 $SCE_1 = (8.28^2 + \dots + 10.86^2) - 106.27^2/10 = 14.096$; $SCE_2 = 5.836$ et $SCE_3 = 6.411$
- Somme des carrés des écarts résiduelle : $SCE_r = 14.096 + 5.8358 + 6.41106 = 26.342$
- Somme des carrés des écarts totale : $SCE_t = T - C = 4447.14 - 4377.07 = 70.072$
- Somme des carrés des écarts factorielle : $SCE_a = SCE_t - SCE_r = 70.072 - 26.342 = 43.730$
ou $SCE_a = (106.27^2/10 + 135.83^2/10 + 120.27^2/10) - 4377.07 = 43.730$

Le tableau d'analyse de la variance :

Source de variation	d.l	SCE	CM	F _{obs}	p-value
Classes d'hybrides	2	43.7302	21.865	22.41	1.86 10 ⁻⁶
Variations résiduelles	27	26.3420	0.9756	-	-
Total	29	70.072	-	-	-

- Pour $\alpha=0.05$, $F_{1-\alpha} = 3.35$ et puisque $F_{obs} \geq F_{1-\alpha}$, on rejette H_0
- Pour $\alpha=0.01$, $F_{1-\alpha} = 5.49$ et puisque $F_{obs} \geq F_{1-\alpha}$, on rejette H_0
- Pour $\alpha=0.001$, $F_{1-\alpha} = 9.02$ et puisque $F_{obs} \geq F_{1-\alpha}$, on rejette H_0

Puisque l'hypothèse nulle est rejetée au niveau $\alpha=0.001$, on conclut à des différences très hautement significatives entre les moyennes des trois classes d'hybrides. Les valeurs critiques de F sont obtenues des tables statistiques, mais elles peuvent être obtenues, entre autres, par la fonction **inverse.loi.F.droite** (α ; 2; 27) du tableur Excel (2 et 27 sont les degrés de liberté de F) ;

La valeur de la *p-valeur* est ici inférieure à 0.0001, ce qui veut dire que l'hypothèse nulle doit être rejetée avec un risque inférieur à 0.001. Cette probabilité est donnée par les logiciels statistiques, mais elle peut être calculée, entre autres, par la fonction **Loi.F.droite** (22,41; 2; 27) du tableur Excel, où 22.41 est la valeur de F observée.

5.5.4. Taille de l'effet du facteur

Nous avons vu qu'un test F significatif indique l'existence de différences significatives entre les moyennes. Or, il est aussi important de donner une mesure qui reflète la taille de l'effet du facteur étudié.

Plusieurs paramètres ont été proposés dans la littérature pour estimer cette taille de l'effet d'un facteur. Parmi les plus connus, on peut citer le éta-carré (η^2) et le oméga-carré (ω^2).

1) Le éta-carré correspond au rapport :

$$\eta^2 = \frac{SCE_a}{SCE_t}$$

qui rappelle le coefficient de détermination R^2 dans un problème de régression. Les deux quantités expriment la part de la variabilité de la variable dépendante qui est imputable à la variable indépendante (c'est-à-dire au facteur). La valeur de η^2 est d'autant plus grande que l'effet mis en évidence est important. Eta-carré présente le plus souvent l'inconvénient de surestimer l'intensité de l'effet du facteur.

2) Le oméga-carré est une sorte d'ajustement de η^2 pour corriger la surestimation de l'effet. Cette quantité se calcule par :

$$\omega^2 = \frac{SCE_a - (p - 1)CM_r}{SCE_t + CM_r}$$

et s'interprète de la même manière que le êta-carré.

Exemple 5.2

Reprenez l'exemple 5.1 et représentez graphiquement les résultats de l'analyse de la variance. Donnez également une mesure de la taille de l'effet du facteur "classes d'hybrides de mandarines".

Solution 5.2

- a) On peut utiliser un diagramme en bâtonnets pour représenter les moyennes des 3 traitements en les accompagnant de la représentation de la variabilité sous forme d'écart-type :

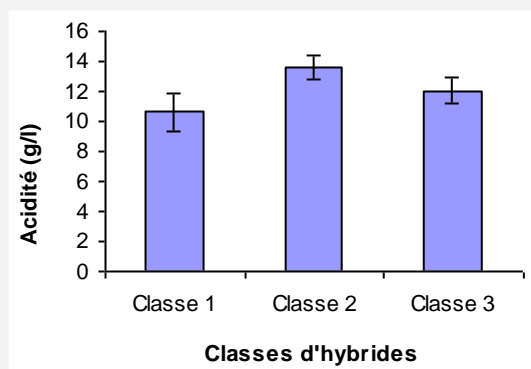


Figure 5.4. Acidités moyennes des trois classes d'hybrides (\pm les écarts-types).

Remarques :

- Si la variable indépendante (facteur) est métrique ou ordinale (doses, températures, etc.), on utilisera de préférence une représentation sous forme de points reliés entre eux par des lignes brisées ;
- La variabilité pourra être aussi représentée par l'erreur-standard (écart-type de la moyenne) au lieu des écarts-types.

b) Taille de l'effet du facteur :

Puisque le test est significatif (rejet de H_0), on peut calculer la valeur de Oméga-carré :

$$\omega^2 = \frac{43.73 - (3 - 1)0.9756}{70.072 + 0.9756} = 0.59.$$

Il s'avère que 59% de la variabilité de l'acidité est attribuée aux classes d'hybrides de mandarines.

5.6. ESTIMATIONS DE PARAMETRES

Plus qu'un test d'hypothèse, l'analyse de la variance peut avoir aussi comme objet l'estimation de paramètres.

1) Dans le cas du modèle fixe, on est souvent amené à estimer la variance commune des populations (σ^2), les moyennes des différentes populations (μ_i), la moyenne générale si l'hypothèse nulle n'est pas rejetée (μ) et les effets des différents niveaux du facteur étudié (α_i).

On peut montrer, par le calcul des espérances mathématiques, que le carré moyen résiduel est une estimation sans biais de la variance commune des populations :

$$\hat{\sigma}^2 = CM_r.$$

L'intervalle de confiance de cette variance au niveau de confiance $(1-\alpha)100\%$ peut être obtenu par :

$$\frac{SCE_r}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{SCE_r}{\chi^2_{\alpha/2}}$$

où la variable χ^2 possède n.-p degrés de liberté.

Les moyennes théoriques μ_i et μ sont estimées à partir des moyennes observées correspondantes \bar{y}_i et \bar{y} , et l'effet du niveau i du facteur étudié est estimé par $\hat{\alpha}_i = \bar{y}_i - \bar{y}$.

L'intervalle de confiance de la moyenne de la population i est donnée par :

$$\hat{\mu}_i = \bar{y}_i \pm t_{1-\alpha/2} \sqrt{\frac{CM_r}{n_i}}$$

et celui de la moyenne générale, lorsque l'hypothèse nulle est vraie, par :

$$\hat{\mu} = \bar{y} \pm t_{1-\alpha/2} \sqrt{\frac{CM_r}{n.}},$$

où la variable t de Student possède n.-p degrés de liberté. On peut également obtenir une estimation de la différence entre les moyennes de deux populations i et i'.

$$(\bar{y}_i - \bar{y}_{i'}) \pm t_{1-\alpha/2} \sqrt{CM_r \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}.$$

2) Dans le cas du modèle aléatoire, on est souvent amené à estimer les deux composantes de la variance σ^2 et σ_A^2 , la composante additionnelle σ_A^2 est la variance des variables aléatoires A_i .

Pour autant que $CM_a \geq CM_r$, on obtient les estimations non biaisées suivantes :

$$\hat{\sigma}^2 = CM_r \quad \text{et} \quad \hat{\sigma}_A^2 = \frac{(p-1)(CM_a - CM_r)}{b}$$

où $b = n. - \frac{1}{n.} \sum_{i=1}^p n_i^2$.

Si les échantillons ont le même effectif n, la formule donnant $\hat{\sigma}_A^2$ se simplifie :

$$\hat{\sigma}_A^2 = \frac{(CM_a - CM_r)}{n}$$

et l'intervalle de confiance de la moyenne générale s'obtient par la relation :

$$\hat{\mu} = \bar{y} \pm t_{1-\alpha/2} \sqrt{\frac{CM_a}{np}}$$

où la variable t de Student possède $p-1$ degrés de liberté.

Exemple 5.3

Reprenez l'exemple 5.1 et estimez les moyennes des trois classes de mandarines, la moyenne générale, les effets du facteur étudié et la variance commune des trois classes. Cherchez également les intervalles de confiances de ces moyennes et de la variance commune ?

Solution 5.3

Sachant que le modèle est fixe et en supposant que les conditions d'application de l'ANOVA sont vérifiées, les moyennes des trois populations et leurs limites de confiance au niveau 95% sont données par :

$$\hat{\mu}_i = \hat{y}_i \pm 2,052 \sqrt{0,9756/10}$$

ou encore :

$\mu_1 = 10.63 \text{ g/l}$ avec comme limites de confiance **9.99** et **11.27** g/l ;

$\mu_2 = 13.58 \text{ g/l}$ avec comme limites de confiance **12.94** et **14.22** g/l ;

$\mu_3 = 12.03 \text{ g/l}$ avec comme limites de confiance **11.39** et **12.67** g/l.

D'autre part, calculer une estimation de la moyenne générale alors que les moyennes théoriques des trois classes ne sont pas égales ne présente pas d'intérêt.

Quant aux effets des trois traitements, ils sont estimés par :

$$\hat{a}_1 = 10.63 - 12.08 = -1.45, \hat{a}_2 = 1.50 \text{ et } \hat{a}_3 = -0.05.$$

La variance commune des populations est estimée par : $\hat{\sigma}^2 = 0.98$ et les limites de confiance sont données par :

$$\frac{26.342}{43.195} < \sigma^2 < \frac{26.342}{14.573}$$

soit un intervalle de confiance au niveau 95% de **[0.61, 1.81]**.

5.7. CONCLUSION

L'ANOVA 1 reste une méthode simple à réaliser et ses résultats sont faciles à interpréter car elle ne fait intervenir qu'une seule variable indépendante (un seul facteur). Il est utile de distinguer entre le modèle fixe et le modèle aléatoire bien que la statistique du test se calcule de la même manière. Il est à rappeler que dans le cas d'un modèle à effets aléatoires, le chercheur s'intéresse à un nombre très élevés de niveaux mais tire au hasard un sous-ensemble de p niveaux.

Nous avons souligné que certaines conditions d'application doivent au préalable être vérifiées, sachant toutefois que l'ANOVA est une méthode robuste dans la mesure où de légères violations de ces conditions d'application n'entraînent que des effets mineurs. Ceci est surtout vrai pour l'hypothèse de la normalité, plus particulièrement lorsque les effectifs des échantillons sont raisonnablement grands. Si les distributions des populations sont en cloche et si le rapport de la plus grande à la plus petite variance ne dépasse pas trois ou quatre, l'ANOVA reste généralement valable. La répercussion d'une assez grande hétérogénéité des variances est

surtout importante lorsque les effectifs des échantillons sont assez différents. Nous reviendrons sur les solutions à adopter lorsque ces conditions d'application sont loin d'être respectées.

D'autre part, lorsque l'ANOVA 1 ne montre pas de différences significatives entre les moyennes, il est souvent intéressant de calculer la puissance ($1-\beta$) du test (*cf.* chapitre 2). Ce problème sera étudié dans le cours de biométrie (2^{ème} année du cycle ingénieur de l'IAV).

CHAPITRE 6

ANALYSE DE LA VARIANCE A DEUX CRITERES DE CLASSIFICATION

6.1. INTRODUCTION

L'objectif de l'analyse de la variance à deux critères de classification est d'étudier l'influence de deux facteurs de variabilité agissant simultanément sur une variable dépendante quantitative. Chacun des deux facteurs présente deux ou plusieurs niveaux.

Les deux facteurs étudiés peuvent être placés sur le même pied, c'est-à-dire qu'ils ont un rôle symétrique. Dans ce cas, le choix des niveaux se fait indépendamment pour chacun des facteurs et on peut s'intéresser à l'effet global de chacun d'eux et à ce qu'on appelle leur interaction. Ce premier type de modèles porte le nom de modèles croisés. Cependant, les deux facteurs de variation peuvent aussi être subordonnés ou emboîtés l'un à l'autre. Dans ce second cas, on parle de modèles hiérarchisés. Ces deux types de modèles seront présentés et étudiés en détail dans les paragraphes 6.2 à 6.4. Dans tous les cas, on distinguera le modèle fixe lorsque les deux facteurs sont fixes et le modèle aléatoire lorsque les deux facteurs sont aléatoires. Lorsque l'un des deux facteurs est fixe et l'autre est aléatoire, on parlera de modèle mixte.

Dans certains cas, le chercheur peut être intéressé par les deux facteurs étudiés à part égale. On peut citer l'exemple de l'étude des effets de la vitamine B₁₂ et d'un antibiotique, lorsqu'ils sont ajoutés à la nourriture, sur le gain moyen quotidien des animaux d'une espèce donnée, ou encore l'exemple de l'étude des effets de traitements hormonaux et du sexe sur la teneur en calcium du sang d'une espèce de poissons. Dans d'autres cas, le chercheur peut être intéressé par un seul facteur, l'autre facteur, assimilé à ce qu'on appelle des blocs, étant subsidiaire. Les blocs ont le rôle de gagner en puissance en réduisant la variabilité aléatoire (moindre variabilité à l'intérieur des blocs et assez de variabilité entre les blocs). C'est le cas, par exemple, lorsque l'expérimentateur souhaite comparer certaines alimentations (facteur 1) en répartissant au préalable les animaux en trois catégories de poids : légers, moyens et lourds (facteur 2). L'objectif étant de constituer des groupes d'animaux, c'est-à-dire des blocs, dont chacun est aussi homogène que possible. Ce dernier sujet relève du domaine de la planification expérimentale et nous renvoyons le lecteur qui s'y intéresse, entre autres, aux livres de Cochran et Cox, [1955], de Federer [1955] et de Dagnelie [2007].

D'autre part, nous verrons que l'analyse de la variance à deux critères de classification peut être réalisée même lorsqu'on ne dispose que d'une seule observation par échantillon, c'est-à-dire d'une seule valeur par combinaison de niveaux des deux facteurs. Ceci ne peut être le cas de l'ANOVA 1 dont la réalisation exige au moins deux observations par échantillon pour pouvoir obtenir une variation résiduelle qui est indispensable pour réaliser le test de signification.

Nous supposons dans la suite de ce chapitre que le plan est complet, c'est-à-dire que nous disposons d'observations pour toutes les combinaisons des niveaux des deux facteurs, et équilibré, c'est-à-dire que les effectifs par combinaison des niveaux des deux facteurs sont constants et égaux à n. Si ce n'est pas le cas, parce que certaines données sont manquantes pour des raisons quelconques (décès, maladies, casse, etc.), il est recommandé de passer par le modèle linéaire généralisé ou de procéder, si le nombre des données manquantes n'est pas élevé, à des estimations de ces données pour faciliter les calculs.

6.2. RELATIONS ENTRE LES CRITÈRES DE CLASSIFICATION

6.2.1. Définitions

En présence de deux critères de classification, on est amené à préciser non seulement la nature de chacun des deux facteurs, c'est-à-dire si le facteur est fixe ou aléatoire (*cf.* ANOVA1), mais aussi le type de leur relation, c'est-à-dire si les deux facteurs sont croisés ou hiérarchisés.

On parle de facteurs croisés lorsqu'aucun d'eux n'est subordonné à l'autre. Il s'agit d'une sorte de disposition factorielle des deux facteurs : chaque niveau de l'un des facteurs apparaît avec chacun des niveaux de l'autre facteur.

Par contre, on dit que les deux facteurs sont hiérarchisés lorsque l'un d'eux est subordonné à l'autre. Si le facteur b est hiérarchisé au facteur a, alors chaque niveau du facteur a est associé à différents niveaux du facteur b.

1° Supposons, en effet, que l'on s'intéresse à la comparaison du pH de deux horizons du sol d'une forêt (horizons O et A) en creusant au hasard et indépendamment quatre profils (P1, P2, P3 et P4) et en prélevant chaque fois deux échantillons de terre sur lesquels le pH est mesuré. Le schéma du tableau de résultats peut se présenter comme suit, où y_{ijk} désigne le pH du k-ième échantillon prélevé dans l'horizon i du profil j :

Horizons	Profils			
	P1	P2	P3	P4
O	y_{111} y_{112}	y_{121} y_{122}	y_{131} y_{132}	y_{141} y_{142}
A	y_{211} y_{212}	y_{221} y_{222}	y_{231} y_{232}	y_{241} y_{242}

On constate d'abord qu'il y a bien deux critères de classification, l'horizon qui est un facteur fixe avec deux niveaux et le profil qui est un facteur aléatoire avec quatre niveaux, car chaque valeur du pH est classée en fonction d'un horizon pédologique et d'un profil. On remarque ensuite que le pH des deux horizons est mesuré chaque fois dans le même profil et chacun des horizons a la même signification pour tous les profils. On ne peut donc pas intervertir les données des profils pour un horizon et pas pour l'autre. Dans ces conditions, on dit que les critères "horizons" et "profils" sont croisés.

2° Reprenons maintenant le même exemple en supposant mesurer cette fois-ci le pH de deux échantillons de chacun des horizons dans quatre profils différents. Le schéma du tableau de résultats peut se présenter comme suit :

Horizons	Profils	pH
O	P1	y_{111} y_{112}
	P2	y_{121} y_{122}
	P3	y_{131} y_{132}
	P4	y_{141} y_{142}
A	P1	y_{211} y_{212}
	P2	y_{221} y_{222}
	P3	y_{231} y_{232}
	P4	y_{241} y_{242}

Il y a toujours les deux mêmes critères de classification, le facteur "horizons" avec deux niveaux et le facteur "profils" avec quatre niveaux. Le profil P1 où l'on a mesuré le pH de l'horizon O et le profil P1 où l'on a mesuré le pH de l'horizon A sont différents : il n'y a aucune correspondance entre eux. On peut donc intervertir sans problème les données des profils pour un même horizon. Le facteur "profils" est donc dit hiérarchisé au facteur "horizons".

6.2.2. Remarques

Deux remarques méritent d'être signalées à propos des modèles hiérarchisés :

- Le facteur "profils" a quatre niveaux dans le cas des deux types de modèles, on peut cependant voir que l'étude fait intervenir huit profils dans le cas du modèle hiérarchisé ;
- Les profils du modèle hiérarchisé peuvent être numérotés de **P1** à **P4** pour le premier horizon et de **P5** à **P8** pour le second horizon. Cette façon de procéder permet d'éviter des problèmes de confusion lorsque la collecte et l'analyse statistique des données ne se font pas par la même personne.

6.3. MODÈLES CROISÉS

6.3.1. Généralités

Deux facteurs croisés ont un rôle interchangeable. Les niveaux des deux facteurs sont croisés et une ou plusieurs observations pour chacune des combinaisons sont obtenues. Le tableau 6.1 donne la forme générale que prend un tableau de données pour deux facteurs a et b de nombres de niveaux respectifs p et q, lorsqu'on dispose de n observations par échantillon (c'est-à-dire par combinaison). La notation y_{ijk} désigne la k-ième observation de l'échantillon prélevé dans la population relative à la combinaison du i-ième niveau du premier facteur avec le j-ième niveau du second facteur.

Tableau 6.1. Forme générale du tableau de données qu'on traite par le modèle croisé d'ANOVA 2.

	Niveaux	Facteur b			
		b ₁	b ₂	...	b _q
Facteur a	a ₁	y ₁₁₁	y ₁₂₁	...	y _{1q1}
		y ₁₁₂	y ₁₂₂	...	y _{1q2}
	
		y _{11n}	y _{12n}	...	y _{1qn}
	a ₂	y ₂₁₁	y ₂₂₁	...	y _{2q1}
		y ₂₁₂	y ₂₂₂	...	y _{2q2}
	
		y _{21n}	y _{22n}	...	y _{2qn}

	a _p	y _{p11}	y _{p21}	...	y _{pq1}
		y _{p12}	y _{p22}	...	y _{pq2}
	
		y _{p1n}	y _{p2n}	...	y _{pqn}

Les observations par combinaison des niveaux des deux facteurs sont prélevées de manière aléatoire et simple, et leur ordre peut être modifié dans la même combinaison.

D'autre part, nous allons devoir distinguer trois modèles (paragraphe 6.3.3 à 6.3.5) :

- un modèle I, appelé aussi **fixe**, lorsque les deux facteurs sont fixes (paragraphe 6.3.3);
- un modèle II, appelé aussi **aléatoire**, lorsque les deux facteurs sont aléatoires (paragraphe 6.3.4) ;
- un modèle III, appelé aussi **mixte**, lorsque l'un des deux facteurs est fixe, l'autre est aléatoire (paragraphe 6.3.5).

Dans tous les cas, nous ne considérons ici que le cas où le plan est complet et équilibré, c'est-à-dire qu'aucune combinaison de facteurs ne manque et que l'effectif des observations de chaque combinaison des niveaux des deux facteur (échantillon) est constant et égal à n.

6.3.2. Exemple de modèle croisé

Une expérience a été mise en place pour étudier les effets de trois niveaux de température et de trois niveaux de pH sur la croissance d'une bactérie dans une culture. Trois répétitions ont été retenues par combinaison. La croissance est mesurée par la densité optique (tableau 6.2).

Tableau 6.2. Densités optiques mesurant la croissance d'une bactérie dans une culture à trois pH et trois températures.

pH	Températures								
	20 °C			30 °C			40 °C		
5	12.6	10.1	14.5	10.4	15.8	13.2	20.9	17.5	22.4
6	18.8	22.0	19.9	28.0	22.0	27.8	30.0	35.0	30.6
7	40.0	39.4	37.4	43.4	43.1	46.4	54.9	58.0	54.6

Dans cette expérience, on a :

- Deux facteurs étudiés simultanément : le facteur "températures" avec trois niveaux et le facteur "pH" avec également trois niveaux. Soit, au total, neuf populations formées par la combinaison des niveaux des deux facteurs (**p=3 et q=3**) ;
- Un échantillon de trois observations par population est prélevé de manière aléatoire et simple (**n=3**) ;
- La variable dépendante (y) à laquelle on s'intéresse est la densité optique ;
- Le facteur "températures" est placé sur le même pied que le facteur "pH". Aucun des deux facteurs n'est en effet subordonné à l'autre. On dit que le modèle d'analyse de la variance est croisé, on parle aussi d'une analyse de la variance factorielle.

6.3.3. Modèle fixe

On considère deux facteurs contrôlés (fixes) indépendants l'un de l'autre, le premier, noté **a** à **p** niveaux et le second, noté **b** à **q** niveaux. Dans chacune des **p x q** populations qui correspondent aux combinaisons des niveaux des deux facteurs, on prélève un échantillon aléatoire et simple de **n** observations.

6.3.3.1. Modèles observé et théorique

a) Modèle observé

Comme dans le cas de l'ANOVA 1, l'écart entre une observation et la moyenne générale peut être scindé en plusieurs composantes :

$$(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{i...} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i...} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.}) \quad (6.1)$$

(1) (2) (3) (4)

où :

$\bar{y}_{...}$ est la moyenne observée générale ;

$\bar{y}_{i...}$ est la moyenne observée du i-ème niveau du premier facteur ;

$\bar{y}_{.j.}$ est la moyenne observée du j-ème niveau du second facteur ;

$\bar{y}_{ij.}$ est la moyenne de l'échantillon prélevé dans la population relative à la combinaison du i-ème niveau du premier facteur avec le j-ème niveau du second facteur.

Dans l'expression (6.1) :

- La quantité (1) désigne l'écart entre la moyenne générale et la moyenne du i-ième niveau du facteur **a** (elle mesure l'effet du facteur **a** au niveau **i**) ;
- La quantité (2) désigne l'écart entre la moyenne générale et la moyenne du j-ième niveau du facteur **b** (elle mesure l'effet du facteur **b** au niveau **j**) ;
- La quantité (3) mesure l'effet d'interaction des deux facteurs à la combinaison du i-ième niveau du facteur **a** avec le j-ième niveau du facteur **b** ;
- La quantité (4) donne l'écart résiduel.

Les moyennes sont calculées par les relations suivantes :

$$\bar{y}_{...} = \frac{1}{npq} \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n y_{ijk}, \quad \bar{y}_{i..} = \frac{1}{nq} \sum_{j=1}^q \sum_{k=1}^n y_{ijk}, \quad \bar{y}_{.j.} = \frac{1}{np} \sum_{i=1}^p \sum_{k=1}^n y_{ijk} \quad \text{et} \quad \bar{y}_{ij.} = \frac{1}{n} \sum_{k=1}^n y_{ijk}.$$

On constate que l'interaction, donnée par la quantité (3) de l'expression 6.1, est nulle lorsque :

$$\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} = 0$$

ou encore lorsque :

$\bar{y}_{ij.} - \bar{y}_{.j.} = \bar{y}_{i..} - \bar{y}_{...}$ c'est-à-dire si l'effet de i est équivalent pour tous les j, ou que :

$\bar{y}_{ij.} - \bar{y}_{i..} = \bar{y}_{.j.} - \bar{y}_{...}$ c'est-à-dire si l'effet de j est équivalent pour tous les i.

En d'autres termes, il y a absence de l'interaction si l'effet d'un facteur sur la variable dépendante demeure le même pour les différents niveaux du second facteur. L'interaction est une mesure de la non-additivité des effets des deux facteurs.

