UNIVERSITÉ DE PARAKOU

Ecole Nationale de Statistique, de Planification et de Démographie (ENSPD)

Département de Statistique Appliquée

Méthodes statistiques relatives à une ou plusieurs moyennes

MASTER1

01 03 2025

Dr. DICKO Aliou

Description

- Les méthodes statistiques relatives à une ou plusieurs moyennes se rapportent au test de conformité d'une moyenne, le test *t* d'égalité de deux moyennes, le test d'égalité de deux variances ainsi que l'analyse de la variance.

- Comme d'autres types de tests statistiques, l'analyse de variance (ANOVA) compare les moyennes de différents groupes et démontre l'existence de différences statistiques entre les moyennes.

- Bien que l'analyse de variance permette d'identifier une différence, elle ne dit pas quels groupes spécifiques sont statistiquement différents les uns des autres.

Description

- Pour faire cette distinction, les tests de comparaisons particulières et multiples de moyennes sont nécessaires.

- L'analyse de variance multivariée (MANOVA) utilise le même cadre conceptuel que l'ANOVA.

- La MANOVA est une extension de l'ANOVA dans laquelle les effets des facteurs sont évalués sur une combinaison de plusieurs variables réponses.

- L'avantage de l'utilisation d'une MANOVA au lieu de plusieurs ANOVA simultanée réside dans le fait qu'elle prend en compte la corrélation entre les variables réponses et permet ainsi une meilleure utilisation des informations provenant des données.

Objectif général de l'ECU

La formation vise à initier les apprenants aux méthodes statistiques pouvant leur permettre de comparer des moyennes à partir d'échantillons aléatoires simples sur la base des tests d'hypothèse.

Objectifs spécifiques de l'ECU

- A la fin de ce cours l'apprenant doit être capable de :
- Tester la conformité d'une moyenne à une valeur théorique ;
- Comparer les moyennes de deux échantillons (indépendants ou appariés);
- Comparer les moyennes de plusieurs groupes (plus de 2) distinguées par un seul critère (facteur)
- Décrire une analyse de la variance lorsqu'il s'agit de deux critères de classification ;
- Expliquer une analyse de la variance pour trois ou plus de trois critères de classification ;
- Utiliser des outils statistiques de comparaisons multiples de moyennes ;
- Décrire les conditions d'application des méthodes statistiques relatives à une ou plusieurs moyennes ;
- Réaliser une analyse de la variance multivariée (MANOVA)/analyse canonique discriminante ;
- Appliquer les méthodes statistiques relatives à une ou plusieurs moyennes sur l'ordinateur

Prérequis

- Très bonne connaissance en statistique descriptive ;
- Bonne connaissance en informatique (notamment le logiciel Excel) ;
- Bonne connaissance en distribution de probabilité d'une variable aléatoire.

Contenu de la formation

- Méthodes relatives à une ou deux moyennes (test de conformité d'une moyenne, test d'égalité de deux variances et ses variantes) ;
- Analyse de la variance à un critère de classification (décomposition des sommes des carrés et des produits d'écarts, différents tests d'égalité de moyenne, importance du facteur étudié ;
- Analyse de la variance à deux critères de classification ;
- Analyse de la variance à trois et plus de trois critères de classification ;
- Comparaisons particulières et multiples de moyennes (structuration des moyennes et autres, comparaison deux à deux et autres contrastes) ;
- Conditions d'application de ces méthodes ;
- Analyse de la variance multivariée / Analyse canonique discriminante (calcul des variables canoniques, test de signification et importance relative des variables canoniques, interprétation des variables canoniques ;
- Application sur ordinateur et interprétation des résultats (d'analyse).

Méthodes d'enseignement/apprentissage

- Cours théorique ;
- Travaux pratiques ;
- Travaux dirigés ;
- TPE (Etude de cas, exposé).

Lieu d'apprentissage

- Ecole/salle de cours ;
- Salle informatique ;

Matériel pédagogique

- Projecteur ;
- Ordinateur ;
- Note de cours ;
- Tableau et craie.

Compétences générales

- Prendre des initiatives et décisions ;
- Croire que l'on peut surmonter tous les obstacles ;
- Mettre en œuvre des ressources organisationnelles pour produire des résultats ;
- Analyser les problèmes pour y trouver des solutions ;
- Apprendre ;
- Être une personne fiable et responsable ;
- Rédiger des rapports.