Examinons, en effet, les graphes de la figure 6.1 où l'on suppose étudier deux facteurs a et b croisés dont les niveaux du premier sont désignés par a1 et a2 et ceux du second par b1 et b2 :

- Sur la figure 6.1 (a), on constate qu'il y a absence des effets principaux des facteurs a et b et il n'y a pas d'interaction ;
- Sur la figure 6.1 (b), on constate qu'il n'y a pas d'effet du facteur a et il n'y a pas d'interaction, mais il y a effet principal du facteur b. L'absence de l'interaction s'explique par la différence de moyennes entre les deux niveaux du facteur b qui reste la même lorsqu'on passe du niveau a1 au niveau a2 du facteur a ;
- Sur la figure 6.1 (c), il y a effets principaux des facteurs a et b et il n'y a pas d'interaction ;
- Sur la figure 6.1 (d), il y a effets principaux des facteurs a et b et il y a interaction. L'interaction s'explique par la différence de moyennes entre les deux niveaux du facteur b qui s'amenuise lorsqu'on passe du niveau a1 au niveau a2 du facteur a ;
- Sur la figure 6.1 (e), on constate qu'il y a effet du facteur a et de l'interaction, mais il n'y a pas effet du facteur b ;
- Sur la figure 6.1 (f), on constate qu'il n'y a pas d'effets ni du facteur a ni du facteur b, mais il y a interaction. Il y a lieu de réaliser deux ANOVA 1 (une pour comparer les moyennes du facteur b pour le niveau a1 du premier facteur et l'autre pour comparer les moyennes du facteur b pour le niveau a2 du premier facteur).

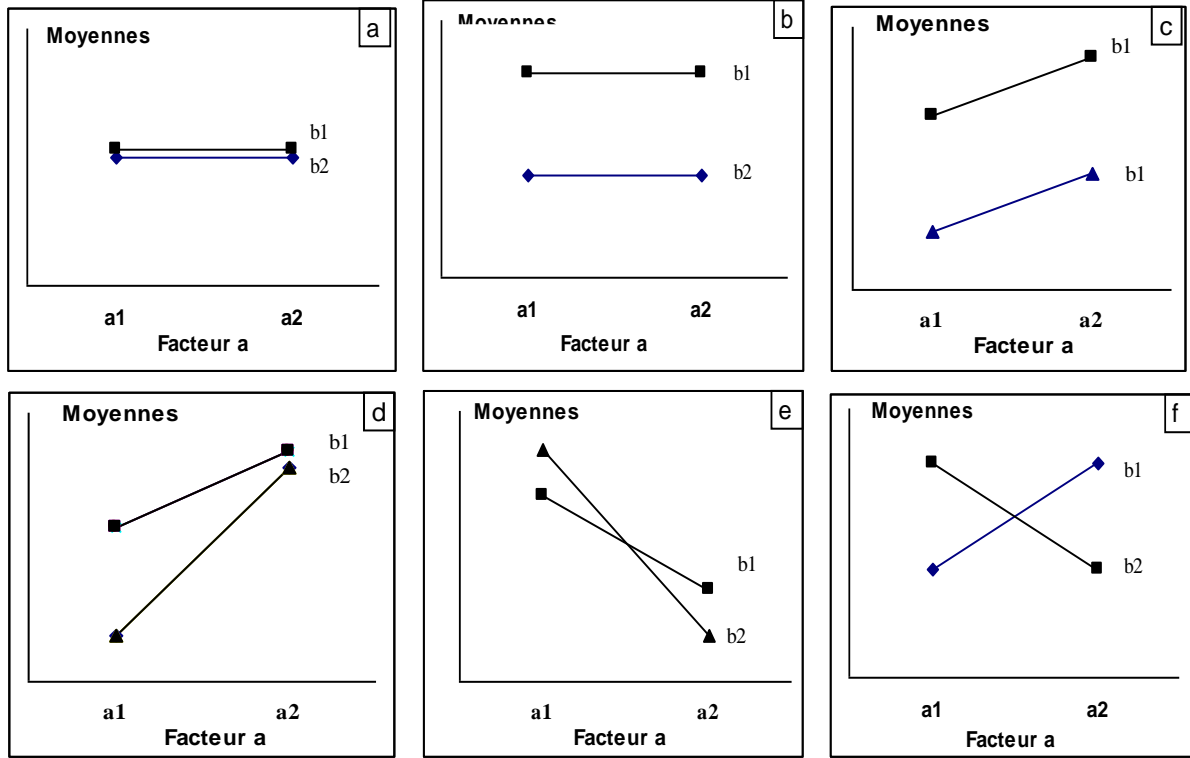


Figure 6.1. Illustration de l'absence de l'interaction (graphes a, b et c) et de la présence de l'interaction (graphes d, e et f) dans le cas de deux facteurs à deux niveaux chacun.

Lorsque l'interaction est significative, une attention particulière doit donc être accordée à l'interprétation des effets des facteurs d'intérêt. En effet, si l'interaction est significative :

- Un effet significatif du facteur b ne signifie pas forcément l'existence de différences entre les niveaux de ce facteur pour chacun des niveaux du facteur a. Cela peut être éventuellement le cas de la figure 6.1 (d) ;
- Un effet non significatif du facteur b peut toutefois avoir de l'effet pour certains niveaux du facteur a. Cela peut être éventuellement le cas de la figure 6.1 (f).

L'équation fondamentale de l'ANOVA 2 s'obtient en élevant au carré les deux membres de la quantité 6.1, puis en sommant pour toutes les observations :

$$\sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 = qn \sum_{i=1}^p (\bar{y}_{i..} - \bar{y}_{...})^2 + pn \sum_{j=1}^q (\bar{y}_{.j.} - \bar{y}_{...})^2 + n \sum_{i=1}^p \sum_{j=1}^q (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2$$

On constate que les sommes des doubles produits ne sont pas considérées puisqu'elles sont nulles dans un tel modèle équilibré, ce qui facilitera l'interprétation ultérieure. Ici, on retrouve en effet les formules des sommes des carrés des écarts. Ces quantités peuvent, avec leurs nombres de degrés de liberté, s'écrire :

$$SCE_t = SCE_a + SCE_b + SCE_{ab} + SCE_r$$

$$pqn - 1 = (p - 1) + (q - 1) + (p - 1)(q - 1) + pq(n - 1) \quad (6.2)$$

où :

- SCE_t représente la somme des carrés des écarts totale (variabilité totale) ;
- SCE_a et SCE_b représentent les sommes factorielles des carrés des écarts relatives à chacun des deux facteurs (variabilités dues au facteur **a** et au facteur **b**) ;
- SCE_{ab} représente la somme des carrés des écarts relative à l'interaction des deux facteurs (variabilité due à l'interaction) ;
- SCE_r représente la somme des carrés des écarts résiduelle (variabilité résiduelle, c'est-à-dire à l'intérieur des échantillons).

Comme en ANOVA 1, on peut calculer les différents carrés moyens en divisant les sommes des carrés des écarts par leurs nombres de degrés de liberté :

$$CM_a = \frac{SCE_a}{p-1} ; CM_b = \frac{SCE_b}{q-1} ; CM_{ab} = \frac{SCE_{ab}}{(p-1)(q-1)} \text{ et } CM_r = \frac{SCE_r}{pq(n-1)} .$$

b) Modèle théorique

Le modèle théorique associé au modèle observé du paragraphe précédent s'écrit :

$$Y_{ijk} = \mu_{..} + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

On peut donc dire que toute valeur observée Y_{ijk} (densité optique de la répétition k correspondant au pH i et à la température j) est la somme de cinq termes :

- une moyenne générale : $\mu_{..}$
- un effet principal du facteur a au i-ième niveau : $a_i = \mu_{i.} - \mu_{..}$
- un effet principal du facteur b au j-ième niveau : $b_j = \mu_{.j} - \mu_{..}$
- un effet de l'interaction entre les niveaux a_i de a et b_j de b : $(ab)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$
- un terme d'erreur aléatoire : $\varepsilon_{ijk} = Y_{ijk} - \mu_{ij}$

Les variables résiduelles ε_{ijk} sont supposées être normales indépendantes de moyenne nulle et de même variance σ^2 . On peut aussi vérifier que :

$$\sum_{i=1}^p a_i = 0, \sum_{j=1}^q b_j = 0 \text{ et } \sum_{i=1}^p (ab)_{ij} = \sum_{j=1}^q (ab)_{ij} = 0 .$$

6.3.3.2. Tests d'hypothèses

1) Trois hypothèses nulles peuvent être testées :

a) la première hypothèse concerne l'absence de l'effet principal du facteur a :

$$H_0 : a_1 = a_2 = \dots = a_p = 0 \text{ contre } H_1 : a_i \neq 0 \text{ pour au moins un } i ;$$

b) la seconde hypothèse concerne l'absence de l'effet principal du facteur b :

$$H'_0 : b_1 = b_2 = \dots = b_q = 0 \text{ contre } H'_1 : b_j \neq 0 \text{ pour au moins un } j ;$$

c) la troisième hypothèse concerne l'absence d'interaction :

$$H''_0 : (ab)_{11} = (ab)_{12} = \dots = (ab)_{pq} = 0$$

contre

$$H''_1 : (ab)_{ij} \neq 0 \text{ pour au moins une combinaison (i,j).}$$

2) Dans le cas où il y a présence d'effet de l'un ou des deux facteurs et/ou de l'interaction, on est souvent amené à fournir une mesure globale de cet effet :

- pour le facteur a, cet effet est mesuré par : $\delta_a^2 = \frac{1}{p-1} \sum_{i=1}^p a_i^2$;
- pour le facteur b, cet effet est mesuré par : $\delta_b^2 = \frac{1}{q-1} \sum_{j=1}^q b_j^2$;
- pour l'interaction, cet effet est mesuré par : $\delta_{ab}^2 = \frac{1}{(p-1)(q-1)} \sum_{i=1}^p \sum_{j=1}^q (ab)_{ij}^2$.

Les hypothèses nulles H_0 , H'_0 et H''_0 peuvent aussi s'écrire :

$$H_0 : \delta_a^2 = 0 \text{ ou } \mu_{.1} = \mu_{.2} = \dots = \mu_{.p}.$$

$$H'_0 : \delta_b^2 = 0 \text{ ou } \mu_{.1} = \mu_{.2} = \dots = \mu_{.q}$$

$$H''_0 : \delta_{ab}^2 = 0$$

3) La réalisation des tests de signification se base sur le calcul des espérances mathématiques et la connaissance des distributions des sommes des carrées des écarts ou des carrés moyens. Lorsque les conditions d'application du modèle sont vérifiées, on obtient les valeurs attendues des carrés moyens suivantes :

- $E(CM_a) = \sigma^2 + qn\delta_a^2$;
- $E(CM_b) = \sigma^2 + pn\delta_b^2$;
- $E(CM_{ab}) = \sigma^2 + n\delta_{ab}^2$;
- $E(CM_r) = \sigma^2$ (pour $n > 1$) ;

et les distributions des différentes sommes des carrés des écarts suivantes :

- sous H_0 , $\frac{SCE_a}{\sigma^2}$ suit une distribution χ^2 à $p-1$ degrés de liberté ;
- sous H'_0 , $\frac{SCE_b}{\sigma^2}$ suit une distribution χ^2 à $q-1$ degrés de liberté ;
- sous H''_0 , $\frac{SCE_{ab}}{\sigma^2}$ suit une distribution χ^2 à $(p-1)(q-1)$ degrés de liberté ;
- $\frac{SCE_r}{\sigma^2}$ suit une distribution χ^2 à $pq(n-1)$ degrés de liberté.

Ces variables aléatoires χ^2 sont indépendantes.

Les espérances mathématiques et les distributions obtenues montrent que :

- Lorsque H_0 est vraie, $E(CM_a) = E(CM_r)$; le test de signification de l'effet principal du facteur **a** se fait donc en comparant le CM_a au CM_r . On a la quantité :

$$F_a = \frac{SCE_a / (p-1)}{SCE_r / (pq(n-1))} = \frac{CM_a}{CM_r}$$

qui est, sous H_0 , une valeur observée d'une variable F de Fisher-Snedecor à $k_1=p-1$ et $k_2=pq(n-1)$ degrés de liberté. On rejette H_0 lorsque $F_a \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=p-1$ et $k_2=pq(n-1)$ degrés de liberté (ou que p-value est inférieure au risque de première espèce (**p-value** < α)) ;

- Lorsque H'_0 est vraie, $E(CM_b) = E(CM_r)$; le test de signification de l'effet principal du facteur b se fait donc en calculant la quantité :

$$F_b = \frac{CM_b}{CM_r}$$

et en rejetant H'_0 lorsque $F_b \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=q-1$ et $k_2=pq(n-1)$ degrés de liberté (**p-value** < α) ;

- Lorsque H''_0 est vraie, $E(CM_{ab}) = E(CM_r)$; le test d'absence d'interaction se fait donc en calculant la quantité :

$$F_{ab} = \frac{CM_{ab}}{CM_r}$$

et en rejetant H''_0 lorsque $F_{ab} \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=(p-1)(q-1)$ et $k_2=pq(n-1)$ degrés de liberté (**p-value** < α).

6.3.4. Modèle aléatoire

Les deux facteurs sont ici aléatoires, c'est-à-dire qu'on a une infinité de niveaux aussi bien pour le facteur **a** que pour le facteur **b**. Un échantillonnage à deux degrés est effectué. On prélève d'abord un échantillon au hasard constitué de **pq** populations puis, au sein de chacune de celles-ci et toujours au hasard, un échantillon de **n** individus est sélectionné.

6.3.4.1. Modèles observé et théorique

Le modèle observé et l'équation de l'analyse de la variance à deux critères de classification s'écrivent de la même manière que dans le cas du modèle fixe (cf. expressions (6.1) et (6.2)).

L'écriture du modèle théorique tient compte du caractère aléatoire des moyennes des différents niveaux des deux critères de classification ($M_{i.}$ et $M_{.j}$) et de la moyenne de chaque combinaison des niveaux des deux facteurs (M_{ij}). Le modèle théorique s'écrit donc :

$$Y_{ijk} = \mu_{..} + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}$$

dans lequel les effets principaux sont donnés par $A_i = M_{i.} - \mu_{..}$ et $B_j = M_{.j} - \mu_{..}$ et le terme d'interaction par $(AB)_{ij} = M_{ij} - M_{i.} - M_{.j} + \mu_{..}$.

On suppose que les quantités A_i , B_j , $(AB)_{ij}$ et ε_{ijk} sont des variables aléatoires normales indépendantes de moyenne nulle et de variances respectives σ_A^2 , σ_B^2 , σ_{AB}^2 et σ^2 . Ces variables sont également indépendantes entre elles.

6.3.4.2. Tests d'hypothèses

1) Comme dans le cas du modèle fixe, on peut tester les effets principaux des facteurs **a** et **b** et l'interaction. L'importance de chacune de ces trois sources de variation sur la variable dépendante est mesurée par la variance correspondante.

Les trois hypothèses nulles qui peuvent être testées sont :

- absence de l'effet du facteur a : $H_0 : \sigma_A^2 = 0$
- absence de l'effet du facteur b : $H'_0 : \sigma_B^2 = 0$
- absence d'interaction : $H''_0 : \sigma_{AB}^2 = 0$.

2) De même, la réalisation des tests de signification tient compte des formules des espérances mathématiques des sommes des carrés des écarts, ou des carrés moyens, et de la connaissance de leurs distributions. Lorsque les conditions d'application du modèle sont vérifiées, on obtient les espérances mathématiques des carrés moyens suivantes :

- $E(CM_a) = \sigma^2 + n\sigma_{AB}^2 + qn\sigma_A^2$;
- $E(CM_b) = \sigma^2 + n\sigma_{AB}^2 + pn\sigma_B^2$;
- $E(CM_{ab}) = \sigma^2 + n\sigma_{AB}^2$;
- $E(CM_r) = \sigma^2$ ($n > 1$) ;

et les distributions des sommes des carrés des écarts suivantes :

- $\frac{SCE_a}{E(CM_a)}$ suit une distribution χ^2 à p-1 degrés de liberté ;
- $\frac{SCE_b}{E(CM_b)}$ suit une distribution χ^2 à q-1 degrés de liberté ;
- $\frac{SCE_{ab}}{E(CM_{ab})}$ suit une distribution χ^2 à (p-1)(q-1) degrés de liberté ;
- $\frac{SCE_r}{E(CM_r)}$ suit une distribution χ^2 à pq(n-1) degrés de liberté ($n > 1$).

Les différentes variables aléatoires χ^2 sont indépendantes.

Les espérances mathématiques et les distributions obtenues montrent que :

- Lorsque H_0 est vraie, $E(CM_a) = E(CM_{ab})$. Le test consiste donc à calculer la quantité :

$$F_a = \frac{CM_a}{CM_{ab}}$$

qui est, sous H_0 , une valeur observée d'une variable F de Fisher-Snedecor, et à rejeter H_0 lorsque $F_a \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1 = p-1$ et $k_2 = (p-1)(q-1)$ degrés de liberté (c'est-à-dire lorsque **p-value** < α);

- Lorsque H_0' est vraie, $E(CM_b) = E(CM_{ab})$. Le test consiste donc à calculer la quantité :

$$F_b = \frac{CM_b}{CM_{ab}}$$

et à rejeter H_0' lorsque $F_b \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=q-1$ et $k_2=(p-1)(q-1)$ degrés de liberté (**p-value** $< \alpha$) ;

- Lorsque H_0'' est vraie, $E(CM_{ab}) = E(CM_r)$. Le test consiste donc à calculer la quantité :

$$F_{ab} = \frac{CM_{ab}}{CM_r}$$

et à rejeter H_0'' lorsque $F_{ab} \geq F_{1-\alpha}$, où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=(p-1)(q-1)$ et $k_2=pq(n-1)$ degrés de liberté (**p-value** $< \alpha$).

6.3.5. Modèle mixte

Dans ce cas, l'un des deux facteurs est fixe, l'autre est aléatoire. Si l'on suppose que c'est le facteur **a** qui est fixe, cela revient à considérer un nombre **p** fini de niveaux pour ce facteur et un nombre infini de niveaux pour le facteur **b**. Un échantillonnage à deux degrés est effectué. Le choix des unités du premier degré concerne le facteur aléatoire, en prélevant au hasard **q** niveaux par exemple. Un échantillon aléatoire et simple d'effectif **n** est ensuite prélevé dans chacune des **pq** populations.

6.3.5.1. Modèles observé et théorique

Le modèle observé reste le même que dans le cas des modèles fixe et aléatoire, tandis que le modèle théorique s'écrit :

$$Y_{ijk} = \mu_{..} + a_i + B_j + (aB)_{ij} + \varepsilon_{ijk}$$

dans lequel les effets du facteur fixe (a_i) sont des constantes alors que les effets du facteur aléatoire (B_j) et de l'interaction ($(aB)_{ij}$) sont des variables aléatoires. En effet, les moyennes des différents niveaux du facteur aléatoire ($M_{.j}$) ainsi que les moyennes des différentes populations (M_{ij}) sont des variables aléatoires. L'interaction est aléatoire puisque l'un des facteurs qui intervient dans son expression est aléatoire.

Dans ces conditions, on a :

$$a_i = \mu_{i.} - \mu_{..} \text{ avec } \sum_{i=1}^p a_i = 0, \quad B_j = M_{.j} - \mu_{..} \text{ et } (aB)_{ij} = M_{ij} - \mu_{i.} - M_{.j} + \mu_{..}$$

On suppose que les variables B_j et $(aB)_{ij}$ sont normales indépendantes de moyenne nulle et de variance respectives σ_B^2 et σ_{aB}^2 , avec $\sum_{i=1}^p (aB)_{ij} = 0$ pour tout j , et indépendantes entre elles. De même, on suppose que les variables ε_{ijk} sont normales indépendantes de moyenne nulle et de variance σ^2 et indépendantes des variables B_j et $(aB)_{ij}$.

6.3.5.2. Tests d'hypothèses

1) Trois hypothèses nulles peuvent être testées :

- absence de l'action du facteur **a** : $H_0 : \mu_1 = \mu_2 = \dots = \mu_p.$
- absence de l'action du facteur **b** : $H'_0 : \sigma_B^2 = 0$
- absence d'interaction : $H''_0 : \sigma_{aB}^2 = 0$

2) La réalisation de ces tests de signification repose sur le calcul des espérances mathématiques des sommes des carrées des écarts ou des carrés moyens et sur la connaissance de leurs distributions théoriques. Lorsque les conditions d'application du modèle sont vérifiées et que l'effectif n est supérieur à un, on obtient les espérances mathématiques des carrés moyens suivantes :

- $E(CM_a) = \sigma^2 + n\sigma_{aB}^2 + qn\delta_a^2$
- $E(CM_b) = \sigma^2 + pn\sigma_B^2$
- $E(CM_{ab}) = \sigma^2 + n\sigma_{aB}^2$
- $E(CM_r) = \sigma^2$

et les distributions des sommes des carrées des écarts suivantes :

- sous H_0 , $\frac{SCE_a}{\sigma^2 + n\sigma_{aB}^2}$ suit une distribution χ^2 à p-1 degrés de liberté
- $\frac{SCE_b}{E(CM_b)}$ suit une distribution χ^2 à q-1 degrés de liberté
- $\frac{SCE_{ab}}{E(CM_{ab})}$ suit une distribution χ^2 à (p-1)(q-1) degrés de liberté
- $\frac{SCE_r}{E(CM_r)}$ suit une distribution χ^2 à pq(n-1) degrés de liberté.

En raisonnant de la même manière que dans le cas des modèles fixes et aléatoire, il ressort d'après les espérances mathématiques et les distributions théoriques obtenues que :

- Le test de l'hypothèse H_0 repose sur le calcul de la quantité :

$$F_a = \frac{CM_a}{CM_{ab}}.$$

H_0 doit être rejetée lorsque $F_a \geq F_{1-\alpha}$ où $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=p-1$ et $k_2=(p-1)(q-1)$ degrés de liberté (ou lorsque **p-value** $< \alpha$) ;

- Le test de l'hypothèse H'_0 repose sur le calcul de la quantité :

$$F_b = \frac{CM_b}{CM_r}.$$

H'_0 doit être rejetée lorsque $F_b \geq F_{1-\alpha}$ où $F_{1-\alpha}$ est obtenue sur la table statistique avec $k_1=q-1$ et $k_2= pq(n-1)$ degrés de liberté (ou lorsque **p-value** $< \alpha$) ;

- Le test de l'hypothèse H_0'' repose sur le calcul de la quantité :

$$F_{ab} = \frac{CM_{ab}}{CM_r}.$$

H_0 doit être rejetée lorsque $F_{ab} \geq F_{1-\alpha}$ où $F_{1-\alpha}$ est obtenue sur la table statistique avec $k_1=(p-1)(q-1)$ et $k_2=pq(n-1)$ degrés de liberté (ou lorsque **p-value** $< \alpha$).

6.3.6. Réalisation de l'analyse de la variance

6.3.6.1. Formules de calculs

1) Les différentes sommes et le facteur de correction sont calculés par les formules suivantes :

- somme pour le $i^{\text{ème}}$ niveau du facteur **a** : $Y_{i..} = \sum_{j=1}^q \sum_{k=1}^n y_{ijk} = \sum_{j=1}^q Y_{ij.}$
- somme pour le $j^{\text{ème}}$ niveau du facteur **b** : $Y_{.j.} = \sum_{i=1}^p \sum_{k=1}^n y_{ijk} = \sum_{i=1}^p Y_{ij.}$
- somme par échantillon : $Y_{ij.} = \sum_{k=1}^n y_{ijk}$
- somme générale : $Y_{...} = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n y_{ijk}$
- somme des carrés : $T = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n y_{ijk}^2$
- facteur de correction : $C = \frac{Y_{...}^2}{npq}$

2) Les sommes des carrées des écarts sont données par :

- somme des carrées des écarts par échantillon : $SCE_{ij} = \sum_{k=1}^n y_{ijk}^2 - Y_{ij.}^2 / n$
- somme des carrées des écarts liée au facteur a : $SCE_a = \frac{1}{qn} \sum_{i=1}^p Y_{i..}^2 - C$
- somme des carrées des écarts liée au facteur b : $SCE_b = \frac{1}{pn} \sum_{j=1}^q Y_{.j.}^2 - C$
- somme des carrées des écarts résiduelle : $SCE_r = \sum_{i=1}^p \sum_{j=1}^q SCE_{ij}$
- somme des carrées des écarts totale: $SCE_t = T - C$
- somme des carrés des écarts de l'interaction : $SCE_{ab} = SCE_t - SCE_a - SCE_b - SCE_r$
qui peut être aussi calculée par : $SCE_{ab} = n \sum_{i=1}^p \sum_{j=1}^q (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$

6.3.6.2. Tableau d'analyse de la variance

Le tableau 6.3 présente l'esquisse du tableau d'analyse de la variance à deux critères de classification lorsque le modèle est croisé. Les trois dernières colonnes rappellent la manière de réaliser les différentes comparaisons selon le modèle d'ANOVA2.

Tableau 6.3. Esquisse du tableau d'analyse de la variance à deux critères de classification dans le cas du modèle croisé.

Sources de variation	Degrés de liberté	SCE	CM	F _{obs}	F _{obs}	F _{obs}
Facteur a	p-1	SCE _a	CM _a			
Facteur b	q-1	SCE _b	CM _b			
Interaction	(p-1)(q-1)	SCE _{ab}	CM _{ab}			
Variation résiduelle	pq(n-1)	SCE _r	CM _r			
Variation totale	pqn-1	SCE_t		(¹)	(²)	(³)

(¹) : les deux facteurs sont fixes (modèle fixe) ;

(²) : les deux facteurs sont aléatoires (modèle aléatoire) ;

(³) : le facteur **a** est fixe et le facteur **b** est aléatoire (modèle mixte).

6.3.6.3. Remarques

1) Si l'interaction n'est pas significative, les modèles d'ANOVA sont dits additifs. Dans ce cas, on peut tester les effets principaux des facteurs par rapport à la quantité :

$$CM_{com} = \frac{SCE_{ab} + SCE_r}{(p-1)(q-1) + pq(n-1)}.$$

qui constitue une estimation commune de la variance σ^2 . Dans ce cas, on a :

$$F_a = \frac{CM_a}{CM_{com}} \quad \text{et} \quad F_b = \frac{CM_b}{CM_{com}}$$

Chacune de ces deux valeurs observées est comparée à la valeur théorique $F_{1-\alpha}$. Celle-ci possède $k_1=p-1$ et $k_2=(p-1)(q-1)+pq(n-1)$ degrés de liberté lorsqu'on teste H_0 et $k_1=q-1$ et $k_2=(p-1)(q-1)+pq(n-1)$ degrés de liberté lorsqu'on teste H_0' .

2) Lorsque l'interaction est significative, il est de coutume pour les modèles fixes de tester les effets simples pour étudier l'effet d'un facteur à chaque niveau de l'autre facteur. Il s'agit de séparer les données par niveau de l'un des facteurs et de réaliser plusieurs ANOVA 1 pour tester chaque fois l'effet de l'autre facteur. Cette approche a cependant l'inconvénient de considérer les données comme si elles concernaient des études séparées.

Dans ce cas, il convient cependant de noter que certains auteurs adoptent quelques modifications pour améliorer les tests de signification des effets par les ANOVA 1 :

- Le carré moyen utilisé pour calculer les valeurs observées de la variable F de Fisher-Snedecor des différentes ANOVA 1 est celui de l'ANOVA 2 et non pas celui qui provient des ANOVA 1. Cette façon de procéder permet d'utiliser un carré moyen basé sur un plus grand nombre de degrés de liberté ;
- Si l'on souhaite obtenir un risque d'erreur de première espèce de l'ordre de α pour l'ensemble des tests, il est souvent recommandé de considérer, pour chaque test, un risque d'erreur de première espèce égal à $\alpha'=\alpha/k$, où k est le nombre de tests à réaliser (cf. exemple 6.1).

3) Lorsqu'on dispose d'une seule observation par combinaison des niveaux des deux critères de classification, c'est-à-dire lorsque n est égal à 1, il n'y a pas de résidus. L'équation de l'analyse de la variance devient :

$$SCE_t = SCE_a + SCE_b + SCE_{ab}.$$

Les degrés de liberté qui correspondent à ces sommes des carrées des écarts sont alors liés par la relation :

$$pq - 1 = (p - 1) + (q - 1) + (p - 1)(q - 1).$$

Il devient donc impossible, d'après le tableau d'analyse de la variance (tableau 6.3), de pouvoir faire certaines comparaisons. Ce problème se pose pour les deux facteurs du modèle fixe et pour le facteur aléatoire du modèle mixte.

Dans ce cas, on peut commencer par tester l'interaction en utilisant le test d'additivité de Tukey (Tukey, 1949). Si celle-ci n'est pas significative, on peut tester les effets principaux des facteurs fixes par rapport à cette interaction.

Le test de Tukey revient à calculer d'abord (Snedecor et Cochran, 1968; Dagnelie, 1980):

$$SCE_{add} = \frac{\left[\sum_{i=1}^p \sum_{j=1}^q y_{ij} (\bar{y}_{i.} - \bar{y}_{..}) (\bar{y}_{.j} - \bar{y}_{..}) \right]^2}{\sum_{i=1}^p (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_{j=1}^q (\bar{y}_{.j} - \bar{y}_{..})^2}$$

et à rejeter l'hypothèse d'absence d'interaction lorsque la valeur observée :

$$F_{obs} = \frac{SCE_{add}}{(SCE_{ab} - SCE_{add}) / k},$$

où $k = (p-1)(q-1)-1$, est supérieure ou égale à $F_{1-\alpha}$ avec $k_1=1$ et $k_2=(p-1)(q-1)-1$ degrés de liberté.

4) En ANOVA 2, on peut être intéressé par des estimations des paramètres. Les effets principaux peuvent être estimés par les formules qui ont été données tout au long de ce chapitre. De même, lorsque le modèle est aléatoire, on peut obtenir des estimations non biaisées des variances des effets principaux :

$$\begin{aligned} \hat{\sigma}_A^2 &= (CM_a - CM_{ab}) / qn \\ \hat{\sigma}_B^2 &= (CM_b - CM_{ab}) / pn \\ \hat{\sigma}_{AB}^2 &= (CM_{ab} - CM_r) / n \quad \text{si } n > 1 \\ \hat{\sigma}^2 &= CM_r \quad \text{si } n > 1. \end{aligned}$$

5) Les tailles des effets associées à chacun des facteurs et à l'interaction peuvent être évaluées en utilisant, comme on l'avait fait pour l'ANOVA 1, différents paramètres. Parmi ceux-ci, on peut citer le η^2 qui correspond au rapport de la somme des carrées des écarts de la source de variation en question à la somme des carrées des écarts totale :

$$\eta_a^2 = \frac{SCE_a}{SCE_t} ; \quad \eta_b^2 = \frac{SCE_b}{SCE_t} \quad \text{et} \quad \eta_{ab}^2 = \frac{SCE_{ab}}{SCE_t}.$$

Eta-carré exprime la part de la variabilité de la variable dépendante qui est imputable à la variable indépendante (facteur).

6) Toute la théorie que nous avons reprise dans les paragraphes précédents concerne des échantillons d'effectifs égaux. Lorsque ces effectifs sont fortement inégaux, on peut utiliser plusieurs approches qui tiennent compte de cette inégalité. Plusieurs logiciels statistiques peuvent être utilisés pour faire correctement les analyses de la variance.

Toutefois, si l'on perd une ou quelques observations au cours d'une expérience pour laquelle on a prévu des échantillons d'une même taille, on pourrait obtenir des estimations pour ces données manquantes pour faciliter les calculs. Deux situations peuvent se présenter :

- 1) Si l'on dispose de plusieurs observations par échantillons ($n > 1$), la valeur manquante peut être estimée par la moyenne de cet échantillon. Si l'on estime k valeurs, on devra diminuer les nombres de degrés de liberté de la variation résiduelle et du total de k unités.

Exemple : Reprenons les données des densités optiques mesurant la croissance d'une bactérie dans une culture à trois pH et trois températures et supposons que le tube à essai correspondant à la deuxième répétition concernant le pH 7 et la température 30 °C a été mal manipulé et n'a pas permis de mesurer la densité optique.

pH	Températures								
	20 °C			30 °C			40 °C		
5	12.6	10.1	14.5	10.4	15.8	13.2	20.9	17.5	22.4
6	18.8	22.0	19.9	28.0	22.0	27.8	30.0	35.0	30.6
7	40.0	39.4	37.4	43.4		46.4	54.9	58.0	54.6

Puisqu'il s'agit du même échantillon, on peut remplacer la donnée manquante par la moyenne de l'échantillon, soit 44.9. Il s'agit d'une donnée fictive qui a été calculée pour faciliter les calculs. Le tableau d'ANOVA 2 prendra les degrés de liberté suivants :

Sources de variation	Degrés de liberté	SCE	CM	F _{obs}
pH	2	SCE _a	CM _a	
Température	2	SCE _b	CM _b	
pH x Température	4	SCE _{ab}	CM _{ab}	
Variation résiduelle	17	SCE _r	CM _r	
Variation totale	25	SCE_t		

- 2) Si l'on dispose d'une seule observation par échantillon ($n=1$), on peut utiliser l'une des nombreuses méthodes disponibles dans la littérature tenant compte notamment du plan expérimental adopté.

L'une des plus connues et qu'on a l'habitude d'utiliser dans le cas des expériences en blocs aléatoires complets, consiste à minimiser le carré moyen de l'interaction. Elle fournit comme estimation de la valeur manquante x_{ij} (Cochran et Cox, 1955; Dagnelie, 1980) :

$$\hat{y}_{ij} = \frac{pY'_{i.} + qY'_{.j} - Y'_{..}}{(p-1)(q-1)},$$

où $Y'_{i.}$ et $Y'_{.j}$ sont les sommes marginales sans valeurs estimées et $Y'_{..}$ est la somme totale.

S'il y a plus d'une valeur manquante, la formule peut être utilisée de manière itérative. Dans le cas de deux valeurs manquantes par exemple, on commence par donner une valeur arbitraire à la première valeur manquante, puis on estime la seconde par la formule. Ensuite, on utilise cette nouvelle estimation pour obtenir une nouvelle estimation de la première valeur manquante. On poursuit le calcul jusqu'à ce que des estimations successives donnent des valeurs très proches pour la même valeur manquante.

Si l'on estime k valeurs, on ne devra oublier de diminuer les nombres de degrés de liberté de l'interaction et du total de k unités.

Exemple : Une expérimentation avait pour but de comparer l'ingestion de quatre aliments de compositions différentes par des moutons. L'expérimentateur dispose de 12 animaux qu'il répartit en trois classes selon leurs poids et a attribué les quatre aliments au hasard aux différents moutons, à l'intérieur de chacun des trois lots. Les résultats sont exprimés en grammes d'aliments ingérés par kilogramme de poids vif. La donnée $y_{4.2}$ du lot des poids moyens recevant l'aliment 4 n'a pas pu être récupérée.

Classe de poids des moutons	Alimentations			
	1	2	3	4
Lot des animaux lourds	40	37.7	52.6	54.6
Lot des animaux moyens	38.8	40.1	54.6	.
Lot des animaux légers	39.3	32.6	53.8	50.8

La répartition en lots permet de contrôler l'hétérogénéité. Les animaux d'un lot sont les plus « homogènes » possibles et les lots sont les plus différents possibles les uns des autres, tout en souhaitant éviter des différences trop importantes pouvant conduire à une interaction entre lots et alimentations. Ici, les lots jouent le rôle de Blocs qui constitue un facteur aléatoire (*cf. cours de biométrie de l'IAV Hassan II*).

La valeur manquante est estimée à 57.2 grammes et le tableau d'ANOVA 2 prendra les degrés de liberté suivants :

Sources de variation	Degrés de liberté	SCE	CM	F _{obs}
Alimentations	3	SCE _a	CM _a	
Lots	2	SCE _b	CM _b	
Alimentations x Lots	5	SCE _{ab}	CM _{ab}	
Variation totale	10	SCE_t		

Exemple 6.1

Reprenez les données du tableau 6.2 et analysez aussi complètement que possible les résultats obtenus ?

Solution 6.1

a) Méthode : ANOVA 2, modèle croisé fixe

- ANOVA 2 : il y a un facteur "température" et un facteur "pH" ($p=3$, $q=3$ et $n=3$)
- Modèle fixe : les deux facteurs sont fixes
- Modèle théorique : $Y_{ijk} = \mu_{..} + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$

b) Conditions d'application :

- Les 9 ($=3 \times 3$) échantillons sont aléatoires, simples et indépendants
- Les distributions des 9 populations sont normales et possèdent la même variance σ^2 .

c) Hypothèses nulles :

H_0'' : absence d'interaction température-pH contre H_1'' : présence d'interaction

H_0 : $\mu_{1.} = \mu_{2.} = \mu_{3.}$ contre H_1 : $\mu_{k.} \neq \mu_{l.}$ pour au moins 2 niveaux de températures

H_0 : $\mu_{.1} = \mu_{.2} = \mu_{.3}$ contre H_1 : $\mu_{.k} \neq \mu_{.l}$ pour au moins 2 niveaux de pH

d) Calcul :

$$Y_{..} = 788.70 ; C = (788.70^2/27) = 23038.80 ; T = (12.6^2 + \dots + 54.6^2) = 28387.95$$

$$SCE_t = T - C = 5349.15$$

$$SCE_1 = 9.74 ; SCE_2 = 14.59 ; \dots ; SCE_9 = 7.09 \text{ et } SCE_r = SCE_1 + SCE_2 + \dots + SCE_9 = 97.81$$

$$SCE_{\text{températures}} = (214.7^2 + 250.1^2 + 323.9^2)/9 - 23038.80 = 689.79$$

$$SCE_{\text{pH}} = (137.4^2 + 234.1^2 + 417.2^2)/9 - 23038.80 = 4487.58$$

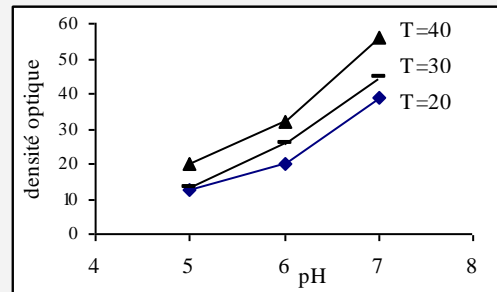
$$SCE_{\text{pH.Température}} = SCE_t + SCE_{\text{température}} - SCE_{\text{pH}} - SCE_r = 73.98$$

e) Tableau d'analyse de la variance :

Source de variation	d.l	SCE	CM	Fobs	p-valeur
Températures	2	689.79	344.89	63.47	0.000
pH	2	4487.58	2243.79	412.94	0.000
Températures x pH	4	73.98	18.49	3.40	0.031
Résidus	18	97.81	5.43		
Total	26	5349.15	-	-	-

Commençons par tester l'interaction : on constate que celle-ci est significative au niveau $\alpha=0.05$ car $F_{0.95}=2.928$ ($p\text{-value} = 0.031$). Cela veut dire que les densités moyennes obtenues avec les trois températures ne sont pas les mêmes aux trois niveaux du pH. L'effet de la température est donc, comme le montre la figure suivante, fonction du niveau choisi pour le pH.

Si nous testons les effets principaux des deux facteurs, nous constatons qu'ils sont très hautement significatifs. Cela veut dire que les densités optiques moyennes ne sont pas toutes égales pour les trois températures, et elles ne sont pas toutes égales pour les trois pH.



Puisque l'interaction est significative, le chercheur peut juger utile de réaliser trois ANOVA 1, une pour chaque niveau du pH, pour comparer les effets simples du facteur "températures". On obtient les résultats suivants :

H_0 : il n'y a pas effet température contre H_1 : il y a effet température

Pour pH=5, l'ANOVA1 donne $F_{obs}^{(1)} = \frac{56.65}{5.434} = 10.42$.

Cette valeur est supérieure à $F_{0.983} = 5.18$ avec 2 et 18 degrés de liberté, ce qui indique un effet significatif de la température sur la croissance des bactéries au pH=5 ($p\text{-value}=0.00098$).

Nous avons utilisé $\alpha'=0.05/3=0.0167$ au lieu de $\alpha=0.05$ car nous avons 3 comparaisons. D'après le graphique, on constate que c'est la température $T=40$ qui donne une densité moyenne plus élevée par rapport aux deux autres températures qui conduisent à la même densité moyenne pour ce pH (ce problème sera traité par les comparaisons multiples de moyennes du chapitre 7) ;

Pour les deux autres pH, on a :

$$F_{obs}^{(2)} = \frac{101.51}{5.434} = 18.68 \text{ pour pH=6 et } F_{obs}^{(3)} = \frac{223.71}{5.434} = 41.17 \text{ pour pH=7,}$$

ce qui conduit, pour chaque pH, au rejet de l'hypothèse d'égalité des densités moyennes obtenues avec les trois températures.

Exemple 6.2

Reprenez l'Exemple 6.1 et donnez une mesure de la taille de l'effet associé à chaque facteur et à l'interaction ?

Solution 6.2

$$\eta_{temp}^2 = \frac{689.79}{5349.15} ; \eta_{pH}^2 = \frac{4487.58}{5349.15} \text{ et } \eta_{temp-pH}^2 = \frac{73.98}{5349.15}.$$

Il s'avère que 13% de la variabilité totale de la croissance des bactéries sont expliqués par la température, 84% par le pH et 1.4% par l'interaction entre la température et le pH.

6.4. MODÈLES HIÉRARCHISÉS

6.4.1. Introduction

Comme nous l'avons vu au paragraphe 6.2, on dit qu'un modèle d'analyse de la variance à deux critères de classification est hiérarchisé lorsque l'un des deux facteurs est subordonné, au lieu d'être croisé, à l'autre facteur. C'est le cas par exemple lorsqu'un aviculteur souhaite comparer les poids des œufs pendus par trois races de poules pondeuses en choisissant, de manière aléatoire et simple, six poules pondeuses par race et pour chacune d'elles il pèse quatre œufs qu'il choisit également au hasard. Dans ce cas, il n'y a pas de correspondances entre les différentes poules des deux races. On dit que le facteur "poules" est subordonné au facteur "races".

D'une manière générale, on considère q niveaux du facteur subordonné (facteur **b**) à l'intérieur de chacun des p niveaux du facteur principal (facteur **a**). Puis n observations à l'intérieur de chacun des q niveaux du facteur subordonné. L'extension à plus de deux facteurs reste facile. Dans le cas de trois facteurs par exemple, on peut considérer un troisième facteur **c** qui est hiérarchisé dans les niveaux du facteur **b** et que ce dernier est à son tour hiérarchisé dans les niveaux du facteur **a**. Le problème devient toutefois moins facile, en présence de plusieurs facteurs, lorsque certains d'entre eux sont hiérarchisés et d'autres sont croisés. On parle dans ce dernier cas de modèles partiellement hiérarchisés.

Dans un modèle d'ANOVA 2 hiérarchisé, le critère subordonné est généralement aléatoire et le critère principal peut être fixe (contrôlé) ou aléatoire. On parle donc généralement de modèles hiérarchisés mixte (paragraphe 6.4.3) ou aléatoire (paragraphe 6.4.4) selon que le facteur principal est fixe ou aléatoire. Dans tous les cas, nous verrons qu'il est impossible d'évaluer l'effet de l'interaction des deux facteurs.

L'écriture des modèles observé et théorique et la réalisation des calculs du modèle hiérarchisé ne peuvent présenter de difficultés majeures pour le lecteur qui a assimilé la théorie des modèles croisés.

6.4.2. Exemple introductif

On souhaite étudier l'effet de trois régions données sur la pousse moyenne des bourgeons de pommiers d'une variété donnée au cours d'une année. Dans chaque région, quatre pommiers ont été choisis au hasard et sur chacun d'entre eux quatre pousses, choisies aussi au hasard, ont été mesurées. Les données obtenues en centimètres sont présentées au tableau 6.4.

Tableau 6.4. Longueurs des pousses de bourgeons de pommiers relevées dans trois régions (en cm).

Régions	Pommiers	Pousse 1	Pousse 2	Pousse 3	Pousse 4
1	1	18.6	16.4	17.6	13.5
	2	23.3	19.2	13.7	15.0
	3	20.4	17.8	19.7	14.9
	4	17.6	14.9	20.0	18.6
2	1	9.9	10.8	14.3	11.9
	2	15.4	13.3	16.5	14.5
	3	14.2	14.4	16.4	13.7
	4	15.7	14.1	17.5	14.1
3	1	14.9	13.5	13.8	15.7
	2	21.8	18.9	19.2	18.3
	3	30.7	26.6	16.5	21.0
	4	21.6	17.8	20.5	20.1

Il s'agit ici d'un modèle hiérarchisé, car il apparaît qu'il n'y a aucune correspondance entre les différents pommiers des différentes régions. Par exemple, la désignation pommier 1 de la région 1 n'a pas le même sens que la désignation pommier 1 de la région 2 ou celle de la région 3. On dit que le facteur "pommiers" est subordonné au facteur "régions". Ce dernier est dit facteur principal.

6.4.3. Modèle mixte

Lorsque le modèle est mixte, c'est-à-dire lorsque le facteur principal est fixe et le facteur subordonné est aléatoire, l'échantillonnage à réaliser doit être à deux degrés. C'est le cas de l'exemple des pommiers où l'on a d'abord sélectionné aléatoirement quatre pommiers dans chaque région, et sur chacun d'entre eux, quatre pousses ont été également choisies au hasard.

Le modèle observé s'écrit donc :

$$(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..}) + (y_{ijk} - \bar{y}_{ij.}), \quad (6.3)$$

et l'équation fondamentale de l'analyse de la variance s'obtient en élevant au carré les deux membres de la quantité 6.3 puis en sommant pour toutes les observations :

$$\sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 = pn \sum_{i=1}^p (\bar{y}_{i..} - \bar{y}_{...})^2 + n \sum_{i=1}^p \sum_{j=1}^q (\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2$$

Ces quantités, qui rappellent les formules des sommes des carrées des écarts, s'écrivent avec leurs nombres de degrés de liberté comme suit :

$$\begin{aligned} SCE_t &= SCE_a + SCE_{b/a} + SCE_r \\ pqn - 1 &= (p - 1) + p(q - 1) + pq(n - 1) \end{aligned}$$

où la quantité $SCE_{b/a}$, qui peut aussi être écrite $SCE_{b(a)}$, représente la variabilité due au facteur subordonné (b) à l'intérieur du facteur principal (a). Les autres termes ont le même sens que dans le cas du modèle croisé.

Les différents carrés moyens s'obtiennent en divisant les sommes des carrées des écarts par leurs nombres de degrés de liberté :

$$CM_a = \frac{SCE_a}{p - 1}; \quad CM_{b/a} = \frac{SCE_{b/a}}{p(q - 1)}; \quad CM_r = \frac{SCE_r}{pq(n - 1)}.$$

Le modèle théorique associé au modèle observé s'écrit :

$$Y_{ijk} = \mu_{..} + a_i + B_{ij} + \varepsilon_{ijk}.$$

où B_{ij} représente une contribution aléatoire associée au $j^{\text{ème}}$ niveau du facteur subordonné correspondant au $i^{\text{ème}}$ niveau du facteur principal. On suppose que les quantités B_{ij} et ε_{ijk} sont des variables aléatoires normales indépendantes de moyenne nulle et de variances respectives σ_B^2 et σ^2 , et qu'elles sont indépendantes entre elles. Les quantités a_i sont les effets du facteur contrôlé, ce sont des constantes telles que :

$$a_i = \mu_{i.} - \mu_{..} \text{ avec } \sum_{i=1}^p a_i = 0.$$

Dans le cas où il y a présence d'effet du facteur a, on est souvent amené à fournir une mesure de cet effet par la quantité :

$$\delta_a^2 = \frac{1}{p-1} \sum_{i=1}^p a_i^2.$$

1) Deux hypothèses peuvent être testées :

– Absence d'effet du facteur principal sur les résultats :

$$H_0 : a_1 = a_2 = \dots = a_p = 0 \quad \text{contre} \quad H_1 : a_i \neq 0 \quad \text{pour au moins un } i.$$

– Absence d'effet du facteur subordonné sur les résultats :

$$H'_0 : \sigma_B^2 = 0 \quad \text{contre} \quad H'_1 : \sigma_B^2 \neq 0.$$

Cette seconde hypothèse est rarement testée par les chercheurs.

2) La réalisation de ces tests repose sur le calcul des espérances mathématiques et la connaissance des distributions théoriques des sommes des carrées des écarts ou des carrés moyens. Lorsque les conditions d'application du modèle sont vérifiées et que l'effectif n est supérieur à 1, on obtient les espérances mathématiques des carrés moyens suivantes :

- $E(CM_a) = \sigma^2 + n\sigma_B^2 + qn\delta_a^2$ (cf. paragraphe 6.3.3.2)
- $E(CM_{b/a}) = \sigma^2 + n\sigma_B^2$
- $E(CM_r) = \sigma^2$

et les distributions des sommes des carrées des écarts suivantes :

- Sous H_0 , $\frac{SCE_a}{\sigma^2 + n\sigma_B^2}$ suit une distribution χ^2 à p-1 degrés de liberté ;
- $\frac{SCE_{b/a}}{\sigma^2 + n\sigma_B^2}$ suit une distribution χ^2 à p(q-1) degrés de liberté ;
- $\frac{SCE_r}{\sigma^2}$ suit une distribution χ^2 à pq(n-1) degrés de liberté.

Les espérances mathématiques et les distributions théoriques obtenues montrent que :

- Le test de l'hypothèse H_0 nécessite le calcul de la statistique :

$$F_a = \frac{CM_a}{CM_{b/a}}$$

et le rejet de cette hypothèse lorsque $F_a \geq F_{1-\alpha}$, où la valeur de $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=p-1$ et $k_2=p(q-1)$ degrés de liberté (**p-value** < α) ;

- Le test de l'hypothèse H'_0 nécessite le calcul de la statistique :

$$F_{b/a} = \frac{CM_{b/a}}{CM_r}$$

et le rejet de cette hypothèse lorsque $F_{b/a} \geq F_{1-\alpha}$, où la valeur de $F_{1-\alpha}$ est lue sur la table statistique avec $k_1=p(q-1)$ et $k_2=pq(n-1)$ degrés de liberté (**p-value** < α) ;

6.4.4. Modèle aléatoire

Dans le cas du modèle aléatoire, l'échantillonnage à réaliser est à trois degrés. Dans l'exemple des pommiers, si le facteur "régions" avait été aléatoire, on aurait réalisé un échantillonnage à trois degrés car les trois régions seraient elles-mêmes choisies au hasard au premier degré.

Le modèle observé est le même que dans le cas du modèle mixte, mais le modèle théorique doit tenir compte du caractère aléatoire des moyennes $M_{i.}$:

$$Y_{ijk} = \mu_{..} + A_i + B_{ij} + \varepsilon_{ijk} ;$$

où $A_i = M_{i.} - \mu_{..}$.