Mode d'évaluation

- Evaluation formative ;
- Devoir de table et oral.

Bibliographie

- Bardos M. (2001). *Analyse Discriminante - Application au risque et scoring financier*. Dunod. 332p.
- Dagnelie P. (2013). Statistique théorique et appliquée, 3ème édition, De Boeck Université. 516 p.
- Dalgaard P. (2008). *Introductory Statistics with R*. 2ème édition, Springer. 364p.
- Lafaye de Micheaux P., Drouilhet R., Liquet B. (2014). *Le logiciel R : Maîtriser le langage, effectuer des analyses (bio) statistiques.* 2ème édition, Springer, Paris. 714 p.
- Lebart L., Morineau A., Piron M. (2000). *Statistique Exploratoire Multidimensionnelle*. 4eme édition, Dunod. 480 p.
- Montgomery D.C., Runger G.C. (2013). *Applied Statistics and Probability for Engineers*. 5ème édition, Wiley. 792p.
- Morgenthaler S. (2007). *Introduction à la statistique.* 3ème édition, Presses polytechniques et universitaires romandes. 385p.
- Paradis E. (2005). *R pour les débutants*. 81p.
- Saporta G. (2011). Probabilités, analyse de données et statistique, 3ème édition, Technip, Paris, France. 622 pages.
- Tenenhaus M. (1996). *Méthodes Statistiques en Gestion*. Editions Dunod. 373 p.
- Walpole R. E. (2007). *Probability & statistics for engineers & scientists*. Editions Upper Saddle River, NJ: Pearson Prentice Hall. 816 p.
- Wonnacott T.H., Wonnacott R.J. (1990). *Introductory Statistics*, 5th ed., Wiley. 711p.

0. Généralités sur les tests

0.1. Introduction

- Un **test d'hypothèse** est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations.

- Les méthodes de **l'inférence statistique** nous permettent de déterminer, avec une probabilité donnée, si les **différences constatées** au niveau des échantillons peuvent être **imputables au hasard** ou si elles sont suffisamment importantes pour signifier que les **échantillons proviennent de populations vraisemblablement différentes**.

0. Généralités sur les tests

0.1. Introduction

- Les **tests paramétriques** requièrent un modèle à fortes contraintes (normalité des distributions ou approximation normale pour des grands échantillons).

Ces hypothèses sont d'autant plus difficiles à vérifier que les effectifs étudiés sont plus réduits.

- Les **tests non paramétriques** sont des tests dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon.

Il n'y a pas d'hypothèse de normalité au préalable.

Les tests paramétriques, quand leurs conditions sont remplies, sont les plus puissants que les tests non paramétriques.

Les tests non paramétriques s'emploient lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles transformations de variables.

Ils peuvent s'utiliser même pour des échantillons de taille très faible.

0. Généralités sur les tests

0.1. Introduction

On distingue les tests suivant :

- Le **test de conformité** consiste à confronter un paramètre calculé sur l'échantillon à une valeur pré-établie. Les plus connus sont certainement les tests portant sur la moyenne, la variance ou sur les proportions. On connaît la loi théorique en général la loi normale.

Par exemple, dans un jeu de dés à 6 faces, on sait que la face 3 a une probabilité de 1/6 d'apparaître. On demande à un joueur de lancer (sans précautions particulières) 100 fois le dé, on teste alors si la fréquence d'apparition de la face 3 est compatible avec la probabilité 1/6. Si ce n'est pas le cas, on peut se poser des questions sur l'intégrité du dé.

- Le **test d'ajustement ou d'adéquation** consiste à vérifier la compatibilité des données avec une distribution choisie a priori. Le test le plus utilisé dans cette optique est le test d'ajustement à la loi normale, qui permet ensuite d'appliquer un test paramétrique.

0. Généralités sur les tests

0.1. Introduction

- Le **test d'homogénéité ou de comparaison** consiste à vérifier que K($K \geq 2$) échantillons (groupes) proviennent de la même population ou, cela revient à la même chose, que la distribution de la variable d'intérêt est la même dans les K échantillons.

Y a-t-il une différence entre le taux de glucose moyen mesuré pour deux échantillons d'individus ayant reçu des traitements différents ?

- Le **test d'indépendance ou d'association** consiste à éprouver l'existence d'une liaison entre 2 variables. Les techniques utilisées diffèrent selon que les variables sont qualitatives nominales, ordinales ou quantitatives.