On suppose que les quantités A_i , B_{ij} et ε_{ijk} sont des variables aléatoires normales indépendantes de moyenne nulle et de variances respectives σ_A^2 , σ_B^2 et σ^2 , ces variables sont également supposées indépendantes entre elles.

1) Les deux hypothèses à tester sont donc :

– Le facteur principal n'agit pas sur les résultats :

$$H_0 : \sigma_A^2 = 0 \quad \text{contre} \quad H_1 : \sigma_A^2 \neq 0$$

– Le facteur subordonné n'a pas d'effet sur les résultats :

$$H'_0 : \sigma_B^2 = 0 \quad \text{contre} \quad H'_1 : \sigma_B^2 \neq 0 .$$

2) Lorsque les conditions d'application du modèle sont vérifiées, on obtient les espérances mathématiques des carrés moyens suivantes :

- $E(CM_a) = \sigma^2 + n\sigma_B^2 + qn\sigma_A^2$
- $E(CM_{b/a}) = \sigma^2 + n\sigma_B^2$
- $E(CMr) = \sigma^2$

et les distributions des sommes des carrées des écarts suivantes :

- sous H_0 , $\frac{SCE_a}{\sigma^2 + n\sigma_B^2}$ suit une distribution χ^2 à p-1 degrés de liberté et $\frac{SCE_{b/a}}{\sigma^2 + n\sigma_B^2}$ suit une distribution χ^2 à p(q-1) degrés de liberté ;
- sous H'_0 , $\frac{SCE_{b/a}}{\sigma^2}$ suit une distribution χ^2 à p(q-1) degrés de liberté et $\frac{SCE_r}{\sigma^2}$ suit une distribution χ^2 à pq(n-1) degrés de liberté.

Les espérances mathématiques et les distributions théoriques des sommes des carrés des écarts obtenues montrent donc que les tests se réalisent de la même manière que dans le cas du modèle mixte (cf. paragraphe 6.4.3).

6.4.5. Réalisation de l'analyse de la variance

6.4.5.1. Formules de calculs

1) Une première manière de procéder consiste à obtenir le même tableau d'analyse de la variance que dans le cas du modèle croisé, sauf que l'on obtient la somme des carrées des écarts liée au facteur subordonné $SCE_{b/a}$ par la relation :

$$SCE_{b/a} = SCE_t - SCE_a - SCE_r.$$

Autrement dit, $SCE_{b/a}$ regroupera les quantités SCE_b et SCE_{ab} du modèle croisé.

2) On peut aussi réaliser p ANOVA 1, une pour chaque niveau du facteur principal, comme si l'on veut "comparer" chaque fois les q niveaux du facteur hiérarchisé, et de regrouper ensuite les résultats. On obtient :

$$SCE_r = \sum_{i=1}^p SCE_{r_i} \text{ et } SCE_{b/a} = \sum_{i=1}^p SCE_{b_i}$$

où SCE_{r_i} et SCE_{b_i} sont les sommes des carrées des écarts résiduelle et factorielle de la i^{ème} ANOVA 1.

La somme des carrées des écarts du facteur principal se calcule par la relation :

$$SCE_a = \frac{1}{qn} \sum_{i=1}^p Y_{i..}^2 - C$$

si les effectifs des échantillons et les nombre des niveaux du facteur hiérarchisé par niveau du facteur principal sont constants ou par :

$$SCE_a = \sum_{i=1}^p \frac{Y_{i..}^2}{n_{i.}} - \frac{Y_{...}^2}{n..}$$

si les effectifs des échantillons et/ou les nombre des niveaux du facteur hiérarchisé par niveau du facteur principal sont différents. Les symboles $n_{i.}$ et $n_{..}$ désignent respectivement le nombre d'observations pour le niveau i du facteur principal et le nombre total d'observations.

6.4.5.2. Tableau d'analyse de la variance

Le tableau d'analyse de la variance ressemble à celui du modèle croisé. Il ne comprend cependant que trois sources de variation (tableau 6.5).

Tableau 6.5. Esquisse du tableau d'analyse de la variance à deux critères de classification, modèle hiérarchisé.

S. de variation	Degrés de liberté	SCE	CM	F _{obs}
Facteur a	p-1	SCE _a	CM _a	
Facteur b/a	p(q-1)	SCE _{b/a}	CM _{b/a}	
Variation résiduelle	pq(n-1)	SCE _r	CM _r	
Variation totale	pqn-1	SCE_t		

Exemple 6.3

Reprenez les données du tableau 6.4 et testez s'il existe des différences significatives entre les longueurs des pousses des bourgeons dans les trois régions ?

Solution 6.3

a) **Méthode** : ANOVA 2, modèle hiérarchisé mixte

- ANOVA 2 : il y a un facteur "régions" et un facteur "pommiers" (p=3, q=4 et n=4)
- Modèle mixte : le facteur "régions" est fixe et le facteur "pommiers" est aléatoire

b) **Conditions d'application** :

- voir les conditions d'application du modèle au paragraphe 6.4.3.

c) **Hypothèse nulle** :

$$H_0: \mu_{1.} = \mu_{2.} = \mu_{3.} \text{ contre } H_1: \mu_i \neq \mu_j \text{ pour au moins deux régions.}$$

d) **Calcul** :

$$Y_{..} = 818.80 ; C = (818.8^2/48) = 13967.36 ; T = 14673.42$$

$$SCE_t = T - C = 706.06$$

$$SCE_1 = 14.628; SCE_2 = 56.860; \dots; SCE_{12} = 7.660 \text{ et } SCE_r = SCE_1 + SCE_2 + SCE_{12} = 266.485$$

$$SCE_{régions} = (281.2^2 + 226.7^2 + 310.9^2)/16 = 227.96$$

$$SCE_{pommiers} = SCE_t - SCE_{régions} - SCE_r = 211.61$$

e) **Tableau d'analyse de la variance** :

Source de variation	d.l	SCE	CM	Fobs	Prob
Régions	2	227.96	113.98	4.85	0.0373
Pommiers (Régions)	9	211.61	23.513		
Résidus	36	266.49	7.402		
Total	47	706.06	-	-	-

La statistique observée dépasse la valeur théorique $F_{0.95} = 4.256$ sous l'hypothèse nulle. On rejette alors l'hypothèse nulle avec un risque de première espèce qui est exactement égal à p-value qui est de 0.0373.

En prenant un risque de 5%, on peut conclure à des différences significatives entre les longueurs moyennes des pousses obtenues dans les trois régions (c'est-à-dire **un effet région significatif**).

Exemple 6.4

Une expérience clinique a été réalisée dans le but de comparer trois traitements améliorant la récupération de la pression artérielle des patients après un type d'opérations. Chaque traitement a été reçu par un certain nombre de patients choisis au hasard. Les pressions ont été mesurées trois fois pour chaque patient à des moments différents de la journée. Analysez aussi complètement que possible les résultats du tableau ci-dessous ?

Traitements	Patients			
1	1	113	85	119
	2	127	103	105
	3	115	96	105
2	1	142	97	96
	2	109	125	112
	3	103	130	111
	4	132	98	104
3	1	110	92	111
	2	97	94	86
	3	103	114	91

Solution 6.4

a) Méthode : ANOVA 2, modèle hiérarchisé mixte (les 10 patients sont différents)

- ANOVA 2 : car il y a un facteur "traitements" et un facteur "patients" ($p=3$, q est variable et $n=3$)

- Modèle mixte : le facteur "traitements" est fixe et le facteur "patients" est aléatoire

b) Conditions d'application :

- voir les conditions du modèle au paragraphe 6.4.3.

c) Hypothèse nulle :

$H_0: \mu_1 = \mu_2 = \mu_3$, contre $H_1: \mu_i \neq \mu_j$, pour au moins 2 traitements.

d) Calcul :

Sommes par traitement : $Y_{1..} = 968$, $Y_{2..} = 1359$, $Y_{3..} = 898$ et total général $Y_{...} = 3225$.

Les trois ANOVA 1, une par traitement, permettent d'obtenir SCE_{r_i} et SCE_{b_i} pour $i=1, 2, 3$:

$$SCE_{b/a} = \sum_{i=1}^p SCE_{b_i} = 76.222 + 37.583 + 253.556 = 367.36$$

$$SCE_r = \sum_{i=1}^p SCE_{r_i} = 1194.00 + 2568.67 + 558.00 = 4320.67 ;$$

$$SCE_a = \left(\frac{968^2}{9} + \frac{1359^2}{12} + \frac{898^2}{9} \right) - \frac{3225^2}{30} = 933.47$$

$$SCE_t = 933.47 + 367.96 + 4320.67 = 5621.50$$

e) Tableau d'analyse de la variance :

Source de variation	d.l	SCE	CM	F _{obs}	p-value
Traitements	2	933.47	466.74	8.89	0.012
Patients (trait)	7	367.36	52.48		
Résidus	20	4320.67	216.03		
Total	29	5621.50	-	-	-

La statistique observée dépasse la valeur théorique $F_{0.95} = 4.737$ sous l'hypothèse nulle. On rejette alors l'hypothèse nulle avec un risque de première espèce qui est exactement égal à la valeur de la p-value qui est de **0.012**.

En prenant un risque de 5%, on peut conclure à des différences significatives entre les pressions artérielles obtenues avec les trois traitements.

6.5 CONCLUSION

Nous avons vu que l'ANOVA2 permet d'étudier les effets de deux facteurs à la fois sur une variable dépendante. Nous avons souligné qu'il faut distinguer entre le modèle croisé lorsque les deux facteurs ont un rôle symétrique et le modèle hiérarchisé lorsque l'un des facteurs est subordonné à l'autre.

L'ANOVA 2 présente plusieurs avantages par rapport à la réalisation de plusieurs ANOVA 1 par niveau de l'un des deux facteurs. Elle permet d'étudier, dans le cas du modèle croisé, l'interaction des deux facteurs qui est parfois parmi les résultats recherchés d'une étude. Ensuite, on peut vérifier qu'on gagne en puissance.

Nous avons donné un exemple de modèles à effets fixes où nous avons testé l'interaction avant les effets principaux des deux facteurs. La présence d'une interaction signifie qu'un facteur exerce des effets différents selon le niveau de l'autre facteur, et, par conséquent, le biologiste est amené à être attentif quant à l'information apportée par l'analyse des effets principaux sur le phénomène étudié. Des tests concernant les effets simples, c'est-à-dire l'effet d'un facteur à chaque niveau de l'autre facteur, ou des analyses par des contrastes peuvent aider à interpréter les résultats.

L'ANOVA 2 est une méthode robuste. Elle peut être utilisée en présence de légères déviations par rapport aux conditions de normalité et d'égalité des variances.

Comme toute analyse de la variance, l'ANOVA 2 peut être complétée par des méthodes de comparaisons multiples de moyennes lorsque l'effet de l'un ou des deux facteurs ou de l'interaction s'avère significatif, pour autant que ces sources de variation soient fixes. Ces méthodes, qui feront l'objet du chapitre 7, permettent de déceler les moyennes qui diffèrent significativement entre elles.

CHAPITRE 7

Procédures de comparaisons multiples de moyennes

7.1. INTRODUCTION

Lorsqu'on teste l'égalité des moyennes de plus de deux populations, l'analyse de la variance nous permet de dire, lorsque le test est significatif, qu'il y a au moins deux moyennes qui diffèrent significativement. Autrement dit, on ne peut savoir quelles moyennes diffèrent de quelles autres parmi toutes les moyennes testées. Or il se peut que les expériences mises en place par le chercheur aient pour but des tests d'hypothèses beaucoup plus spécifiques sur les moyennes. Cette question peut être traitée par les méthodes de comparaisons multiples de moyennes¹.

Les méthodes de comparaisons multiples de moyennes n'ont du sens qu'avec des facteurs de nature fixe. Elles nous renseignent, après un certain nombre de tests, sur les éventuels groupes homogènes de moyennes, c'est-à-dire sur les groupes dont les moyennes ne diffèrent pas significativement entre elles.

On distingue ce qu'on appelle les tests de comparaisons *a priori* et les tests de comparaisons *a posteriori* (*post-hoc*) selon que les tests de comparaison sont planifiés avant ou après la collecte des données. Dans le premier cas, les comparaisons, déterminées préalablement par l'expérimentateur, sont dirigées par la théorie, et, par conséquent, leur nombre est petit. Tandis que les comparaisons *a posteriori* sont formulées lorsque les données sont collectées et les moyennes sont examinées et rangées par l'expérimentateur, et, par conséquent, les comparaisons entre toutes les paires de moyennes sont justifiées dans ce cas. On retient que si les comparaisons sont planifiées à l'avance la probabilité de commettre au moins une erreur de type I est plus petite que si les comparaisons n'interviennent qu'à posteriori, sauf si l'on planifie toutes les comparaisons possibles à l'avance. Généralement, on réalise les comparaisons planifiées à l'avance au moyen d'une analyse par contraste.

Mais, peut-on dire que l'on ne peut utiliser les tests de comparaisons multiples de moyennes que si le test F d'ANOVA est significatif ? Cette condition n'est en effet requise que pour quelques tests de comparaisons multiples. Ce sujet controversé de la nécessité ou non d'un test F global significatif au préalable a été d'ailleurs repris par plusieurs auteurs, notamment par Wlcox (1987) et Howell (2007). Certains d'entre eux estiment que la persistance de l'examen du test F de l'ANOVA avant la réalisation des tests de comparaisons multiples de moyennes relève de la tradition. Il peut arriver dans des cas très limites que l'ANOVA ne montre pas de différences significatives entre les moyennes, alors que les résultats des comparaisons multiples indiquent que certaines moyennes sont différentes.

Par ailleurs, l'expérimentateur non familier peut rencontrer des problèmes lorsqu'il souhaite utiliser ces méthodes. D'abord, il doit opérer un choix devant une panoplie de procédures proposées par les logiciels statistiques. Ensuite, il doit être attentif quant à l'interprétation des résultats lorsqu'il y a des chevauchements entre des groupes homogènes de moyennes.

¹ Les contrastes seront traités au cours de la deuxième année du cycle ingénieur.

L'objectif de ce chapitre sera d'abord d'exposer, après la présentation d'un exemple introductif qui nous servira à illustrer les calculs (paragraphe 7.2), la problématique du contrôle du risque d'erreur de première espèce lorsqu'on utilise des méthodes de comparaisons multiples de moyennes (paragraphe 7.3). Ce sont d'ailleurs les différentes approches adoptées pour contrôler ce risque d'erreur sans trop sacrifier de puissance qui ont fait que les méthodes de comparaison proposées dans la littérature sont nombreuses. Le paragraphe 7.4 nous permettra de distinguer les comparaisons *a priori* et les comparaisons *a posteriori*. L'objectif sera ensuite de présenter quelques-unes de ces méthodes, en les classant en méthodes de comparaisons des moyennes deux à deux (paragraphe 7.5) et en méthodes de comparaisons des moyennes par groupes (paragraphe 7.6). Nous exposerons enfin la comparaison des moyennes d'un certain nombre de traitements à la moyenne d'un traitement témoin (paragraphe 7.7).

Dans tous les cas, nous supposerons que les conditions de base de l'analyse de variance sont toujours vérifiées tant en ce qui concerne le caractère aléatoire et simple des échantillons que la normalité et l'égalité des variances des populations parents. Nous supposerons aussi, sauf mention particulière, que les échantillons prélevés sont tous de même effectif.

7.2. EXEMPLE INTRODUCTIF

Un chercheur souhaite tester l'effet de la race sur le rendement à l'abattage des bovins mâles. Il dispose des valeurs du tableau 7.1, exprimant le rendement poids mort/poids vif en pour-cent, en considérant un échantillon aléatoire et simple par race de quatre animaux ayant pratiquement le même âge. Il souhaite procéder aux comparaisons multiples de moyennes.

Tableau 7.1. Rendements poids mort/poids vif) x 100, obtenus pour des échantillons d'animaux appartenant à cinq races.

Race 1	Race 2	Race 3	Race 4	Race 5
83.3	85.1	82.6	83.9	80.2
83.7	82.9	82.7	80.9	78.1
85.0	85.9	83.0	82.8	82.1
86.5	84.2	81.3	82.4	80.8

Ces données nous serviront à illustrer les calculs des différentes méthodes de comparaisons multiples de moyennes qui seront traitées au fil des paragraphes de ce chapitre.

7.3. RISQUE DE PREMIÈRE ESPÈCE

7.3.1. Généralité

Lorsqu'on aborde le sujet des procédures de comparaisons multiples de moyennes, on évoque presque systématiquement la problématique du risque de commettre des erreurs de première espèce. Nous avons vu, au paragraphe 5.2.1, que la multiplication du nombre de tests *t* de Student peut faire augmenter de manière importante la probabilité de commettre au moins une erreur de première espèce, c'est-à-dire le risque de déclarer au moins deux moyennes comme différentes, lorsqu'en réalité elles sont identiques.

En effet, lorsqu'on évoque les comparaisons multiples de moyennes, on est amené à distinguer ce qu'on appelle le risque d'erreur par comparaison que nous noterons α (paragraphe 7.3.2) et le risque global d'erreur que nous noterons α_g (paragraphe 7.3.3), ce dernier peut être aussi appelé aussi taux d'erreur de l'ensemble.

7.3.2. Risque d'erreur par comparaison

Le risque d'erreur par comparaison¹ est la probabilité de commettre une erreur de première espèce dans une comparaison donnée. Si l'on réalise, par exemple, une seule comparaison entre deux moyennes, en testant l'hypothèse élémentaire :

$$H_0^{ij} : \mu_i = \mu_j \quad \text{où } i \neq j \quad (7.1)$$

par le test t de Student au niveau $\alpha=0.05$, le risque d'erreur pour cette comparaison sera égal à 0.05. Autrement dit, au maximum dans 5% des cas, on rejettera l'hypothèse nulle à tort.

7.3.3. Risque global d'erreur

Lorsqu'on réalise un ensemble de comparaisons entre plusieurs moyennes, le risque global d'erreur de première espèce² est la probabilité de commettre au moins une erreur de première espèce parmi toutes les comparaisons en question.

Considérons l'une des approches des comparaisons multiples des moyennes qui consiste à tester chacune des hypothèses élémentaires avec un risque d'erreur par comparaison α . Dans ce cas, le risque global d'erreur de première espèce (α_g) est la probabilité de déclarer au moins une hypothèse élémentaire comme fausse lorsque l'hypothèse globale est vraie.

Ce risque global d'erreur (α_g) est plus grand que le risque d'erreur par comparaison (α) et les deux risques sont égaux dans le cas où l'expérience ne comporte qu'une seule comparaison.

Lorsque les comparaisons deux à deux sont indépendantes les unes des autres, le risque global d'erreur est donné par (cf. paragraphe 5.2.1) :

$$\alpha_g = 1 - (1 - \alpha)^c \quad (7.2)$$

où c représente le nombre de comparaisons à réaliser. Le tableau 7.2 montre que les valeurs du risque global d'erreur augmentent considérablement avec l'augmentation du nombre de comparaisons. En particulier, si l'expérimentateur réalise 10 comparaisons, il y a 40% de chance de commettre au moins une erreur de première espèce, c'est-à-dire de trouver une différence qui en réalité n'existe pas.

Tableau 7.2. Variation du risque global (α_g) avec le nombre de comparaisons indépendantes (c) lorsque le risque par comparaison est $\alpha=0.05$.

c	1	5	10	15	21	45
α	0.05	0.05	0.05	0.05	0.05	0.05
α_g (%)	5,0%	22,6%	40,1%	53,7%	65,9%	90,1%

Dans les comparaisons multiples de moyennes, les différentes comparaisons ne sont pas dans la réalité tout à fait indépendantes. On retient que lorsqu'on réalise c comparaisons ayant chacune une probabilité α d'erreur de type I, la probabilité de commettre au moins une erreur de type I ne peut jamais dépasser $c \cdot \alpha$. On peut donc noter que les limites du α_g sont :

$$\alpha \leq \alpha_g \leq c \cdot \alpha \quad (7.3)$$

Dans la plupart des situations, cependant, la quantité α_g , calculée selon l'expression 7.2, représente une approximation raisonnable de α_g .

¹ En anglais : *comparisonwise error rate*.

² En anglais : *overall error rate*.

Dans les comparaisons multiples, l'exactitude de la conclusion générale dépend de l'ensemble des comparaisons. Il faudra donc veiller à maîtriser le risque global d'erreur. C'est pourquoi, les différentes procédures de comparaisons multiples cherchent, en adoptant des approches différentes, à contrôler ce risque global d'erreur. Mais, rappelons-nous, le contrôle du risque de l'erreur de première espèce et la puissance sont deux choses antagonistes (cf. paragraphe 2.2.3). Il vaut donc mieux penser aux méthodes qui assurent un meilleur équilibre entre le risque de l'erreur de première espèce et la puissance.

7.4. COMPARAISONS A PRIORI ET A POSTERIORI

7.4.1. Généralités

Certains expérimentateurs ont parfois tendance à réaliser un grand nombre de tests d'hypothèses se rapportant aux comparaisons multiples de moyennes. Ce nombre élevé de tests se répercute négativement sur la puissance et ne peut donc être justifié théoriquement, sauf peut-être lorsqu'on vise à explorer au maximum les données recueillies. Ceci amène à distinguer deux types d'approches de comparaisons :

- Les comparaisons *a posteriori* qui ne sont précisées qu'après examen des données collectées ;
- Les comparaisons *a priori* qui sont planifiées avant que les données ne soient collectées.

7.4.2. Les comparaisons *a posteriori*

Les comparaisons *a posteriori*¹ ne sont pas planifiées, car le chercheur n'a pas d'idées claires quant aux différences qu'il compte observer. Les comparaisons sont donc effectuées après avoir collecté les données et examiné les moyennes. On comprend donc que les comparaisons entre les moyennes de n'importe quel groupe sont justifiées.

Ces comparaisons sont souvent moins puissantes par rapport aux comparaisons *a priori*, mais elles permettent de déceler toutes les différences significatives entre les moyennes.

7.4.2. Les comparaisons *a priori*

Dans ce cas, les hypothèses à tester sont énoncées avant la collecte des données en se basant sur l'expérience de l'expérimentateur ou sur des considérations théoriques. Ces tests permettent éventuellement d'obtenir des confirmations empiriques de certaines hypothèses théoriques en se servant des données réelles.

Comme ces comparaisons sont planifiées à l'avance, leur nombre est souvent réduit par rapport au nombre total de comparaisons possibles. De ce fait, le risque global d'erreur à contrôler ne concerne que les comparaisons planifiées. Elles présentent aussi l'avantage d'avoir une puissance élevée. Leur inconvénient est qu'elles sont sélectives dans la mesure où certains groupes de moyennes ne sont pas comparés bien que cela puisse être intéressant.

Il existe différents types de procédures *a priori*. Il s'agit généralement de contrastes qui permettent de réaliser des comparaisons spécifiques puisque les hypothèses sont formulées préalablement à la collecte des données. Un contraste constitue une combinaison linéaire des p moyennes où l'on compare les moyennes de certains groupes à d'autres (le problème des contrastes ne sera pas traité dans ce chapitre).

¹ En anglais : *post-hoc tests*.

Si le nombre de comparaisons à réaliser est assez élevé, les méthodes *a priori* n'ont plus l'avantage de puissance et, dans ce cas, certains auteurs conseillent même d'effectuer les comparaisons par les méthodes *a posteriori*.

7.5. MÉTHODES DE COMPARAISONS DES MOYENNES DEUX À DEUX

7.5.1. Introduction

Les méthodes de comparaisons des moyennes deux à deux consistent à réaliser les comparaisons de toutes les moyennes prises deux à deux. Pour p moyennes, cela revient à tester la signification des $p(p-1)/2$ hypothèses suivantes :

$$H_0^{ij} : \mu_i = \mu_j \quad \text{où } 1 \leq i < j \leq p. \quad (7.4)$$

Le principe consiste à calculer la différence entre les deux moyennes observées :

$$d_o = |\bar{y}_i - \bar{y}_j|$$

où \bar{y}_i et \bar{y}_j sont des estimations des moyennes μ_i et μ_j des deux populations, et à rejeter l'hypothèse nulle lorsque cette différence dépasse la valeur critique suivante :

$$v_c = \delta_{(\alpha)}^{(p,v)} \sqrt{2CM_e/n}$$

dans laquelle :

- δ est une valeur théorique relative à la distribution d'échantillonnage de la méthode utilisée. Sa valeur est donnée dans des tables en fonction du nombre de degrés de liberté v du CM_e , du niveau de signification α et parfois, aussi, du nombre de moyennes p ;
- CM_e est le carré moyen résiduel issu de l'ANOVA 1 à v degrés de liberté. Dans le cas, d'une analyse de la variance à deux ou plusieurs critères de classification, on le remplace par le carré moyen qui a servi à la comparaison des moyennes du facteur fixe, c'est-à-dire par le carré moyen résiduel si le modèle est fixe ou par le carré moyen de l'interaction si le modèle est mixte ;
- n est le nombre total d'observations à partir duquel est calculée chacune des p moyennes.

La quantité $\sqrt{2CM_e/n}$ est une estimation de l'erreur-standard de la différence entre deux moyennes.

D'un point de vue pratique, on ne réalise pas toutes les comparaisons des moyennes prises deux à deux. On commence par ordonner les p moyennes observées par ordre croissant puis on applique la méthode de comparaison des moyennes d'une manière séquentielle. Autrement dit, on teste chaque fois l'hypothèse qui fait intervenir les deux moyennes les plus éloignées possibles, et si cette hypothèse n'est pas rejetée, il est inutile de tester les hypothèses qui la composent.

Dans la littérature, on trouve plusieurs méthodes de comparaisons des moyennes deux à deux. Quatre d'entre-elles seront exposées. Il s'agit des méthodes de la plus petite différence significative (PPDS), de Tukey, de Bonferroni et Dunn-Sidak (paragraphe 7.5.2 à 7.5.5).

7.5.2. Méthode de la plus petite différence significative

Connue beaucoup plus par la procédure LSD¹ de Fisher [Fisher, 1935] ou PPDS, cette méthode fut l'une des plus anciennes à être utilisée pour localiser les différences entre les moyennes. Elle consiste à comparer les moyennes deux à deux par des tests t de Student, chacun avec un niveau de signification α .

La valeur critique est appelée la plus petite différence significative. Elle fait intervenir la valeur t de la distribution de Student à ν degrés de liberté. Elle est donnée par :

$$v_c = t_{1-\alpha/2}^{(\nu)} \sqrt{\frac{2CM_e}{n}}$$

si l'effectif n des échantillons est constant, et par :

$$v_c = t_{1-\alpha/2}^{(\nu)} \sqrt{CM_e \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

lorsque les deux moyennes à comparer sont calculées respectivement à partir de n_i et n_j observations.

Le test PPDS est très puissant, mais le risque global d'erreur est contrôlé en imposant simplement un test F de l'ANOVA significatif, ce qui fait que l'application pour un ensemble de moyennes entraîne un risque global d'erreur α_g bien supérieur au niveau α relatif à chacun des tests (les comparaisons ne sont pas en effet indépendantes et le nombre de comparaisons est important). Cet inconvénient, confirmé par plusieurs études basées sur des simulations [Bernhardson, 1975; Carmer et Swanson, 1973], incite à interpréter les résultats de cette méthode avec prudence et fait que de nombreux statisticiens la déconseillent sauf dans le cas de trois moyennes avec un test F de l'ANOVA significatif.

On retient donc qu'il s'agit d'un test qui requiert un F d'ANOVA significatif, qu'il est puissant mais très libéral.

Exemple 7.1

Reprenez les données du tableau 7.1 et réalisez les comparaisons multiples de moyennes en utilisant la méthode de la plus petite différence significative ($\alpha=0.05$) ?

Solution 7.1

a) Méthode : comparaisons multiples de moyennes par la procédure PPDS

- Nous avons cinq races ($p=5$), quatre observations par échantillon ($n=4$)
- L'ANOVA 1 donne un carré moyen résiduel $CM_e=1.726$ auquel correspond $\nu=15$ degrés de liberté.

b) Conditions d'application :

- les cinq échantillons sont aléatoires, simples et indépendants
- les distributions des cinq populations sont normales
- les cinq variances théoriques sont égales : le test de Bartlett donne p-valeur=0.810 (le test d'égalité des cinq variances ne peut être rejetée au niveau $\alpha=0.05$)

¹ En anglais : *least significant difference*.

c) Calculs :

- La valeur critique est :

$$v_c = 2.1314 \sqrt{\frac{(2)(1.72633)}{4}} = 1.980 \text{ ou } 1.98$$

- Les moyennes sont rangées par ordre croissant et toute différence entre deux moyennes dépassant 1.98 doit être déclarée comme significative. On a :

$$|\bar{x}_1 - \bar{x}_5| = 4.325 > 1.98 \rightarrow \text{groupe non homogène ;}$$

$$|\bar{x}_1 - \bar{x}_3| = 2.225 > 1.98 \rightarrow \text{groupe non homogène ;}$$

$$|\bar{x}_1 - \bar{x}_4| = 2.125 > 1.98 \rightarrow \text{groupe non homogène ;}$$

$$|\bar{x}_1 - \bar{x}_2| = 0.100 < 1.98 \rightarrow \text{groupe homogène ;}$$

$$|\bar{x}_2 - \bar{x}_5| = 4.225 > 1.98 \rightarrow \text{groupe non homogène ;}$$

$$|\bar{x}_2 - \bar{x}_3| = 2.0125 > 1.98 \rightarrow \text{groupe non homogène ;}$$

etc.

On obtient, si l'on relie les moyennes non significativement différentes par un trait continu :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625

Il y a donc formation de trois groupes homogènes de moyennes : [Race1, Race2], [Race3, Race4] et [Race5]. Le rendement est significativement plus élevé chez les races 1 et 2 que chez les races 3 et 4, et il est significativement plus élevé chez les races 3 et 4 que chez la race 5.

Plusieurs logiciels présentent les résultats en marquant les moyennes non significativement différentes par la même lettre :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625
a				
	b	b		
			c	c

7.5.3. Méthode de Tukey

Cette méthode, due à Tukey [1951, 1953] et connue aussi par la méthode de la différence franchement significative¹, compare les moyennes deux à deux et elle est n'est applicable que pour des échantillons de mêmes effectif n. Le test de Tukey utilise la loi des écarts studentisés et considère deux moyennes comme différentes si leur différence, en valeur absolue, égale ou dépasse la quantité :

$$v_c = Q_{1-\alpha}^{(p,v)} \sqrt{\frac{CM_e}{n}}$$

où la valeur de Q peut être consultée dans des tables en fonction du niveau α , du nombre de moyennes considérées dans l'étude (p) et du nombre de degrés de liberté (v) du CM_e . La quantité v_c est appelée la plus petite amplitude significative, elle est la même pour toutes les comparaisons.

¹ En anglais : *honestly significant difference (HSD)*.

La procédure de Tukey compare tous les appariements de groupes possibles et indique ceux qui présentent des différences statistiquement significatives. Comme pour la méthode de la plus petite différence significative (PPDS), on commence par ranger les moyennes par ordre croissant ou décroissant et on relie par une ligne les moyennes qui ne diffèrent pas.

Le test de Tukey permet de bien contrôler le risque global d'erreur au niveau α pour l'ensemble des comparaisons deux à deux. Nous pouvons être sûrs que l'ensemble de toutes les comparaisons a collectivement un taux d'erreur de α (le risque global d'erreur α_g est égale à α (à 5% par exemple)). Il permet aussi le calcul d'intervalles de confiance pour les différences entre les moyennes. Il est préféré par de nombreux utilisateurs en raison du contrôle qu'il exerce sur α .

Le test de Tukey est conservateur. Si une différence entre deux moyennes s'avère significative par ce test, elle sera sûrement significative par la méthode PPDS. Ce conservatisme n'est pas excessif en comparaison avec d'autres méthodes contrôlant le risque global d'erreur.

Le contrôle du risque global d'erreur s'associe à une perte de puissance, c'est-à-dire à une incapacité de mettre en évidence les vraies différences entre les moyennes, surtout lorsque le nombre de comparaisons est élevé.

Lorsque les échantillons ne sont pas de mêmes effectifs, on peut utiliser la méthode de Tukey-Kramer, qui est une extension approximative de la méthode de Tukey. On utilise en effet une valeur critique légèrement modifiée :

$$v_{ij} = Q_{1-\alpha}^{(p, \nu)} \sqrt{\frac{CM_e}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

pour tester si deux moyennes sont différentes. Certains auteurs proposent de prendre la moyenne harmonique des effectifs comme l'effectif de chaque échantillon.

Exemple 7.2

Reprenez les données du tableau 7.1 et réalisez les comparaisons multiples de moyennes en utilisant la méthode de Tukey ($\alpha=0.05$).

Solution 7.2

a) Méthode : comparaisons multiples de moyennes par la procédure de Tukey

b) Conditions d'application : (cf. exemple 7.1)

c) Calculs :

La plus petite amplitude significative est :

$$v_c = 4.367 \sqrt{\frac{1.72633}{4}} = 2.8689 \text{ ou } 2.87$$

Après avoir rangé les moyennes par ordre croissant et déclaré toute différence entre deux moyennes dépassant 2.869 comme significative, on obtient :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625

Il y a formation de deux groupes homogènes de moyennes qui se chevauchent : [Race1, Race2, Race4, Race3] et [Race4, Race3, Race5]. On peut voir que le rendement est significativement plus élevé chez les races 1 et 2 que chez la race 5.

7.5.4. Méthode de Bonferroni

La méthode de Bonferroni, proposée par Fisher et appelée aussi méthode de Dunn-Bonferroni, utilise une transformation de la méthode PPDS pour contrôler le risque global d'erreur en utilisant des niveaux de signification respectant l'inégalité de Bonferroni. Elle consiste à fixer un risque d'erreur par comparaison très petit de telle sorte que le risque global d'erreur reste raisonnable. En effet, on sait, d'après la relation 7.3, que :

$$\alpha_g \leq c \cdot \alpha ,$$

c'est-à-dire que le risque global d'erreur est au plus égal à $c \cdot \alpha$. Dans ces conditions, si chacune des hypothèses 7.1 est testée au niveau $\alpha' = \alpha/c$, alors le risque global d'erreur, c'est-à-dire le risque de commettre au moins une erreur de première espèce, ne peut dépasser le niveau α .

Une différence entre deux moyennes est donc déclarée comme significative, selon cette méthode, si elle égale ou dépasse la valeur critique suivante :

$$v_c = t_{1-\frac{\alpha}{2c}}^{(\nu)} \sqrt{\frac{2CM_e}{n}}$$

si les effectifs qui ont permis de calculer les deux moyennes sont égaux, ou encore :

$$v_c = t_{1-\frac{\alpha}{2c}}^{(\nu)} \sqrt{CM_e \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

si ces effectifs ne sont pas égaux. On constate que l'utilisation de la méthode de Bonferroni nécessite la consultation de tables très détaillées de la distribution t de Student.

Si on planifie, par exemple, quatre comparaisons et que l'on souhaite que le risque global d'erreur ne dépasse pas 5%, on utilisera un risque d'erreur par comparaison de $0.05/4=0.0125$. Cela veut dire que si on réalise 4 comparaisons, chacune avec une probabilité de l'erreur de type I de $\alpha=0.0125$, la probabilité d'avoir au moins une erreur de type I ne peut jamais dépasser $4*0.0125=0.05$.

D'autre part, il faut dire que la méthode de Bonferroni est très conservatrice dans la mesure où le risque global d'erreur réel α_g est beaucoup plus petit que le niveau fixé α . Ce contrôle très sévère du risque global d'erreur s'associe à une perte de puissance, c'est-à-dire à une incapacité de mettre en évidence les vraies différences entre les moyennes, surtout lorsque le nombre de comparaisons est élevé.

Le test s'applique aux deux types de comparaisons (planifiées et post-hoc), mais recommandé pour des comparaisons planifiées où le nombre de comparaisons à effectuer est petit.

Plusieurs variantes du test de Bonferroni ont été proposés dans le but d'améliorer la puissance. On peut citer, à titre d'exemples, les tests de Dunn-Sidak et de Bonferroni-Holm. Ce dernier test est moins conservateur et plus puissant puisque, après chaque comparaison, le seuil est ajusté au nombre de comparaisons restantes

Exemple 7.3

Reprenez les données du tableau 7.1 et réalisez, à titre d'exemple, les comparaisons multiples de moyennes en utilisant la méthode de Bonferroni ($\alpha=0.05$) (On se rappelle que ce test n'est pas recommandé pour un grand nombre de comparaisons).

Solution 7.3

a) Méthode : comparaisons multiples de moyennes par la procédure de Bonferroni

b) Conditions d'application : (cf. exemple 7.1)

b) Calculs :

Risque d'erreur par comparaison : $\alpha' = 0.05/10 = 0.005$ et $t_{1-0.05/(2*10)} = 3.286$.

La valeur critique est : $v_c = 3.286 \sqrt{\frac{(2)(1.72633)}{4}} = 3.053$.

Après avoir rangé les moyennes par ordre décroissant et déclaré toute différence entre deux moyennes dépassant **3.053** comme significative, nous obtenons :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625

La conclusion est ici la même que dans le cas des comparaisons par la méthode de Tukey.

7.5.5. Méthode de Dunn-Sidak

Il s'agit d'une légère variante du test de Bonferroni proposée par Sidak (1967). Elle a le même principe, mais utilise l'inégalité :

$$\alpha_g \leq 1 - (1 - \alpha)^c$$

où c désigne toujours le nombre de comparaisons. Elle permet un très léger gain de puissance statistique, c'est-à-dire une légère augmentation de la capacité à déclarer un test significatif lorsque l'hypothèse nulle est fausse.

Pour se protéger, la méthode consiste à tester chacune des hypothèses élémentaires au niveau de signification $\alpha' = 1 - (1 - \alpha)^{1/c}$. Autrement dit, pour un ensemble de six comparaisons élémentaires et $\alpha=0.05$, elle utilise un risque de $\alpha'=0.0085$.

Une différence entre deux moyennes est donc déclarée significative si elle égale ou dépasse la valeur critique suivante :

$$v_c = t_{1-\alpha'/2}^{(v)} \sqrt{\frac{2CM_e}{n}},$$

qui utilise la distribution t de Student.

Ce test est très conservateur dans la mesure où le risque global d'erreur réel α_g est plus petit que le niveau fixé, mais il est relativement moins conservateur que le test de **Bonferroni**.

Exemple 7.4

Reprenez les données du tableau 7.1 et réalisez les comparaisons multiples de moyennes en utilisant la méthode de Dunn-Sidak ?

Solution 7.4

Risque d'erreur par comparaison : $\alpha' = 1 - (1 - 0.05)^{1/10} = 0.005116$.

La valeur critique est :

$$v_c = t_{0.9974} \sqrt{\frac{(2)(1.72633)}{4}} = (3.27)(0.929) = 3.04$$

On vérifie que les groupes homogènes sont les mêmes que dans le cas de la méthode de Bonferroni.

7.6. MÉTHODES DE COMPARAISONS DES MOYENNES PAR GROUPES**7.6.1. Introduction**

Les méthodes de comparaisons des moyennes par groupes, ou méthodes séquentielles, ont pour but de comparer des sous-ensembles des p moyennes en utilisant les hypothèses de formes :

$$H_0^s : \mu_i = \mu_j = \dots = \mu_s \quad \text{avec } 2 \leq s \leq p$$

où s est le nombre de moyennes du groupe, appelé aussi nombre d'échelons entre les moyennes à comparer, p étant comme toujours le nombre total de moyennes.

Les moyennes sont toujours rangées préalablement par ordre croissant. Si deux moyennes sont adjacentes, on dit qu'elles sont séparées par deux échelons, et si deux moyennes sont séparées par une troisième, on dit qu'elles sont séparées par trois échelons et ainsi de suite.

Le principe de ces méthodes consiste à rejeter l'hypothèse d'homogénéité des s moyennes du groupe lorsque l'amplitude observée de ce groupe dépasse l'amplitude critique suivante :

$$A_c = \delta_{\alpha_s}^{(p, \nu)} \sqrt{\frac{CM_e}{n}}$$

dans laquelle la valeur de la distribution utilisée δ est donnée dans des tables statistiques en fonction du nombre d'échelons (s), du nombre de degrés de liberté (ν) et aussi du niveau de signification α_s qui peut dépendre à son tour de s pour certaines méthodes. On constate donc que, contrairement aux méthodes de comparaisons deux à deux (cf. paragraphe 7.5), la valeur de l'amplitude critique dépend ici du nombre s de moyennes du groupe.

Pour éviter des conclusions aberrantes, les comparaisons se réalisent de manière séquentielle, en adoptant la procédure suivante :

- On range les p moyennes par ordre croissant ;
- On commence par comparer les moyennes du groupe limité par la plus grande moyenne et la plus petite moyenne ;
- Si l'hypothèse précédente est rejetée, on passe au test des deux hypothèses dont chacune concerne les $s-1$ moyennes les plus extrêmes ;

- Si un, au moins, des tests des hypothèses précédentes est significatif, on continue les tests en adoptant les règles suivantes :
 - si une hypothèse concernant un groupe de moyennes est rejetée, on teste les hypothèses qui la composent en commençant toujours par les hypothèses qui font intervenir le plus grand nombre de moyennes ;
 - si une hypothèse concernant un groupe de moyennes est acceptée, on considère ce groupe comme homogène et on n'effectue pas d'autres tests sur les hypothèses qui la composent.

La manière de calculer la valeur critique donne lieu à plusieurs méthodes de comparaison par groupes. Nous verrons dans les paragraphes 7.6.2 et 7.6.3 deux d'entre-elles, la méthode de Newman-Keuls et la méthode de Duncan.

7.6.2. Méthode de Newman-Keuls

La méthode de Newman-Keuls, appelée aussi méthode de Student-Newman-Keuls (SNK) ou de la plus petite amplitude significative, a été proposée par Newman [1939] puis indépendamment par Keuls [1952]. Elle ressemble beaucoup à la méthode de Tukey dans la mesure où elles utilisent la statistique des écarts studentisés, mais la méthode de Newman-Keuls utilise différentes valeurs critiques pour différentes paires de comparaisons de moyennes. Elle s'exécute en effet de manière séquentielle en utilisant des valeurs critiques qui dépendent du nombre s de moyennes du groupe dont on souhaite tester l'homogénéité.

Selon cette méthode, une hypothèse concernant l'homogénéité d'un groupe de moyennes séparées par s échelons est rejetée, au niveau de signification α , si l'amplitude observée de ce groupe (la différence entre la plus grande moyenne et la plus petite moyenne du groupe $|\bar{y}_i - \bar{y}_j|$) égale ou dépasse l'amplitude suivante :

$$v_c = q_{1-\alpha}^{(s,v)} \sqrt{\frac{CM_e}{n}}.$$

Cette valeur est aussi connue par la plus petite amplitude significative, elle doit être calculée pour chacune des valeurs du nombre d'échelons s et du nombre de degrés de liberté v de CM_e .

Lorsqu'on teste l'hypothèse globale qui fait intervenir toutes les moyennes, c'est-à-dire lorsque $s=p$, les méthodes de Newman-Keuls et de Tukey sont équivalentes. Mais, comme la valeur critique diminue avec la diminution du nombre d'échelons, le test de Newman-Keuls est moins conservateur en déclarant plus de différences, mais plus puissant que la méthode de Tukey. Le test de Tukey a l'avantage de maintenir le niveau de l'erreur de type I égal au niveau α choisi et permet le calcul d'intervalles de confiance pour les différences entre les moyennes, ce qui n'est pas le cas pour le test de Newman-Keuls.

Bien que la procédure de Newman-Keuls ait en effet été conçue pour contrôler le risque global d'erreur, elle ne le contrôle pas complètement et, sous certaines conditions, le taux d'erreur peut être assez élevé. Elle est parfois qualifiée de méthode libérale ou de modérément libérale, mais moins libérale que celles de Duncan et de la PPDS. D'autres auteurs soulignent qu'elle assure un certain équilibre entre les deux risques d'erreur, mais elle reste tout de même une méthode très controversée.

Une version légèrement modifiée de l'amplitude pour être utilisée afin de tenir compte de la non égalité des effectifs des échantillons est donnée par :

$$v_c = q_{1-\alpha}^{(s,v)} \sqrt{\frac{CM_e}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

Exemple 7.5

Reprenez les données du tableau 7.1 et réalisez les comparaisons multiples de moyennes en utilisant la procédure de Newman-Keuls (SNK).

Solution 7.5

a) Méthode : comparaisons multiples de moyennes par la procédure de *Newman-Keuls*

b) Conditions d'application : (cf. exemple 7.1)

c) Calculs :

On calcule la valeur critique (v_c) en fonction du nombre de moyennes du groupe (s) (tableau 7.3).

Tableau 7.3. Valeurs théoriques et critiques de la méthode de SNK en fonction du nombre de moyennes du groupe.

s	2	3	4	5
$q_{1-\alpha}^{(s,v)}$	3.01	3.67	4.08	4.37
v_c	1.98	2.41	2.68	2.87

Remarquons que nous retrouvons la valeur critique de la PPDS pour les groupes de deux moyennes et celle de Tukey pour les groupes de cinq moyennes.

On examine successivement les différents groupes après avoir rangé les moyennes par ordre croissant.

Groupe de 5 moyennes (il y a un seul) :

$|\bar{x}_1 - \bar{x}_5| = 4.325 > 2.87 \rightarrow$ le groupe des 5 moyennes ne peut donc être considéré comme homogène. Ce résultat est en accord avec la conclusion de l'ANOVA1.

Groupes de 4 moyennes (il y en a deux) :

$|\bar{x}_2 - \bar{x}_5| = 4.525 > 2.68 \rightarrow$ rejeter $H_0^4 : \mu_5 = \mu_3 = \mu_4 = \mu_2$

$|\bar{x}_1 - \bar{x}_3| = 2.225 < 2.68 \rightarrow$ ne pas rejeter $H_0^4 : \mu_3 = \mu_4 = \mu_2 = \mu_1$

Puisque cette dernière hypothèse n'est pas significative, nous n'allons pas tester les hypothèses qui la composent.

Groupes de 3 moyennes :

$|\bar{x}_4 - \bar{x}_5| = 2.2 < 2.41 \rightarrow$ ne pas rejeter $H_0^3 : \mu_5 = \mu_3 = \mu_4$

Il en résulte ce qui suit :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625

Pour cet exemple, le résultat des comparaisons multiples est le même que celui qui est trouvé en utilisant les méthodes de Tukey et de Bonferroni.

7.6.3. Méthode de Duncan

Duncan [1955] a développé une méthode similaire à celle de Newman-Keuls d'un point de vue application. Si deux moyennes sont séparées par s échelons, Duncan définit le niveau de protection comme étant $(1 - \alpha)^{s-1}$. Il en découle que la probabilité de rejeter erronément l'égalité de deux moyennes séparées par s échelons est donnée par $\alpha_s = 1 - (1 - \alpha)^{s-1}$.

On constate que le niveau de signification α_s dépend du nombre d'échelons s . Il est égal à α lorsque le groupe comporte deux moyennes ($s=2$) et il augmente avec l'augmentation de s , pour atteindre son maximum pour l'hypothèse globale, c'est-à-dire lorsque s est égal à p . Ce niveau de signification élevé pour l'hypothèse globale fait que celle-ci peut être rejetée par la méthode de Duncan alors qu'elle peut être, en même temps, acceptée par l'analyse de la variance. Il faut aussi signaler que le niveau de signification maximum de la méthode de Duncan augmente avec l'augmentation du nombre total de moyennes p de l'étude.

Selon la méthode de Duncan, l'hypothèse d'homogénéité d'un groupe de s moyennes est rejetée si la différence entre ses moyennes extrêmes égale ou dépasse la valeur critique :

$$v_c = q_{1-\alpha}^{(s,v)} \sqrt{\frac{CM_e}{n}},$$

où les valeurs théoriques $q_{1-\alpha}^{(s,v)}$ sont obtenues à partir de tables statistiques en fonction de α , de s et du nombre de degrés de liberté v de CM_e . le dénominateur n désigne toujours l'effectif de chaque échantillon.

D'autre part, il faut retenir que des études basées sur des simulations ont montré que la méthode de Duncan, bien que libérale, reste toutefois, dans l'ensemble, moins libérale que la méthode PPDS. De même, il faut noter que ce test est assez populaire auprès des expérimentateurs en raison de sa grande puissance : il y a une grande probabilité de déclarer une différence lorsqu'il y a réellement une différence entre les moyennes.

Exemple 7.6

Reprenez les données du tableau 7.1 et réalisez les comparaisons multiples de moyennes en utilisant la procédure de Duncan.

Solution 7.6

a) Méthode : comparaisons multiples de moyennes par la procédure de Duncan

b) Conditions d'application : (cf. exemple 7.1)

c) Calculs :

On commence par calculer les valeurs critiques (v_c) en fonction du nombre de moyennes à comparer du groupe (tableau 7.4).

Tableau 7.4. Valeurs théoriques et critiques de la méthode de Duncan en fonction du nombre s de moyennes du groupe.

s	2	3	4	5
$q_{0.95}(s, v)$	3.01	3.16	3.25	3.31
v_c	1.98	2.076	2.135	2.176

On commence par ordonner les moyennes par ordre croissant et on examine successivement les différents groupes de moyennes :

Groupe de 5 moyennes (il y a un seul) :

$|\bar{x}_1 - \bar{x}_5| = 4.325 > 2.176 \rightarrow$ rejeter $H_0^5 : \mu_5 = \mu_3 = \mu_4 = \mu_2 = \mu_1$. Le groupe des 5 moyennes ne peut donc être considéré comme homogène. Ce résultat est, dans cet exemple, en accord avec la conclusion de l'ANOVA1.

Groupes de 4 moyennes (il y en a deux):

$|\bar{x}_2 - \bar{x}_5| = 4.525 > 2.135 \rightarrow$ rejeter $H_0^4 : \mu_5 = \mu_3 = \mu_4 = \mu_2$

$|\bar{x}_1 - \bar{x}_3| = 2.225 > 2.135 \rightarrow$ rejeter $H_0^4 : \mu_3 = \mu_4 = \mu_2 = \mu_1$

Groupes de 3 moyennes :

$|\bar{x}_1 - \bar{x}_4| = 2.125 > 2.076 \rightarrow$ rejeter $H_0^3 : \mu_4 = \mu_2 = \mu_1$

$|\bar{x}_2 - \bar{x}_3| = 2.125 > 2.076 \rightarrow$ rejeter $H_0^3 : \mu_3 = \mu_4 = \mu_2$

$|\bar{x}_4 - \bar{x}_5| = 2.200 > 2.076 \rightarrow$ rejeter $H_0^3 : \mu_5 = \mu_3 = \mu_4$

Groupes de 2 moyennes :

$|\bar{x}_1 - \bar{x}_2| = 0.100 < 1.98 \rightarrow$ ne pas rejeter $H_0^2 : \mu_2 = \mu_1$

$|\bar{x}_2 - \bar{x}_4| = 2.025 > 1.98 \rightarrow$ rejeter $H_0^2 : \mu_2 = \mu_4$

$|\bar{x}_4 - \bar{x}_3| = 0.100 < 1.98 \rightarrow$ ne pas rejeter $H_0^2 : \mu_4 = \mu_3$

$|\bar{x}_3 - \bar{x}_5| = 2.100 > 1.98 \rightarrow$ rejeter $H_0^2 : \mu_3 = \mu_5$

Il en résulte ce qui suit :

Race5	Race3	Race4	Race2	Race1
80.300	82.400	82.500	84.525	84.625

La conclusion est la même que celle qui découle en utilisant la méthode de la PPDS.