Est-ce que la distribution de la couleur des yeux observée dans la population française fréquences est indépendante du sexe des individus ?

0. Généralités sur les tests

0.2. Principe des tests

0.2.a. Méthodologie

- Le **principe** des tests d'hypothèse est de **poser une hypothèse** de travail et de **prédire les conséquences de cette hypothèse** pour la population ou l'échantillon. On **compare ces prédictions avec les observations** et l'on **conclut en acceptant ou en rejetant l'hypothèse** de travail à partir de règles de décisions objectives.
- Définir les hypothèses de travail, constitue un élément essentiel des tests d'hypothèses de même que vérifier les conditions d'application de ces dernières.
- Différentes étapes doivent être suivies pour tester une hypothèse :

(1) définir l'hypothèse nulle, notée H0, à contrôler ;

(2) choisir une statistique pour contrôler H0 ;

(3) définir la distribution de la statistique sous l'hypothèse « H0 est réalisée » ;

(4) définir le niveau de signification du test **α** et la région critique associée ;

(5) calculer, à partir des données fournies par l'échantillon, la valeur de la statistique ;

(6) prendre une décision concernant l'hypothèse posée.

0. Généralités sur les tests

0.2. Principe des tests

0.2.b. Hypothèse nulle - hypothèse alternative

- ***L'hypothèse nulle*** notée H0 est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage. Cette hypothèse est formulée dans le but d'être rejetée.

- ***L'hypothèse alternative*** notée H1 est la "négation" de H0, elle est équivalente à dire « H0 est fausse ». La décision de rejeter H0 signifie que H1 est réalisée ou H1 est vraie.

- **Remarque** : Il existe une dissymétrie importante dans les conclusions des tests. **En effet, la décision d'accepter H0 n'est pas équivalente à « H_0 est vraie et H_1 est fausse ». Cela traduit seulement l'opinion selon laquelle, il n'y a pas d'évidence nette pour que H_0 soit fausse**. Un test conduit à rejeter ou à ne pas rejeter une hypothèse nulle jamais à l'accepter d'emblée.

0. Généralités sur les tests

0.2. Principe des tests

0.2.b. Hypothèse nulle - hypothèse alternative.

- La nature de H0 détermine la façon de formuler H_1 et par conséquent la nature unilatérale ou bilatérale du test. On parle de ***test bilatéral*** lorsque l'hypothèse alternative se "décompose en deux parties". Par exemple si H_0 consiste à dire que la population estudiantine avec une fréquence de fumeurs p est représentative de la population globale avec une fréquence de fumeurs p_0, on pose alors : $\mathbf{H_0 : p = p_0}$ **et** $\mathbf{H_1 : p \neq p_0}$. Le test sera bilatéral car on considère que la fréquence p peut être supérieure ou inférieure à la fréquence p_0.

- On parle de ***test unilatéral*** lorsque l'hypothèse alternative se "compose d'une seule partie". Par exemple si l'on fait l'hypothèse que la fréquence de fumeurs dans la population estudiantine p est supérieure à la fréquence de fumeurs dans la population p0, on pose alors **H0 : p = p0 et H1 : p > p0**. Le test sera unilatéral car on considère que la fréquence p ne peut être que supérieure à la fréquence p0. Il aurait été possible également d'avoir : $H_0 : p = p_0$ et $H_1 : p < p_0$

0. Généralités sur les tests

0.2. Principe des tests

0.2.c. Statistique et niveau de signification

- Une **statistique** est une fonction des variables aléatoires représentant l'échantillon.
- Le choix de la statistique dépend de la nature des données, du type d'hypothèse que l'on désire contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées . . . .
- La valeur numérique de la statistique obtenue pour l'échantillon considéré permet de distinguer entre H0 vraie et H0 fausse.
- Connaissant la loi de probabilité suivie par la statistique S sous l'hypothèse H0, il est possible d'établir une valeur seuil, S_{seuil} de la statistique pour une probabilité donnée appelée le niveau de signification α du test.
- La région critique $Rc = f(S_{seuil})$ correspond à l'ensemble des valeurs telles que : $P(S \in R_c) = \alpha$. Selon la nature unilatérale ou bilatérale du test, la définition de la région critique varie.