7.7. MÉTHODE DE DUNNETT

Le test de Dunnett s'utilise lorsqu'il s'agit de comparer un traitement, souvent appelé traitement témoin ou contrôle, à un ensemble de traitements. Ces derniers ne sont pas comparés entre eux.

Le traitement témoin peut être une variété végétale utilisée dans la région, une race animale locale, un placebo, etc.

Si le nombre total de traitements étudiés, en comptant le témoin, est p , le test de Dunnett consiste à réaliser $p - 1$ comparaisons. Cela revient à tester la signification des $p - 1$ hypothèses suivantes :

$$H_0 : \mu_i = \mu_{\text{témoin}} \quad \text{où } i = 1, 2, \dots, p - 1.$$

contre les hypothèses $H_1 : \mu_i \neq \mu_{\text{témoin}}$, dans lesquelles le témoin est le p -ième traitement.

Le principe du test consiste à calculer la différence, en valeur absolue, entre la moyenne observée du traitement i et celle du témoin :

$$A_o = |\bar{y}_i - \bar{y}_{\text{témoin}}|$$

et à considérer le test comme significatif chaque fois que cette différence égale ou dépasse la valeur critique suivante :

$$v_c = d_{1-\alpha/2}^{(p-1, v)} \sqrt{\frac{2CM_e}{n}}$$

dans laquelle les valeurs de $d_{1-\alpha/2}^{(p-1, v)}$ sont lues dans des tables en fonction du nombre de degrés de liberté v , du niveau de signification α et du nombre de comparaisons $(p-1)$. Ces valeurs sont calculées de telle sorte que le risque global d'erreur soit maintenu au niveau α pour les $p-1$ tests. Le terme n désigne l'effectif de chacun des échantillons.

Exemple 7.7

Reprenez les données du tableau 7.1 en supposant que la race 5 constitue un témoin, quelles sont les races qui diffèrent significativement de ce témoin ?

Solution 7.7

a) Méthode : comparer quatre moyennes à un témoin par la méthode de Dunnett

b) Conditions d'application : (cf. exemple 7.1)

c) Calculs :

La valeur critique est : $v_c = 2.73 \sqrt{\frac{(2)(1.72633)}{4}}$ ou 2.536

Comparons cette valeur aux différentes amplitudes observées :

$$|\bar{x}_1 - \bar{x}_5| = 4.325 > 2.536 ; |\bar{x}_2 - \bar{x}_5| = 4.225 > 2.536 ; |\bar{x}_4 - \bar{x}_5| = 2.200 < 2.536 \text{ et } |\bar{x}_3 - \bar{x}_5| = 2.100 < 2.536.$$

Il ressort que les races 1 et 2 ont un pourcentage moyen de rendement significativement supérieur à celui du témoin.

7.8. CONCLUSION

Si un facteur fixe s'avère avoir un effet significatif suite à l'utilisation de l'ANOVA, on peut conclure qu'au moins les moyennes de deux modalités de ce facteur ne sont pas égales. On n'est pas renseigné sur les comparaisons possibles entre les moyennes de deux ou plusieurs modalités de ce facteur. Pour y arriver, il faut un recours aux méthodes de comparaisons multiples de moyennes. Toutefois, le choix de la méthode de comparaisons multiples convenant à une situation donnée ne devrait pas être chose facile.

Il faut se rappeler que lorsqu'on réalise un ensemble de comparaisons entre plusieurs moyennes, le risque global d'erreur de première espèce, qui est la probabilité de commettre au moins une erreur de première espèce parmi toutes les comparaisons en question, augmente considérablement avec l'augmentation du nombre de comparaisons. Ce sont d'ailleurs les

différentes approches adoptées pour contrôler ce risque d'erreur qui ont fait que les méthodes de comparaisons proposées dans la littérature sont nombreuses. On ne peut en effet se contenter de calculer le risque d'erreur d'une comparaison individuelle comme si elle était seule, il faut minimiser le risque d'erreur de l'ensemble des comparaisons.

1° Si on a des hypothèses sur les moyennes qui ont été formulées préalablement à la collecte des données sur la base de considérations théoriques ou d'expériences précédentes, l'expérimentateur s'intéresse souvent à quelques comparaisons dites *a priori* qui correspondent à des questions précises portant sur les moyennes. Généralement, on réalise les comparaisons planifiées au moyen d'une analyse par contraste : il s'agit de poser quelques questions portant sur des combinaisons linéaires particulières des moyennes théoriques des modalités d'un facteur fixe (sujet non traité dans ce chapitre). Dans ce premier cas, où le nombre de comparaisons est habituellement bien plus petit que le nombre de comparaisons possibles, la probabilité de commettre au moins une erreur de type I est plus petite que si les comparaisons n'interviennent qu'*a posteriori*. Selon les cas, les hypothèses sont testées en utilisant le test t de Student, le test F de Fisher, le test de Dunn-Bonferroni, le test de Bonferroni-Holm qui est moins conservatif et plus puissant puisque le seuil est ajusté au nombre de comparaisons restantes après chaque comparaison ou par le test de Dunnett.

L'inconvénient des méthodes *a priori* réside dans le risque de ne pas tester des éventuelles différences intéressantes survenues à des endroits non prévues, mais les comparaisons présentent l'avantage de la puissance.

2° Par contre, si les hypothèses à tester ne sont claires qu'à l'issue de la collecte et l'examen des données et que l'on compte réaliser un grand nombre de comparaisons, on opte pour des tests *post-hoc*. Pour des comparaisons qui ne sont pas par paires de moyennes, on peut utiliser le test de Scheffe (1953), mais habituellement il s'agit de procéder à toutes les comparaisons possibles deux à deux des moyennes.

Les chercheurs ont l'habitude de recourir à l'un de ces tests *post-hoc* après avoir rejeté l'hypothèse nulle d'égalité des moyennes en utilisant l'ANOVA. Toutefois, le test F de l'ANOVA et les tests *post-hoc* utilisent différentes méthodes pour déterminer la signification et ils peuvent conduire à des conclusions différentes. Ces différences se produisent généralement dans des situations particulières et, dans ces cas, il est possible de signaler les résultats *post-hoc* significatifs.

Les tests *post-hoc* sont nombreux et il n'y a pas de consensus universel sur le meilleur à utiliser. Les méthodes les plus efficaces sont celles qui tentent d'assurer le meilleur équilibre entre le risque global d'erreur et la puissance. C'est-à-dire celles qui minimisent le risque de commettre une erreur de première espèce tout en gardant une probabilité raisonnable de déceler une vraie différence. Certains de ces tests sont assez conservateurs dans la mesure où ils ont tendance à rejeter les hypothèses nulles d'égalité des moyennes plus difficilement que prévu théoriquement. Ce conservatisme s'accompagne de la perte de puissance, c'est-à-dire que ces tests sont moins susceptibles de détecter une différence entre les moyennes des groupes qui existe réellement.

Pour obtenir un risque d'erreur global de première espèce spécifié, les procédures *post-hoc* abaissent le niveau de signification de toutes les comparaisons individuelles. Par exemple, pour aboutir à un risque d'erreur global de première espèce de 5% pour un ensemble de comparaisons, la procédure utilise un niveau de signification individuelle beaucoup plus faible.

Dans le cas où l'on opte pour la réalisation de toutes les comparaisons par paires, qui compare toutes les associations de groupes possibles, on a souvent tendance à conseiller le test de Tukey

ou encore le test REGWQ (Ryan-Einot-Gabriel-Welsch). La méthode de Tukey est la plus courante, elle a l'avantage de maintenir le niveau de l'erreur de type I égal au niveau α choisi et permet le calcul d'intervalles de confiance pour les différences entre les moyennes. Il convient cependant d'éviter de réaliser les comparaisons par des tests qualifiés de libéraux, car ils conduisent à un risque global de première espèce assez élevé. Dans ce sens, on peut citer, comme exemples, le test de la plus petite différence significative qualifié de laxiste sauf dans le cas de trois moyennes pour autant que le test F de l'ANOVA soit significatif et, dans une moindre mesure, le test de Duncan. Ceux-ci risquent de ressortir des différences qui, en réalité, n'existent pas. Quant au test de Newman-Keuls, on le classe tantôt parmi les tests libéraux, tantôt parmi les tests intermédiaires à risque global d'erreur modéré. Pour certains auteurs, le test de Newman-Keuls constitue le compromis, alors que d'autres le déconseillent en estimant qu'il donne lieu à une certaine inflation du risque global d'erreur.

Dans tous les cas, si on a à utiliser un test *post-hoc*, il est recommandé de définir la méthode de comparaisons multiples de moyennes dès le départ, car il est déconseillé d'essayer différentes méthodes et choisir celle qui produit les résultats voulus.

Par ailleurs, il faut dire que le chercheur n'a pas toujours besoin de toutes les comparaisons par paires de moyennes, mais d'un sous-ensemble. Cette réduction permet à la procédure d'utiliser un risque d'erreur par comparaison plus élevé pour atteindre le risque d'erreur global spécifié, ce qui augmente la puissance statistique. Ainsi, il se peut que l'expérience soit planifiée pour comparer un groupe de contrôle à $p-1$ autres groupes. Dans ce cas, on utilise le test de Dunnett qui ne nécessite que $p-1$ comparaisons, car on n'a pas besoin de comparer les groupes les uns aux autres. De même, si l'objectif de l'étude est d'identifier le ou les groupes les meilleurs, on n'aura peut-être pas besoin aussi dans ce cas de comparer tous les groupes possibles et on peut recourir à des méthodes telles que la méthode de Gupta.

L'expérimentateur utilisant ces méthodes doit être aussi attentif quant à l'interprétation des résultats lorsqu'il y a des chevauchements entre les groupes homogènes de moyennes.

CHAPITRE 8

TESTS D'AJUSTEMENT

8.1. INTRODUCTION

Lorsqu'on dispose d'un échantillon aléatoire et simple d'unités classées ou non en plusieurs modalités en fonction d'une variable donnée, il est assez souvent question de savoir si une loi théorique de probabilité peut représenter au mieux cette distribution de fréquences ou cette série de valeurs prises par cette variable. Le type de lois, dont il question ici, est en général suggéré par la nature du phénomène étudié.

Il s'agit en effet, dans ce cas, de problèmes de tests d'ajustement qui permettent de tester l'adéquation entre une distribution expérimentale, qui est obtenue à partir de l'échantillon, et la distribution de la loi de probabilité servant de modèle théorique.

Nous nous souvenons en particulier que nous étions amenés, à maintes reprises, notamment lorsque nous avons cherché à tester l'homogénéité des variances ou l'égalité des moyennes, à supposer la normalité des populations, ou des résidus, sans procéder à des tests. C'est pourquoi, on trouve, en particulier, dans la littérature statistique de nombreux tests consacrés à la normalité d'une distribution.

Nous présenterons, dans ce chapitre, certains des tests d'ajustement des plus connus par les expérimentateurs. Nous exposerons d'abord le test khi-deux de Pearson qui s'applique aux distributions discrètes et continues. Nous distinguerons entre le cas de lois complètement définies où le test s'applique sans estimation de paramètres statistiques de la distribution de la population, et le cas de lois incomplètement définies où l'on est amené à estimer des paramètres statistiques dont dépend la distribution de la population à partir de l'échantillon (paragraphe 8.2). Ce test est construit à partir du regroupement des observations dans des classes.

Nous présenterons ensuite le test de Kolmogorov-Smirnov qui s'applique aux distributions continues et qui est basé sur la comparaison de la fonction cumulative de fréquences de l'échantillon avec la fonction de répartition de la population (paragraphe 8.3).

Au paragraphe 8.4, nous exposerons deux tests particuliers assez utilisés par les biologistes. Le premier est le test de l'indice de dispersion pour la distribution de Poisson, le second est le test de Shapiro-Wilk pour la distribution normale.

8.2. LE TEST D'AJUSTEMENT KHI-DEUX DE PEARSON

8.2.1. Cas de lois complètement définies

8.2.1.1. Exemple introductif

On a observé les couleurs d'un échantillon aléatoire et simple de 886 fleurs d'une certaine sorte de pois et on a trouvé 237 qui sont rouges, 465 qui sont roses et 184 qui sont blanches. Peut-on rejeter l'hypothèse nulle de Mendel, au niveau de signification $\alpha=5\%$, selon lequel cette sorte de pois produit les fleurs rouges, roses et blanches dans des proportions à long terme de 25%, 50% et 25%.

8.2.1.2. Réalisation du test

D'une manière générale, soit une population pratiquement infinie dont les individus sont classés en p modalités a_1, a_2, \dots, a_p d'un caractère A qualitatif. Il peut aussi s'agir d'un caractère quantitatif à partir duquel on a établi des modalités. A chaque modalité a_i du caractère A correspond une probabilité P_i . On souhaite tester l'hypothèse nulle suivante :

H_0 : la population possède la distribution de probabilité de la forme supposée,
contre **H_1 : la population ne possède pas la distribution de probabilité de la forme supposée.**

Pour tester cette hypothèse, on prélève dans cette population un échantillon aléatoire et simple de n individus que l'on classe selon le caractère qualitatif A . Soit n_i la fréquence observée pour la modalité a_i . La fréquence théorique, dite aussi attendue ou calculée, de la modalité a_i est donnée par nP_i . Il s'agit de la fréquence qui serait observée pour la modalité a_i si l'hypothèse nulle était vraie. Le test revient donc à mesurer en quelques sortes les écarts entre les effectifs observés n_i et les effectifs théoriques nP_i ($i=1, \dots, p$).

Pour l'exemple introductif des fleurs de pois, on a :

1) Si H_0 est vraie, les effectifs théoriques seront :

$$\begin{aligned}(0.25)(886) &= 221.5 && \text{pour les fleurs rouges} \\ (0.50)(886) &= 443.0 && \text{pour les fleurs roses} \\ (0.25)(886) &= 221.5 && \text{pour les fleurs blanches}\end{aligned}$$

2) Les écarts entre les effectifs observés et les effectifs théoriques sont :

$$\begin{aligned}237 - 221.5 &= 15.5 && \text{pour les fleurs rouges} \\ 465 - 443.0 &= 22.0 && \text{pour les fleurs roses} \\ 184 - 221.5 &= -37.5 && \text{pour les fleurs blanches}\end{aligned}$$

3) Pour avoir une idée sur l'importance des écarts, on ne somme pas directement les écarts puisque cette somme est nulle, mais on calcule d'abord le carré de chaque écart :

$$(n_i - nP_i)^2$$

puis on le pondère par l'effectif théorique pour obtenir son importance relative (c'est-à-dire pour maintenir les écarts en proportion). Enfin, on somme les contributions des différentes modalités.

On obtient la statistique suivante :

$$\chi_{obs}^2 = \sum_{i=1}^p \frac{(n_i - nP_i)^2}{nP_i} \quad \text{ou} \quad \boxed{\chi_{obs}^2 = \sum_{i=1}^p \frac{n_i^2}{nP_i} - n},$$

qui suit approximativement, lorsque l'effectif n est suffisamment grand, une distribution khi-deux à $k=p-1$ degrés de liberté.

L'approximation n'est généralement considérée comme satisfaisante que lorsque les effectifs théoriques nP_i sont supérieurs ou égaux à 5. Certains auteurs estiment qu'on peut tolérer une classe de A avec un effectif théorique inférieur à 5 mais supérieur à 1. Il est aussi possible de regrouper certaines classes contiguës pour obtenir des quantités nP_i supérieures ou égales à 5.

4) Plus la quantité χ_{obs}^2 est grande et plus il y a désaccord entre les fréquences observées et celles attendues selon l'hypothèse nulle. L'hypothèse nulle est rejetée lorsque :

$$\chi_{obs}^2 \geq \chi_{1-\alpha}^2$$

avec $k=p-1$ degrés de liberté où p est le nombre de classes après un éventuel regroupement.

Exemple 8.1

Reprenez l'exemple des couleurs des fleurs de pois et testez si l'hypothèse de Mendel peut être rejetée au niveau de signification 5% .

Solution 8.1

a) **Méthode** : Test khi-carré d'ajustement à une loi complètement définie

b) **Conditions d'application**

- La population est infinie ;
- L'échantillon des 886 fleurs est prélevé de manière aléatoire et simple dans la population ;
- Les effectifs attendus sont supérieurs ou égaux à 5 (à vérifier).

c) **Hypothèses**

H_0 : l'hypothèse de Mendel est vérifiée

contre H_1 : l'hypothèse de Mendel n'est pas vérifiée

d) **Calcul**

Le tableau 8.1 donne les principaux paramètres pour obtenir la valeur observée du khi-carré.

Tableau 8.1. Principaux paramètres pour obtenir la valeur observée du khi-carré.

Classe	n_i	P_i	nP_i	$n_i - nP_i$	$(n_i - nP_i)^2 / nP$
Fleurs rouges	237	0.25	221.5	+15.5	1.085
Fleurs roses	465	0.50	443.0	+22.0	1.093
Fleurs blanches	184	0.25	221.5	-37.5	6.349
Total	886	1.00	886	0	8.526

On a $\chi_{obs}^2 = 8.53$ et $\chi_{0.95}^2 = 5.99$ avec 2 degrés de liberté. Puisque la valeur observée est supérieure à la valeur critique, l'hypothèse nulle doit être rejetée au niveau $\alpha = 0.05$.

En prenant un risque de 5%, on peut dire que l'hypothèse de Mendel n'est pas vérifiée.

8.2.2. Cas de lois incomplètement définies

8.2.2.1. Exemples introductifs

Deux exemples seront traités dans ce cas de lois incomplètement spécifiées.

1) Le premier concerne le nombre de certaines taches particulières présentes sur les poumons des taureaux d'une région donnée (données IAV Hassan II). On souhaite savoir si le nombre de ces taches se distribue selon la loi de Poisson, en faisant des observations sur un échantillon de 70 animaux dans les abattoirs de la région (tableau 8.2).

Tableau 8.2. Répartition des taureaux en fonction du nombre de taches détectées dans les poumons.

Nombre de taches	Nombre de taureaux
0	5
1	14
2	17
3	14
4	10
5	7
6	3
Total	70

2) Le second exemple concerne les données d'un échantillon de 429 caisses de farine remplies par une machine en bon état de fonctionnement (données IAV Hassan II). En fonction des données regroupées dans la distribution de fréquences présentée par le tableau 8.3, on souhaite savoir si l'on peut dire que le remplissage se fait selon une loi normale ($\alpha=0.05$).

Tableau 8.3. Répartition de 429 caisses de farine en fonction de leur poids (en kg).

Classes de poids (kg)	Effectifs
48.5-48.8	1
48.9-49.2	12
49.3-49.6	26
49.7-50.0	64
50.1-50.4	83
50.5-50.8	96
50.9-51.2	62
51.3-51.6	55
51.7-52.0	21
52.1-52.4	8
52.5-52.8	1
Total	429

8.2.2.2. Réalisation du test

Dans le cas de ces deux exemples, comme dans bien d'autres cas semblables, il arrive que la loi de probabilité à tester ne soit pas complètement définie, c'est-à-dire que ses paramètres ne sont pas choisis *a priori*. Il s'avère nécessaire, dans ces conditions, de commencer par obtenir des estimations des paramètres inconnus de cette loi à partir des données de l'échantillon.

La "distance" du khi-deux s'exprime toujours sous forme d'une moyenne pondérée des d'écarts quadratiques entre les fréquences observées (n_i) et les fréquences théoriques estimées $n\hat{p}_i$:

$$\chi_{obs}^2 = \sum_{i=1}^p \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \quad \text{ou} \quad \chi_{obs}^2 = \sum_{i=1}^p \frac{n_i^2}{n\hat{p}_i} - n$$

qui suit approximativement, lorsque l'effectif **n** est suffisamment grand, une distribution khi-deux à **k=p-1-r** degrés de liberté, où **p** est le nombre de classes après un éventuel regroupement et **r** est le nombre de paramètres estimés.

L'hypothèse nulle est rejetée lorsque :

$$\chi_{obs}^2 \geq \chi_{1-\alpha}^2$$

avec **k=p-1-r** degrés de liberté.

La loi du khi-deux dépend du nombre de classes et de la méthode d'estimation des paramètres. Ces estimations s'obtiennent habituellement par la méthode du khi-deux minimum, qui rendent la valeur de khi-deux aussi petite que possible. Lorsque les effectifs sont assez grands, on retrouve généralement les expressions obtenues par la méthode du maximum de vraisemblance. Dans les cas, par exemple, des distributions de Poisson et de la normale, nous utiliserons la moyenne observée pour estimer la moyenne théorique. De même, nous utiliserons, dans le cas de la distribution normale, la variance observée pour estimer la variance théorique :

$$\hat{m} = \frac{1}{n} \sum_{i=1}^p n_i x_i \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2$$

Exemple 8.2

Reprenez les données du tableau 8.2 et réalisez un test d'ajustement à la distribution de Poisson ?

Solution 8.2

a) Méthode : Test khi-carré d'ajustement à une distribution de Poisson

b) Conditions d'application

- La population est pratiquement infinie ;
- L'échantillon des 70 taureaux est prélevé de manière aléatoire et simple dans la population ;
- Les effectifs attendus sont supérieurs ou égaux à 5 (à vérifier).

c) Hypothèses

H_0 : la population du nombre de taches est distribuée selon une loi de Poisson

contre H_1 : la population du nombre de taches n'est pas distribuée selon une loi de Poisson

d) Calcul

- Rappelons que la loi de probabilité de Poisson dépend d'un seul paramètre qui est la moyenne :

$$P_i = \frac{e^{-\mu} \mu^i}{i!} \quad (i \geq 0)$$

On commence par estimer le paramètre inconnu μ :

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^7 n_i x_i = 2.614 \text{ taches/ taureau;}$$

- On obtient ensuite les probabilités estimées et on complète le tableau 8.4 pour obtenir la valeur observée du khi-deux

Tableau 8.4. Principaux paramètres permettant le calcul du khi-deux observé dans le cas d'ajustement à une loi de Poisson.

Classe	n_i	\hat{P}_i	$n\hat{P}_i$	$n_i^2 / n\hat{P}_i$
0	5	0.0732	5.125	5.125
1	14	0.1914	13.399	13.399
2	17	0.2502	17.515	17.515
3	14	0.2180	15.263	15.263
4	10	0.1425	9.975	9.975
5	7	0.0745	5.216	5.216
6	3	0.0325	2.273	2.273
>6	0	0.0176	1.234	1.234
Total	70	1.0000	70.000	70.833

- Nous avons toléré une classe avec une fréquence attendue inférieure à 5 ;
- Puisqu'on a estimé un seul paramètre, qui est la moyenne, la valeur théorique du khi-deux est calculée avec 5 degrés de liberté (il s'agit de $7-1-1=5$, car il y a 7 classes après regroupement). On a :

$$\chi_{obs}^2 = 70.833 - 70 = 0.833 \quad \text{et} \quad \chi_{0.95}^2 = 11.1.$$

En prenant un risque de 5%, il n'y a pas suffisamment de preuve pour rejeter l'hypothèse nulle. Ce résultat est attendu dès le calcul des fréquences attendues, car celles-ci sont très proches des fréquences observées.

Exemple 8.3

Reprenez les données du tableau 8.3 et dites si l'on peut dire que le remplissage des caisses se fait selon une loi normale ($\alpha=0.05$) ?

Solution 8.3

a) Méthode : test khi-carré d'ajustement à une distribution de Gauss

b) Conditions d'application

- La population est pratiquement infinie ;
- L'échantillon des 429 caisses est prélevé de manière aléatoire et simple dans la population ;
- Les effectifs attendus sont supérieurs ou égaux à 5 (à vérifier).

c) Hypothèses

H_0 : le poids des casses est distribuée selon une loi de Gauss

contre H_1 : le poids des casses n'est pas distribuée selon une loi de Gauss

d) Calcul

- La loi de probabilité de la distribution normale dépend de deux paramètres : la moyenne et l'écart-type. Ils sont obtenus, pour le cas de cette distribution, par :

$$\bar{x} = \frac{1}{429} \sum_{i=1}^{11} n_i x_i = 50.585 \text{ kg/caisse}$$

et

$$s = \sqrt{\frac{1}{n} \left[\sum_{i=1}^{11} n_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{11} n_i x_i \right)^2 \right]} = 0.723 \text{ kg/caisse}$$

où x_i est le point central de la classe i (moyenne des deux limites de la classe).

- Le calcul des probabilités, et donc des fréquences attendues, passe par la détermination des valeurs réduites :

$$u_i = (x_i' - 50.585) / 0.723$$

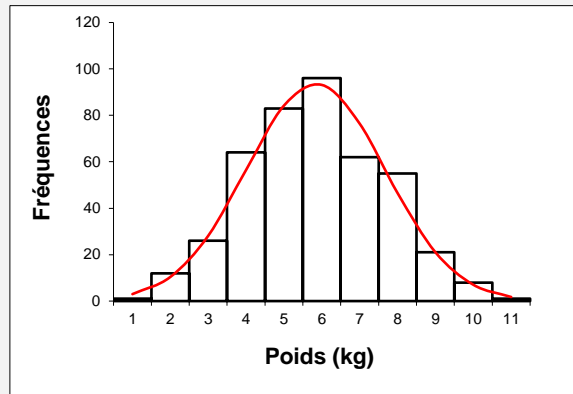
où x_i' est la limite supérieure de la classe i . Pour la deuxième classe, par exemple, on obtient :

$$u_2 = (49.25 - 50.585) / 0.723 = -1.8465.$$

Le tableau 8.5 donne le détail des calculs et la figure ci-dessous représente la distribution normale ajustée.

Tableau 8.5. Principaux paramètres permettant le calcul du khi-deux observé pour l'ajustement à la distribution normale.

Classe	n_i	u	$\varphi(u)$	\hat{P}_i	$n\hat{P}_i$	
<48.4	0	-2.9530	0.0016	0.0016	0.68	3.519
48.5-48.8	1	-2.3997	0.0082	0.0066	2.84	
48.9-49.2	12	-1.8465	0.0324	0.0242	10.385	
49.3-49.6	26	-1.2932	0.0980	0.0656	28.123	
49.7-50.0	64	-0.7400	0.2297	0.1317	56.496	
50.1-50.4	83	-0.1867	0.4259	0.1963	84.205	
50.5-50.8	96	0.3665	0.6430	0.2171	93.125	
50.9-51.2	62	0.9198	0.8212	0.1781	76.423	
51.3-51.6	55	1.4730	0.9296	0.1085	46.535	
51.7-52.0	21	2.0263	0.9786	0.0490	21.023	
52.1-52.4	8	2.5795	0.9951	0.0164	7.045	9.167
52.5-52.8	1	3.1328	0.9991	0.0041	1.751	
> 52.8	0		1.0000	0.0009	0.371	
Total	429	-	-	1.0000	429	



Pour satisfaire la condition de fréquences attendues d'au moins 5, nous avons regroupé les fréquences des deux premières classes et des trois dernières et nous avons toléré une classe de fréquence inférieure à 5. La valeur observée du khi-deux est:

$$\chi_{obs}^2 = \frac{1^2}{3.519} + \frac{12^2}{10.385} + \dots + \frac{9^2}{9.167} - 429 = 7.58.$$

Puisqu'on a estimé deux paramètres, la valeur théorique du khi-deux est calculée avec 7 (=10-1-2) degrés de liberté. On obtient la valeur $\chi_{0.95}^2 = 14.1$ qui est supérieure à la valeur observée.

Conclusion : En prenant un risque de 5%, on accepte l'hypothèse de la normalité de la population.

8.3. TEST DE KOLMOGOROV-SMIRNOV

Ce test d'ajustement est une approche non paramétrique permettant de tester si un échantillon peut être considéré comme extrait d'une population qui a une distribution théorique continue spécifiée. Il est basé sur la comparaison de la fonction cumulative de fréquences de l'échantillon N' (fonction de répartition observée) avec la fonction de répartition théorique F (cf. cours de la 2^{ème} année du cycle préparatoire de l'IAV).

Le test de Kolmogorov-Smirnov peut s'utiliser dans le cas de petits échantillons. Il s'applique lorsque la distribution est continue et complètement définie. Les paramètres de la distribution présumée ne sont pas estimés à partir de l'échantillon mais spécifiés à l'avance.

Si l'hypothèse nulle est vraie, les deux fonctions, empirique et théorique, sont proches. L'adéquation de la fonction N' à la fonction F est mesurée au moyen de la distance de Kolmogorov-Smirnov.

A chaque point x_i , on calcule la différence entre N' et F et on cherche l'écart maximum :

$$D_{obs} = \max_{(i=1, \dots, n)} |N'(x_i) - F(x_i)| = \max_{(i=1, \dots, n)} \left(\left| F(x_i) - \frac{i-1}{n} \right|; \left| \frac{i}{n} - F(x_i) \right| \right)$$

qui suit une loi de « Kolmogorov-Smirnov » de paramètre n . L'hypothèse nulle qui stipule que la variable suit la loi que l'on a fixée est rejetée lorsque :

$$D_{obs} \geq D_{1-\alpha}.$$

Les valeurs critiques sont fournies par des tables en fonction de n et le risque α de se tromper. Ces tables sont données en annexes pour des valeurs de n et α pour un test bilatéral.

Pour réaliser le test, on suit les étapes suivantes :

- 1) Trier les valeurs observées x_i par ordre croissant ;
- 2) Calculer $N'_{i-1} = P(X < x_i) = (i - 1)/n$ (les valeurs inférieures de la distribution observée) ;
- 3) Calculer les probabilités $F(x_i)$;
- 4) Calculer les écarts $|F(x_i) - (i - 1)/n|$;
- 5) Calculer $N'_i = P(X \leq x_i) = i/n$ (les valeurs supérieures de la distribution observée) ;
- 6) Calculer les écarts $|i/n - F(x_i)|$;
- 7) Retenir le plus grand écart mesurant la différence entre les distributions empirique et théorique ;
- 8) Rejeter l'hypothèse nulle lorsque $D_{obs} \geq D_{1-\alpha}$.

Remarques :

- 1) En plus de l'adéquation à une loi de distribution fixée, ce test peut également être utilisé pour comparer deux distributions, en vérifiant si leurs fonctions de répartition sont similaires ;
- 2) L'ajustement à une distribution de Gauss, de Poisson, uniforme ou exponentielle par le test de Kolmogorov-Smirnov est disponible par certains logiciels statistiques (SPSS ou autres) ;
- 3) Dans le cas d'un test de la normalité d'une distribution, lorsque la moyenne et l'écart-type de cette distribution normale présumée ne sont pas connus mais estimés à partir de l'échantillon, le test statistique de Lilliefors peut être utilisé pour estimer la *p-value*. Plusieurs logiciels statistiques utilisent cette correction de signification en la signalant.

Exemple 8.4

On dispose des poids de 15 caisses de farine prélevées de manière aléatoire et simple dans le stock d'une minoterie. En utilisant les valeurs obtenues ci-dessous, vérifiez, en utilisant le test de Kolmogorov-Smirnov au seuil de 5%, pour décider si la distribution de l'échantillon est en adéquation avec la loi normale de moyenne 50 kg et d'écart-type 2.5 kg.

45,72	45,91	46,58	47,76	48,90	49,32	49,79	50,15
50,39	51,66	52,09	52,78	52,80	53,26	53,31	

Solution 8.4

a) Méthode : test de normalité par la méthode de Kolmogorov-Smirnov

b) Conditions d'application

- La population est pratiquement infinie ;
- L'échantillon des 15 caisses est tiré de manière aléatoire et simple ;
- La distribution est continue et entièrement définie : ces deux conditions sont vérifiées (Poids est une variable continue et les deux paramètres sont fixés *a priori*).

c) Hypothèses : la distribution de l'échantillon est en adéquation avec la loi normale de moyenne 50 kg et d'écart-type 2.5 kg.

d) Calcul

- Le tableau 8.6 donne le détail des calculs nécessaires au test de Kolmogorov-Smirnov ;
- L'écart maximum des deux dernières colonnes est : **0.147** (test bilatéral). Cette valeur est inférieure à la valeur de la table au niveau de signification $\alpha=5\%$ qui est de $D_{0,95}=0.338$, elle est même inférieure à la valeur de la table au niveau de signification $\alpha=20\%$ car $D_{0,80}=0.266$.

Conclusion : L'hypothèse de normalité de la population normale ne peut être rejetée au niveau de signification $\alpha=5\%$.

Remarque : Si on n'avait pas défini la moyenne et l'écart-type, ces deux paramètres seraient estimés par 50.028 et 2.645 respectivement et l'écart maximum serait de 0.131 avec une p-value de 0.200. Dans ce cas, la p-value a subi une correction de signification de Lilliefors (utilisez un logiciel statistique pour vérifier ces résultats).

Tableau 8.6. Principaux paramètres permettant le calcul de l'écart maximum pour la réalisation du test de Kolmogorov-Smirnov.

Poids	i	N'(i)	N'(i-1)	u(x _i)	F(x _i)	$\left \frac{i}{n} - F(x_i) \right $	$\left F(x_i) - \frac{i-1}{n} \right $
45,72	1	0,067	0,000	-1,7120	0,0434	0,0232	0,0434
45,91	2	0,133	0,067	-1,6360	0,0509	0,0824	0,0157
46,58	3	0,200	0,133	-1,3680	0,0857	0,1143	0,0477
47,76	4	0,267	0,200	-0,8960	0,1851	0,0815	0,0149
48,90	5	0,333	0,267	-0,4400	0,3300	0,0034	0,0633
49,32	6	0,400	0,333	-0,2720	0,3928	0,0072	0,0595
49,79	7	0,467	0,400	-0,0840	0,4665	0,0001	0,0665
50,15	8	0,533	0,467	0,0600	0,5239	0,0094	0,0573
50,39	9	0,600	0,533	0,1560	0,5620	0,0380	0,0287
51,66	10	0,667	0,600	0,6640	0,7467	0,0800	0,1467
52,09	11	0,733	0,667	0,8360	0,7984	0,0651	0,1318
52,78	12	0,800	0,733	1,1120	0,8669	0,0669	0,1336
52,80	13	0,867	0,800	1,1200	0,8686	0,0020	0,0686
53,26	14	0,933	0,867	1,3040	0,9039	0,0295	0,0372
53,31	15	1,000	0,933	1,3240	0,9072	0,0928	0,0261

8.4. QUELQUES AUTRES METHODES

8.4.1. Test de l'indice de dispersion pour la distribution de Poisson

Outre le test d'ajustement khi-deux du paragraphe 8.3, on trouve dans la littérature plusieurs autres tests efficaces pour tester l'ajustement à une distribution de Poisson. Parmi ceux-ci, on peut citer le test de l'indice de dispersion¹.

En effet, dans le domaine écologique, on peut avoir trois types de distributions spatiales des individus dans un espace géographique (figure 8.1) :

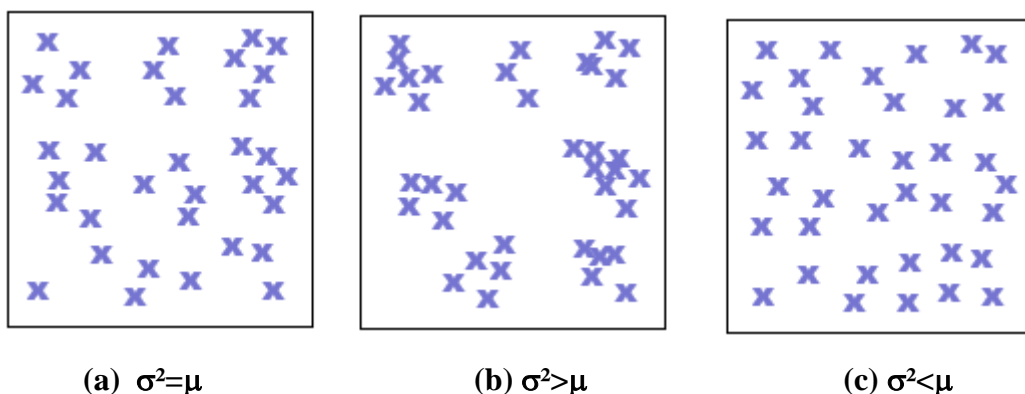


Figure 8.1. Types de répartitions spatiales des individus ((a): répartition aléatoire, (b) : sur-dispersion, (c): sous-dispersion).

¹ En anglais : index of dispersion test.

- Sur la figure 8.1 (a), la répartition spatiale des individus est purement **aléatoire** (distribution aléatoire ou de Poisson) ;
- Sur la figure 8.1 (b), les individus ont tendance à se regrouper. On parle de sur-dispersion ou de distribution agrégée. Il y a des valeurs extrêmes ;
- Sur la figure 8.1 (c), les individus ont tendance à se repousser mutuellement. On parle de sous-dispersion ou de contagion négative (distribution régulière).

Pour la distribution de Poisson, l'indice théorique $I = \sigma^2 / \mu$ auquel correspond l'indice estimé :

$$I_e = \frac{\hat{\sigma}^2}{\bar{x}}$$

est égal à l'unité. On peut donc utiliser cette valeur estimée pour avoir une idée sur la dispersion de la distribution par rapport à la distribution de Poisson.

Si la population est sous-dispersée, la variance sera beaucoup plus petite que la moyenne et donc l'indice de dispersion estimé sera plus proche de 0. Si la population est sur-dispersée, la variance sera beaucoup plus grande que la moyenne et donc l'indice de dispersion estimé sera beaucoup plus grand que 1.

Le test consiste à calculer la quantité :

$$\chi_{obs}^2 = (n-1)I_e$$

qui, sous l'hypothèse nulle et lorsque la moyenne est assez grande (supérieure à 2 selon certains auteurs et beaucoup plus grande selon d'autres), suit une distribution khi-deux à n-1 degrés de liberté. Ce test est bilatéral, on rejette l'hypothèse d'une compatibilité avec une distribution de Poisson lorsque :

$$\chi_{obs}^2 \leq \chi_{\alpha/2}^2$$

c'est-à-dire dans le cas où la population est sous-dispersée, et lorsque :

$$\chi_{obs}^2 \geq \chi_{1-\alpha/2}^2$$

c'est-à-dire dans le cas où la population est sur-dispersée (agrégée).

Exemple 8.5

Reprenez les données du tableau 8.2 et réalisez un test ajustement à une distribution de Poisson, en utilisant le test basé sur l'indice de dispersion ($\alpha=0.05$) ?

Solution 8.5

a) Méthode : ajustement à une distribution de Poisson par le test de l'indice de dispersion

b) Conditions d'application :

- La population est pratiquement infinie
- L'échantillon des 70 taureaux est tiré de manière aléatoire et simple
- La moyenne est supérieure à 2

c) Hypothèses : cf. exemple 8.2.

d) Calcul

La valeur observée de la variable khi-deux est : $\chi_{obs}^2 = (70-1) \frac{2.501}{2.614} = 66.02$.

On a :

$$\chi_{0.025}^2 < 66.02 < \chi_{1-\alpha/2}^2 \text{ avec } \chi_{0.025}^2 = 47.92 \text{ et } \chi_{1-\alpha/2}^2 = 93.86,$$

ce qui ne permet pas de rejeter l'hypothèse nulle. L'indice de dispersion observé :

$$I_e = 2.501/2.614 = 0.96$$

est d'ailleurs très proche de 1.

Conclusion : En prenant un risque de 5%, on dit qu'il n'y a pas suffisamment de preuve pour rejeter l'hypothèse nulle, la distribution de Poisson constitue un ajustement pour les données observées.

8.4.2. Test de Shapiro-Wilk pour la distribution normale

Le test de Shapiro-Wilk permet de tester la normalité d'une population. On l'utilise souvent lorsque la taille de l'échantillon est faible. Les étapes du test consistent à :

- Ordonner les n observations par ordre croissant :

$$y_1 \leq y_2 \leq \dots \leq y_{n-1} \leq y_n$$

- Calculer la quantité T_1 par la relation :

$$T_1 = SCE_y$$

- Calculer les différences :

$$d_1 = y_n - y_1$$

$$d_2 = y_{n-1} - y_2$$

...

$$d_i = y_{n-i+1} - y_i$$

...

- Calculer la quantité T_2 par la relation :

$$T_2 = \left[\sum_{i=1}^{\lfloor n/2 \rfloor} a_n(i) d_i \right]^2$$

où les coefficients $a_n(i)$ sont les scores normaux (tableau 8.7) et $\lfloor n/2 \rfloor$ correspond à la partie entière de $n/2$. Si n est pair, il y a $n/2$ différences et si n est impair, il y a également $\lfloor n/2 \rfloor$ différences, c'est-à-dire que l'observation médiane n'intervient pas ;

- Obtenir la valeur observée en calculant le rapport :

$$W_{obs} = \frac{T_2}{T_1}$$

- Rejeter, au risque α , l'hypothèse de normalité lorsque :

$$W_{obs} < W_\alpha$$

où les valeurs de W_α sont données par le tableau 8.8.

Exemple 8.6

L'échantillon des douze valeurs suivantes peut-il être considéré comme extrait d'une population normale ($\alpha=0.01$) ?

9.1 3.2 45.2 23.9 21.9 20.2 11.9 86.1 32.7 18.7 11.6 28.4

Solution 8.6

a) Méthode : ajustement à une normale par la méthode de Shapiro-Wilk

b) Conditions d'application :

- L'échantillon des douze observations est prélevé de manière aléatoire et simple

c) Hypothèses : cf. exemple 8.3.

d) Calcul

On ordonne les valeurs par ordre croissant et on obtient les différences d_i :

$$d_1 = y_{12} - y_1 = 86.1 - 3.2 = 82.9 ;$$

$$d_2 = y_{11} - y_2 = 45.2 - 9.1 = 36.1 ;$$

$$d_3 = 32.7 - 11.6 = 21.1 ;$$

$$d_4 = 28.4 - 11.9 = 16.5 ;$$

$$d_5 = 23.9 - 18.7 = 5.2 ; \text{ et}$$

$$d_6 = 21.9 - 20.2 = 1.7 .$$

$$T_1 = SCE_y = 5351.003$$

$$T_2 = [(82.9)(0.5475) + (36.1)(0.3325) + (21.1)(0.2347) + (16.5)(0.1586) + (5.2)(0.0922) + (1.7)(0.0303)]^2 = [65.491]^2$$

ce qui donne une valeur observée de Shapiro-Wilk :

$$W_{obs} = \frac{T_2}{T_1} = \frac{(65.491)^2}{5351.003} = 0.80$$

Comme la valeur critique vaut 0.859, on est amené à rejeter l'hypothèse nulle puisque $W_{obs} < W_{0.05}$.

La p-value obtenue par les logiciels statistiques est de 0.0075.

Conclusion : En prenant un risque de 1%, on ne peut pas accepter l'hypothèse de la normalité de la population-parent.

Exercice 8.7

Refaites le même exercice au moyen du test de Kolmogorov-Smirnov, en utilisant un logiciel statistique pour la correction de p-value.

Tableau 8.7. Table des coefficients a_i du test de Shapiro-Wilk

[illegible][illegible][illegible]

Tableau 8.8. Table des valeurs critiques du test de Shapiro-Wilk

	Risque 5%	Risque 1%
n	$W_{0,05}$	$W_{0,01}$
5	0,7 62	0.686
6	0.988	0.713
7	0.803	0.730
8	0.818	0.749
9	0.829	0.764
10	0.842	0.781
11	0.850	0.792
12	0.859	0.805
13	0.866	0.814
14	0.874	0.825
15	0.881	0.835
16	0.887	0.844
17	0.892	0.851
18	0.897	0.858
19	0.901	0.863
20	0.905	0.868
21	0.908	0.873
22	0.911	0.878
23	0.914	0.881
24	0.916	0.884
25	0.918	0.888
26	0.920	0.891
27	0.923	0.894
28	0.924	0.896
29	0.926	0.898
30	0.927	0.900
31	0.929	0.902
32	0.930	0.904
33	0.931	0.906
34	0.933	0.908
35	0.934	0.910
36	0.935	0.912
37	0.936	0.914
38	0.938	0.916
39	0.939	0.917
40	0.940	0.919
41	0.941	0.920
42	0.942	0.922
43	0.943	0.923
44	0.944	0.924
45	0.945	0.926
46	0.945	0.927
47	0.946	0.928
48	0.947	0.929
49	0.947	0.929
50	0.947	0.930

CHAPITRE 9

TESTS D'INDÉPENDANCE

9.1. INTRODUCTION

Les tests d'indépendance permettent de tester la vraisemblance d'une absence de liaison entre deux variables qualitatives dans la population à partir d'un échantillon aléatoire et simple prélevé dans celle-ci.

Les données sont des fréquences réparties en classes croisant les modalités des deux variables qualitatives et consignées dans un tableau de contingence. On peut facilement voir, si la dimension du tableau n'est pas grande, comment une variable dépend de l'autre, mais la signification de cette dépendance reste à vérifier par un test d'indépendance.

Trois tests d'indépendance seront présentés dans le paragraphe 9.2 de ce chapitre, il s'agit du test khi-deux de Pearson, du test khi-deux du rapport de vraisemblance et du test exact de Fisher. Nous utiliserons aussi la statistique de Mantel-Haenszel pour tester la présence ou non d'une relation linéaire, mais cette statistique ne convient que pour les variables qualitatives ordinales.

Les tests d'indépendance que nous utilisons habituellement nous ne donnent pas d'information sur l'intensité de la relation entre les deux variables. Celle-ci peut être renseignée par le calcul de coefficients de mesures d'association (paragraphe 9.3).

Si les deux variables sont mesurées selon une échelle ordinale, on peut aussi utiliser des coefficients qui renseignent à la fois sur l'intensité et le sens de la relation linéaire entre les deux variables. Parmi ces coefficients, nous présenterons le coefficient gamma de Goodman et Kruskal (paragraphe 9.4).

9.2. PRESENTATION DES TESTS D'INDEPENDANCE

9.2.1. Introduction

Soient X et Y deux variables qualitatives à p et q modalités respectivement, et soit une population pratiquement infinie dont les individus sont classés en pq classes en fonction de ces deux variables. Supposons que l'on dispose d'un échantillon aléatoire et simple tiré de cette population et que les résultats du dénombrement sont consignés dans une distribution de fréquences à deux dimensions, en adoptant les notations suivantes :

- n_{ij} : fréquence observée pour la classe croisant la i-ième modalité de la variable i avec la j-ième modalité de la variable j ;
- $n_{i.}$: fréquence de la modalité i de la première variable ;
- $n_{.j}$: fréquence de la modalité j de la deuxième variable ;
- n : effectif total de l'échantillon.

Les tests d'indépendance ont pour but de vérifier si les deux variables qualitatives (appelées aussi critères ou caractères) sont indépendantes ou bien s'influencent entre elles. Quatre de ces tests seront exposés après la présentation des données d'un exemple qui nous servira à illustrer les calculs.

9.2.2. Exemple introductif

Des informations ont été collectées sur un échantillon aléatoire et simple de 225 exploitants agricoles selon qu'ils sont bénéficiaires ou non-bénéficiaires d'un projet de développement donné et selon le degré de demande de conseils techniques à d'autres exploitants agricoles. Les données étaient d'abord collectées et enregistrées dans un tableau semblable au tableau 9.1 avant d'être regroupées dans un tableau de contingence croisant les deux caractères (tableau 9.2).

Peut-on dire que les bénéficiaires du projet s'insèrent davantage que les non-bénéficiaires dans une démarche collective d'échanges de conseils techniques avec d'autres agriculteurs ? Autrement dit, y a-t-il un lien entre la catégorie d'exploitants et le degré de demande de conseils techniques ?

Tableau 9.1. Catégorie et degré de demande de conseils techniques de 225 exploitants agricoles

Agriculteur	Catégorie	Conseil
1	Bénéficiaire	Souvent
2	Non bénéficiaire	Parfois
3	Bénéficiaire	Souvent
.	.	.
20	Non bénéficiaire	Parfois
21	Bénéficiaire	Souvent
22	Bénéficiaire	Parfois
.	.	.
72	Non bénéficiaire	Jamais
73	Bénéficiaire	Jamais
74	Bénéficiaire	Jamais
.	.	.
225	Non bénéficiaire	Parfois

Tableau 9.2. Répartitions de 225 exploitants agricoles en fonction de leur catégorie et du degré de demande de conseils techniques.

Catégorie d'exploitants	Degré de demande de conseils techniques			Total
	Souvent	Parfois	Jamais	
Bénéficiaires	19	52	27	98
Non-bénéficiaires	18	51	58	127
Total	37	103	85	225

Dans le cas de cet exemple, la catégorie d'exploitants et le degré de demande de conseils techniques constituent les deux variables qualitatives (ou les deux critères) dont on souhaite tester l'indépendance. La première variable possède deux modalités et la seconde variable en possède trois.

9.2.3. Hypothèse nulle

L'hypothèse nulle qu'on souhaite tester est :

H_0 : les deux critères sont indépendants
contre H_1 : les deux critères ne sont pas indépendants.

Ces deux hypothèses peuvent aussi s'écrire :

$$H_0 : P_{ij} = P_{i.}P_{.j} \quad \text{contre} \quad H_1 : P_{ij} \neq P_{i.}P_{.j} \quad (i=1,2, \dots, p \text{ et } j=1,2, \dots, q),$$

où $P_{i.}$ et $P_{.j}$ sont les probabilités marginales et P_{ij} est la probabilité à deux dimensions. Nous verrons que ces quantités peuvent être estimées respectivement par $n'_{i.}$, $n'_{.j}$ et n'_{ij} .

D'une manière générale, si les fréquences marginales sont égales ou presque égales, la présence ou l'absence d'association entre les deux variables peuvent être examinées en analysant les distributions des différentes modalités en regardant en lignes (ou en colonnes) le tableau de fréquences. On essaie ainsi de comprendre la structure du tableau en cherchant à localiser les endroits où existent d'éventuels points d'attraction et/ou de répulsion entre les modalités.

Mais lorsque les fréquences marginales sont inégales, la comparaison entre deux modalités de la même variable sur la base des fréquences absolues n'a pas de sens. Par exemple, si l'on souhaite savoir qui a tendance à demander souvent des conseils auprès des autres exploitants (les bénéficiaires ou les non-bénéficiaires), on ne compare pas les fréquences 19 et 18, mais bien 19 sur 98 et 18 sur 127.

En effet, si l'on souhaite comparer deux modalités en lignes, il est plus commode de faire cette comparaison en se basant sur des fréquences conditionnelles, c'est-à-dire sur la répartition en pourcentages à l'intérieur de chaque ligne. On obtient ce qu'on appelle des profils associés aux lignes (tableau 9.3). Ce tableau permet de remarquer que les profils associés à la catégorie des exploitants sont différents. On a par exemple 45.7% des non-bénéficiaires qui déclarent ne jamais demander des conseils techniques à d'autres exploitants agricoles contre seulement 27.6% chez les bénéficiaires.

On peut également faire une analyse semblable en calculant les profils associés aux colonnes.

L'analyse des tableaux des profils laisse penser à une non-indépendance entre la catégorie d'exploitants et le degré de demande de conseils techniques. Cela reste à confirmer par un test d'indépendance.

Tableau 9.3. Profils associés aux lignes obtenus à partir du tableau 9.2.

Catégorie d'exploitants	Degré de demande de conseils techniques			Total
	Souvent	Parfois	Jamais	
Bénéficiaires	19.4	53.1	27.6	100
Non-bénéficiaires	14.2	40.2	45.7	100

9.2.4. Test khi-deux de Pearson

Le principe du test khi-deux de Pearson consiste à comparer les effectifs (fréquences) réellement observés n_{ij} des différentes classes aux effectifs théoriques ou attendus $n\hat{P}_{ij}$. Les effectifs attendus sont ceux qui seraient obtenus dans le cas de l'indépendance des deux critères, c'est-à-dire lorsque l'hypothèse nulle est vraie.

Le test se réalise en calculant la quantité :

$$\chi_{obs}^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - n\hat{P}_{ij})^2}{n\hat{P}_{ij}} \quad (1)$$

$$= \sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n\hat{P}_{ij}} - n \quad (2)$$

et en rejetant l'hypothèse nulle lorsque :

$$\chi_{obs}^2 \geq \chi_{1-\alpha}^2$$

avec $k=(p-1)(q-1)$ degrés de liberté. Ce test n'est qu'approximatif, il n'est valable que si l'effectif total de l'échantillon est suffisamment grand (> à 40 ou 50) et que les effectifs attendus des différentes cellules du tableau sont supérieurs ou égaux à 5. Si ces conditions ne sont pas réunies, on utilisera le test exact de Fisher.

Dans l'expression (1), la quantité \hat{P}_{ij} est une estimation de la probabilité d'obtenir une observation possédant la modalité i du premier caractère et la modalité j du second caractère lorsqu'il y a indépendance, c'est-à-dire lorsque l'hypothèse nulle est vraie. On peut en déduire l'effectif attendu :

$$n\hat{P}_{ij} = nn_{i.}n_{.j} = \frac{n_{i.}n_{.j}}{n}. \quad (3)$$

La valeur χ^2_{obs} prend des valeurs positives, elle est nulle sous l'hypothèse nulle, et sa valeur, en cas d'association "parfaite" entre les deux variables dépend de l'effectif de l'échantillon et du nombre de modalités.

Comme nous l'avons signalé, le test n'est qu'approché. L'approximation n'est satisfaisante que lorsque l'effectif global est grand et que les fréquences attendues des différentes classes (cellules du tableau) sont toutes supérieures ou égales à 5. Si cette règle, qui fait à peu près l'unanimité des statisticiens, n'est pas vérifiée, il y a lieu de procéder à des regroupements de modalités d'un caractère, si cela un sens, pour obtenir des fréquences attendues au moins égales à 5 ou passer par le test exact.

D'après certains auteurs, on peut tolérer la présence d'une fréquence attendue comprise entre 1 et 5 lorsqu'il y a 6 à 10 classes et 2 fréquences comprises entre 1 et 5 lorsqu'il y a plus de 10 classes. D'après d'autres auteurs, l'utilisation des tests du Khi-deux peut être considérée inappropriée si une fréquence attendue est inférieure à 1 ou si la fréquence attendue est inférieure à 5 dans plus de 20% des cas.

Yates a suggéré une correction de continuité consistant à soustraire 0,5 des écarts positifs et à ajouter 0,5 aux écarts négatifs dans le calcul de la différence entre les fréquences, et ce, avant d'élever au carré dans la formule de la statistique du khi-deux :

$$\chi^2_{obs} = \sum_{i=1}^p \sum_{j=1}^q \frac{(|n_{ij} - n\hat{P}_{ij}| - 0.5)^2}{n\hat{P}_{ij}}.$$

On constate que cette correction réduit la valeur du Khi-deux et augmente la valeur du p-value et, dans certains cas, la correction s'ajuste trop fort ce qui rend son utilisation limitée. Certains utilisateurs recourent à cette correction lorsqu'il y a présence de fréquences attendues inférieures à 5, tout en étant supérieures à 3. Mais, généralement, on ne devrait probablement l'utiliser que pour des tests faisant intervenir deux variables à deux modalités chacune (2x2) où l'on considère acceptables les fréquences attendues inférieures à 5 si la correction de Yates est appliquée.

D'autre part, il faut faire attention à l'interprétation des résultats d'un test d'indépendance :

- Le rejet de l'hypothèse nulle, c'est-à-dire l'acceptation de l'existence d'une dépendance, n'implique pas nécessairement l'existence d'une relation directe de cause à effet entre les deux critères considérés ;
- Une valeur élevée du khi-deux permet de rejeter l'hypothèse d'indépendance avec une grande sécurité sans toutefois prouver que l'intensité de l'association entre les deux critères est très forte. Lorsqu'il existe une liaison entre les deux critères, la valeur du khi-deux augmente notamment avec l'augmentation de l'effectif de l'échantillon.

Exemple 9.1

Reprenez les données du tableau 9.2 et vérifiez si l'on peut dire que la catégorie d'exploitants et le degré de demande de conseils techniques sont liés ($\alpha=0.05$) ?

Solution 9.1

a) **Méthode** : test khi-carré d'indépendance

b) **Conditions d'application**

- la population est pratiquement infinie
- l'échantillon des 225 exploitants est prélevé de manière aléatoire et simple
- les effectifs attendus sont égaux ou supérieurs à 5 (hypothèse vérifiée)

c) **Hypothèses**

H_0 : la catégorie d'exploitants et le degré de demande de conseils sont indépendants
contre H_1 : les deux critères ne sont pas indépendants

d) **Calcul**

Le tableau 9.4 reprend les effectifs observés n_{ij} et attendus $n\hat{P}_{ij}$ (en gras) des différentes classes.

Tableau 9.4. Fréquences observées et attendues (en gras) obtenues pour les différentes classes.

Catégorie d'exploitants	Degré de demande de conseils			Total
	Souvent	Parfois	Jamais	
Bénéficiaires	19 16.12	52 44.86	27 37.02	98
Non-bénéficiaires	18 20.88	51 58.14	58 47.98	127
Total	37	103	85	225

- La valeur observée du khi-deux :

$$\chi_{obs}^2 = \frac{19^2}{16.12} + \frac{52^2}{44.86} + \dots + \frac{58^2}{47.98} - 225 = 7.73$$

est supérieure à la valeur critique $\chi_{0.95}^2=5.99$ à $k=2$ degrés de liberté (p -valeur=0.021). L'hypothèse nulle doit donc être rejetée au niveau $\alpha=0.05$.

e) **Conclusion**

En prenant un risque de 5%, la catégorie d'exploitants et le degré de demande de conseils technique ne peuvent être considérés comme indépendants.

9.2.5. Test khi-deux du rapport de vraisemblance¹

Ce test permet également de vérifier s'il y a indépendance entre deux variables qualitatives, en faisant intervenir les rapports entre les fréquences observées et les fréquences attendues. La statistique du test se calcule par la relation :

$$\chi_{ML}^2 = 2 \sum_{i=1}^p \sum_{j=1}^q n_{ij} \ln \frac{n_{ij}}{n\hat{P}_{ij}}. \quad (4)$$

L'hypothèse d'indépendance est rejetée lorsque cette valeur observée dépasse la valeur théorique $\chi_{1-\alpha}^2$ à $(p-1)(q-1)$ degrés de liberté. L'approximation par la distribution khi-deux est valide lorsque l'effectif de l'échantillon est assez grand.

Le test du rapport de vraisemblance est moins fiable que le test khi-deux lorsqu'on a des effectifs qui ne sont pas assez grands. Les valeurs de χ_{ML}^2 et de χ_{obs}^2 sont proches si l'effectif est assez grand ou si l'on est "proche" de l'indépendance.

¹ En anglais: *likelihood ratio chi-square*.

Exemple 9.2

Reprenez les données du tableau 9.2 et vérifiez si l'on peut dire que la catégorie d'exploitants et le degré de demande de conseils techniques sont liés en utilisant le test khi-deux du rapport de vraisemblance ($\alpha=0.05$) ?

Solution 9.1

a) **Méthode** : test khi-carré du rapport de vraisemblance

b) **Conditions d'application** : voir exemple 9.1

c) **Hypothèses** : voir exemple 9.1

d) **Calcul**

La valeur du khi-deux du rapport de vraisemblance :

$$\chi_{ML}^2 = 2((19)(0.1647) + \dots + (58)(0.1897)) = 7.86$$

est supérieure à la valeur critique $\chi_{0.95}^2 = 5.99$ à 2 degrés de liberté (p-valeur=0.020). L'hypothèse nulle doit donc être rejetée au niveau $\alpha=0.05$.

e) **Conclusion** : voir exercice 9.1.

9.2.6. Test exact de Fisher

Les tests d'indépendance présentés aux paragraphes 9.2.4 et 9.2.5 se basent sur le théorème central limite. Ils sont valides lorsque les effectifs sont assez grands. Pour des effectifs petits ou lorsqu'une ou plusieurs cases du tableau ont une fréquence attendue inférieure à 5, on peut utiliser le test exact de Fisher pour tester l'indépendance. Il s'agit d'un test à marges fixées.

Lorsque le tableau de contingence est de dimensions 2 x 2, c'est-à-dire que chaque variable a deux modalités, le test exact de Fisher reste relativement simple à réaliser. Pour des tableaux de dimensions supérieures à 2 x 2, on peut appliquer le test exact de Fisher généralisé (test de Fisher-Freeman-Halton). Ce test exige un calcul long et compliqué. Il se peut aussi que ce test ne puisse être calculé pour certains tableaux. Dans ce dernier cas, on peut essayer une méthode hybride basée sur les règles de Cochran (Mehta et Patel, 1986).

9.2.7. Statistique de Mantel-Haenszel pour relation linéaire

Les valeurs des statistiques du test chi-deux de Pearson et du test du rapport de vraisemblance ne changent pas si on permute les lignes ou les colonnes du tableau de contingence. Ces tests supposent l'absence d'ordre dans les valeurs des deux variables.

Lorsque les deux variables sont mesurées selon des échelles ordinales, on peut utiliser le test de Mantel-Haenszel pour l'association linéaire. Ce test permet de tester si l'on peut affirmer que l'association entre les variables est linéaire. Il se calcule selon la relation suivante :

$$\chi_{MH}^2 = (n-1)r^2 \quad (5)$$

dans laquelle r est le coefficient de corrélation de Pearson entre les deux variables en numérotant les modalités de la première variables par 1, 2, ..., p et celles de la seconde variable par 1, 2, ..., q .

On conclut à l'existence d'une relation linéaire significative entre les deux variables si :

$$\chi_{MH}^2 > \chi_{1-\alpha}^2$$

où la variable $\chi_{1-\alpha}^2$ possède un degré de liberté. Un exemple de ce test sera donné au paragraphe 9.4.

9.2.8. Remarque

Dans le cas du test d'indépendance, nous avons vu que l'on choisit un échantillon aléatoire et simple de n individus et que l'on répartit ensuite ces individus en fonctions des deux critères de classification. Or il se peut, pour certains problèmes, que les effectifs marginaux de l'une des deux variables soient fixés au préalable. Ce genre de répartition se rencontre souvent lorsque les individus sont soumis à certains traitements.

Pour ce type d'échantillonnage, on est souvent amené à tester si la distribution des valeurs de l'une des variables (Y par exemple) est homogène dans chacune des populations de l'autre variable (X par exemple).

L'hypothèse en question est en effet une hypothèse d'homogénéité de distributions. Le test se réalise mathématiquement de la même manière que le test d'indépendance.

Exemple 9.3

Cinq usines fabriquent des boîtes de ton. Les boîtes peuvent être examinées pour vérifier s'ils répondent à certaines exigences de qualités. Des échantillons aléatoires de 80, 100, 70, 90 et 120 boîtes provenant respectivement des usines A, B, C, D et E ont permis de repérer des boîtes de qualités insuffisantes, respectivement au nombre de 7, 8, 6, 8 et 11. Pouvez-vous conclure que les cinq fabrications sont de qualités différentes ($\alpha=0.05$) ?

Solution 9.3

a) **Méthode** : test d'homogénéité de populations

b) **Conditions d'application**

- les effectifs marginaux des cinq échantillons sont fixés et prélevés dans populations pratiquement infinies
- chaque échantillon est prélevé de manière aléatoire et simple
- les effectifs attendus sont supérieurs à 5 (condition vérifiée)

c) **Hypothèses**

H_0 : les cinq fabrications sont de la même qualité
contre H_1 : les cinq fabrications ne sont pas de la même qualité

d) **Calcul**

Le tableau 9.5. reprend les effectifs observés n_{ij} et attendus $n\hat{p}_{ij}$ des différentes classes.

Tableau 9.5. Fréquences observées et attendues (en gras) obtenues pour les différentes classes.

Usine	Qualité		Total
	Bonne	Mauvaise	
A	73 73.04	7 6.96	80
B	92 91.30	8 8.70	100
C	64 63.91	6 6.09	70
D	82 82.17	8 7.83	90
E	109 109.57	11 10.43	120
Total	420	40	460

1) On a $\chi^2_{obs}=0.10$ et $\chi^2_{0.95}=9.49$ avec $k=4$ degrés de liberté (p-valeur=0.999). L'hypothèse nulle ne peut être rejetée au niveau $\alpha=0.05$ (on constate que les effectifs observés et théoriques sont très proches).

2) En prenant un risque de 5%, on peut dire qu'il n'y a pas de différence de qualité entre les boîtes de ton produites par les cinq fabrications.

9.3. MESURES D'ASSOCIATION

9.3.1. Introduction

Nous avons vu que le test d'indépendance permet de tester l'existence d'une relation entre deux variables qualitatives ayant respectivement p et q modalités. Lorsque l'hypothèse d'indépendance est rejetée, on conclut à l'existence d'association entre les deux variables. La valeur de la statistique khi-deux ne peut être interprétée comme une mesure de l'intensité de la relation entre les deux variables, le test khi-deux est en effet influencé par d'autres facteurs, notamment par la taille de l'échantillon.

Trois des coefficients qui permettent d'évaluer le degré de l'intensité de la relation entre deux variables seront passés en revue : le coefficient phi, le coefficient de contingence et le coefficient phi de Cramer.

9.3.2. Le coefficient Phi

Le coefficient ϕ permet de remédier à l'influence de l'effectif de l'échantillon dans le calcul de la statistique chi-deux. Il se calcule par la relation suivante :

$$\phi = \sqrt{\frac{\chi_{obs}^2}{n}} \quad (6)$$

dans laquelle χ_{obs}^2 est la valeur observée du khi-carré de Pearson du test d'indépendance (cf. paragraphe 9.2.4) et n est l'effectif total de l'échantillon. La valeur minimale de phi est zéro, elle indique l'absence de lien entre les deux variables, tandis que la valeur maximale dépend de la dimension du tableau de contingence.

Le coefficient phi (ϕ) est surtout utilisé dans le cas où les deux variables sont dichotomiques, c'est-à-dire lorsque chacune des deux variables ne prend que deux modalités (tableaux de dimensions 2 x 2). Dans ce cas, il représente le coefficient de corrélation lorsqu'on attribue la valeur 0 à l'une des modalités de chaque variable et la valeur 1 à l'autre modalité. La valeur du coefficient ϕ varie dans ce cas sur une échelle allant de 0 à 1. Une valeur de 1 signifie un lien presque parfait et une valeur nulle indique l'absence de lien entre les deux variables. En général, plus la valeur de ϕ est proche de 1, plus le lien est fort, et plus sa valeur est proche de 0, plus le lien est faible.

Pour ceux qui utilisent le logiciel statistique SAS, ils peuvent constater que le phi est donné avec un signe. Celui-ci est négatif si l'association se trouve suivant l'anti-diagonale.

9.3.3. Le coefficient de contingence²

Le coefficient de contingence C permet également de mesurer l'intensité de la liaison qui existe entre deux variables qualitatives. Il se calcule par la formule :

$$C = \sqrt{\frac{\chi_{obs}^2}{\chi_{obs}^2 + n}} \quad (7)$$

dans laquelle χ_{obs}^2 est la valeur observée du khi-carré de Pearson du test d'indépendance (cf. paragraphe 9.2.4) et n est l'effectif total de l'échantillon. La valeur de ce coefficient est nulle

² En anglais: *contingency coefficient*.

lorsqu'il y a indépendance et, dans le cas d'un lien parfait, la valeur de C dépend de la dimension du tableau de contingence. Elle n'approche la valeur 1 que dans le cas de tableaux de grandes dimensions (p et q grands). Certains auteurs ne recommandent ce coefficient que pour les tableaux de dimensions d'au moins 5x5. Pour des tableaux de petites dimensions, le coefficient C tend à surestimer le niveau d'association.

Certains chercheurs utilisent ce qu'on appelle le coefficient de contingence corrigé qui a l'avantage de varier entre 0 et 1. Celui-ci est donné par la relation :

$$C_{\text{corrigé}} = C / C_{\text{max}}$$

dans laquelle C_{max} peut être lue dans la table donnée par Champion [1970].

9.3.4. Le coefficient phi de Cramer

Le coefficient phi de Cramer s'interprète facilement. Lorsqu'il est égal à 0, il y a indépendance entre les deux variables et lorsqu'il est égal à 1, il y a un lien parfait. Ce coefficient se calcule par la relation :

$$\phi_c = \sqrt{\frac{\chi_{obs}^2}{n(k-1)}} \quad (8)$$

dans laquelle k est la plus petite valeur entre le nombre de lignes et de colonnes ($k=\min(p,q)$). Les autres paramètres gardent la même signification que dans le cas de la statistique phi. Lorsque k est égal à deux, le phi de Cramer est le même que le phi classique.

Exemple 9.4

Reprenez les données du tableau 9.2 et mesurez l'intensité de la relation entre la catégorie d'exploitants et le degré de demande de conseils techniques ?

Solution 9.4

a) **Méthode** : calcul de coefficients de mesures d'association

b) **Calcul** :

on obtient :

$$\text{Coefficient Phi : } \phi = \sqrt{\frac{7.733}{225}} = 0.185 \text{ avec une p-value approximative de 0.021}$$

$$\text{Coefficient de contingence : } C = \sqrt{\frac{7.733}{7.733+225}} = 0.182 \text{ avec une p-value approximative de 0.021}$$

$$\text{Coefficient phi de Cramer } \phi_c = \sqrt{\frac{7.733}{225(2-1)}} = 0.185 \text{ avec une p-value approximative de 0.021}$$

Ces valeurs permettent de qualifier la relation entre la catégorie d'exploitants et le degré de demande de conseils techniques de faible.

9.4. DIRECTION DE LA RELATION

9.4.1. Introduction

Différentes mesures statistiques permettent de décrire le sens de la relation entre deux variables qualitatives lorsqu'il existe un ordre dans les modalités de chacune d'elles. Parmi ces mesures, nous présentons le coefficient gamma.

9.4.2. Le coefficient Gamma

Le coefficient gamma, appelé aussi gamma de Goodman et Kruskal, permet d'obtenir de l'information sur l'intensité et le sens de la relation linéaire de deux variables ordinales.

Ce coefficient mesure le surplus du nombre de paires d'observations en concordances par rapport au nombre de paires d'observations en discordances, rapporté au nombre total de paires. Il ne tient pas compte des ex-aequo.

Le nombre d'observations C_{ij} en concordance avec celles d'une cellule ij s'obtient en sommant les fréquences de toutes les cellules du coin supérieur gauche et du coin inférieur droit du tableau de fréquences par rapport à cette cellule :

$$C_{ij} = \sum_{s>i} \sum_{t>j} n_{st} + \sum_{s<i} \sum_{t<j} n_{st} \quad (9)$$

et le nombre de paires concordantes est donné par :

$$C = \frac{1}{2} \sum_i \sum_j n_{ij} C_{ij} . \quad (10)$$

De même, le nombre d'observations D_{ij} en discordance avec celles d'une cellule ij s'obtient en sommant les fréquences de toutes les cellules du coin supérieur droit et du coin inférieur gauche du tableau de fréquences par rapport à cette cellule :

$$D_{ij} = \sum_{s>i} \sum_{t<j} n_{st} + \sum_{s<i} \sum_{t>j} n_{st} \quad (11)$$

et le nombre de paires discordantes est :

$$D = \frac{1}{2} \sum_i \sum_j n_{ij} D_{ij} . \quad (12)$$

La valeur du coefficient gamma s'obtient ensuite par la relation :

$$\gamma = \frac{C - D}{C + D} . \quad (13)$$

Cette valeur varie de -1 à +1. Le signe indique si l'association est positive ou négative et la grandeur indique l'intensité de l'association. Dans le cas d'une indépendance entre les deux variables, le coefficient gamma vaut zéro, mais si $\gamma=0$ on ne peut conclure à une indépendance car il peut y avoir une relation non linéaire.

Lorsque l'effectif des observations est élevé, gamma est distribué approximativement selon une loi normale. Aussi, certains logiciels statistiques fournissent l'erreur-standard asymptotique et parfois le degré de signification.

Exemple 9.5

Un sondage a été réalisé auprès d'un échantillon de 244 agriculteurs choisis de manière aléatoire et simple pour connaître leur degré de satisfaction quant aux services offerts par le crédit agricole. En outre, ces agriculteurs ont été répartis en trois classes selon leur dotation en facteurs fixes (SAU, matériel agricole, etc.) en petits, moyens et grands agriculteurs.

En fonction des résultats obtenus (tableau 9.6), peut-on conclure à un lien entre le degré de satisfaction des exploitants et la taille de leur exploitation ? Si oui, indiquez l'importance et le sens de ce lien ?

Tableau 9.6. Répartitions d'un échantillon de 244 exploitants en fonction du degré de satisfaction vis-à-vis du crédit agricole et de la taille de l'exploitation.

Satisfaction	Classe des exploitations			Total
	Petite	Moyenne	Grande	
Faible	33	24	19	76
Moyenne	31	46	26	103
Elevée	8	14	43	65
Total	72	84	88	244

Solution 9.5

a) **Méthode** : évaluation de l'intensité et du sens de la relation entre deux variables qualitatives

b) **Conditions d'application** (pour tester l'indépendance) :

- L'échantillon des 244 agriculteurs est aléatoire et simple et de grande taille (condition vérifiée)
- Les fréquences attendues sont supérieures ou égales à 5 (condition vérifiée)

c) **Hypothèse nulle** :

H_0 : Les deux variables sont indépendantes

H_1 : Les deux variables ne sont pas indépendantes

d) **Calcul** :

1) La valeur du khi-deux de Pearson : $\chi^2_{obs} = 40.205$ (p-value= 0.000). On conclut à un lien très hautement significatif entre le degré de satisfaction des exploitants et la taille de leur exploitation.

2) La statistique du khi-deux de Mantel-Haenszel est donnée par :

$$\chi^2_{MH} = (244-1)(0.3327)^2 = 26.90$$

et la p-valeur est de 0.000. Il y a donc présence d'une relation linéaire significative entre les deux variables.

3) Le coefficient gamma est donné par :

$$\gamma = \frac{9658 - 3757}{9658 + 3757} = 0.44 \text{ avec p-value approximative de 0.000.}$$

Cette valeur traduit que la relation linéaire est modérée et positive. Les exploitants des plus grandes exploitations expriment une satisfaction élevée alors que ceux des petites exploitations expriment une satisfaction faible.

9.5. CONCLUSION

Nous avons passé en revue différentes statistiques permettant de vérifier s'il y a indépendance ou non entre deux variables qualitatives. Le test du rapport de vraisemblance est moins fiable que le test khi-deux lorsqu'on a des nombres qui ne sont pas assez grands. Le test khi-deux est généralement utilisé lorsque le test exact de Fisher ne peut être calculé.

L'analyse des tableaux de profils permettent une certaine description de la relation entre les deux variables, mais, pour des tableaux avec des variables comportant plusieurs modalités, une analyse exploratoire plus intéressante pourra être obtenue en utilisant une méthode statistique multidimensionnelle appelée analyse factorielle des correspondances ou AFC (*cf.* cours d'analyse des données dispensé en 2^{ième} année du cycle ingénieur de IAV Hassan II).

D'autre part, nous avons souligné que l'intensité de la relation entre deux variables qualitatives peut être évaluée par l'un des coefficients de mesures d'association. Une variable est en effet plus ou moins fortement influencée par une autre. Trois de ces mesures ont été présentées.

Enfin, nous avons vu que, si les deux variables sont mesurées selon des échelles ordinales, on peut utiliser certaines mesures pour obtenir de l'information sur l'intensité et le sens de la relation linéaires entre les deux variables. Parmi ces mesures, nous avons présenté le coefficient gamma.

Il reste à signaler que les logiciels statistiques permettent, selon leurs richesses fonctionnelles, d'obtenir tous ou une partie des tests et des mesures d'association présentés dans ce chapitre. Les données à traiter peuvent être fournies aux logiciels statistiques sous formes d'un tableau de fréquences (lignes x colonnes) ou sous formes de données brutes résultant d'une enquête (individus x variables). Dans ce dernier cas, le logiciel permet d'obtenir à la fois le tableau de contingence, les tests statistiques et les mesures d'association souhaités.