0. Généralités sur les tests

0.2. Principe des tests

0.2.c. Statistique et niveau de signification

Test	Unilatéral $H_0 : t = t_0$		Bilatéral $H_0 : t = t_0$
Hypothèse alternative	$H_1 : t > t_0$	$H_1 : t < t_0$	$H_1 : t \neq t_0$
Niveau de signification	$\mathbb{P}(S > S_{seuil}) = \alpha$	$\mathbb{P}(S < S_{seuil}) = \alpha$	$\mathbb{P}(\lvert S \rvert > S_{seuil}) = \alpha$

0. Généralités sur les tests

0.3. Risques d'erreur

Il existe **deux stratégies** pour prendre une décision en ce qui concerne un test d'hypothèse : la première stratégie fixe à priori la valeur du seuil de signification α et la seconde établit la valeur de la probabilité critique α_{obs} à posteriori.

Règle de décision 1 :

Sous l'hypothèse « H_0 est vraie » et pour un seuil de signification α fixé :

- si la valeur de la statistique S_{obs} calculée appartient à la région critique alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H_1 est acceptée ;
- si la valeur de la statistique S_{obs} n'appartient pas à la région critique alors l'hypothèse H_0 ne peut être rejetée.

0. Généralités sur les tests

0.3. Risques d'erreur

Remarque :

Le choix du niveau de signification ou risque α est lié aux conséquences pratiques de la décision ; en général on choisira α = 0, 05, 0, 01 ou 0, 001.

Règle de décision 2 :

La probabilité critique α telle que $P(S \geq S_{obs}) = \alpha_{obs}$ est évaluée

- si $\alpha_{obs} \geq \alpha$ l'hypothèse H_0 est acceptée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est trop important ;
- si $\alpha_{obs} < \alpha$ l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est très faible.

C'est cet α_{obs} que les logiciels d'analyse affichent

0. Généralités sur les tests

0.3. Risques d'erreur

Définition 1.
On appelle risque d'erreur de première espèce la probabilité de rejeter H0 et d'accepter H1 alors que H0 est vraie.

Ceci se produit si la valeur de la statistique de test tombe dans la région de rejet alors que l'hypothèse H0 est vraie. La probabilité de cet évènement est le niveau de signification α. On dit aussi que le niveau de signification est la probabilité de rejeter l'hypothèse nulle à tort.

Remarque : La valeur du risque **α** doit être fixée a priori par l'expérimentateur et jamais en fonction des données. C'est un compromis entre le risque de conclure à tort et la faculté de conclure.

0. Généralités sur les tests

0.3. Risques d'erreur

- **Exemple** : Si l'on cherche à tester l'hypothèse qu'une pièce de monnaie n'est pas « truquée », nous allons adopter la règle de décision suivante :

- H0 : la pièce n'est pas truquée

– est acceptée si $X \in [40, 60]$

– rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$ avec X « nombre de faces » obtenus en lançant 100 fois la pièce. Le risque d'erreur de première espèce est $\alpha = P(B(100, 1/2) \in [40, 60])$.

0. Généralités sur les tests
0.3. Risques d'erreur

Définition 2. On appelle risque d'erreur de seconde espèce, notée β la probabilité de rejeter H1 et d'accepter H0 alors que H1 est vraie.

Ceci se produit si la valeur de la statistique de test ne tombe pas dans la région de rejet alors que l'hypothèse H1 est vraie.

Remarque : Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique sous l'hypothèse H1.

0. Généralités sur les tests

0.3. Risques d'erreur

Exemple : Si l'on reprend l'exemple précédent de la pièce de monnaie, et que l'on suppose la probabilité d'obtenir **face** est de 0, 6 pour une pièce truquée. En adoptant toujours la même règle de décision :

H0 : la pièce n'est pas truquée

– est acceptée si $X \in [40, 60]$

– rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$

avec X « nombre de faces » obtenues en lançant 100 fois la pièce. Le risque de seconde espèce est $\beta = P(B(100, 0, 6) \in [40, 60])$.

0. Généralités sur les tests

0.4. Puissance d'un test

Rappelons que les tests ne sont pas faits pour « démontrer » H_0 mais pour « rejeter » H_0.

L'aptitude d'un test à rejeter H0 alors qu'elle est fausse constitue la puissance du test.

Définition 3.

On appelle puissance d'un test, la probabilité de rejeter H0 et d'accepter H1 alors que H1 est vraie. Sa valeur est $1 - \beta$

La puissance d'un test est fonction de la nature de H1, un test unilatéral est plus puissant qu'un test bilatéral.

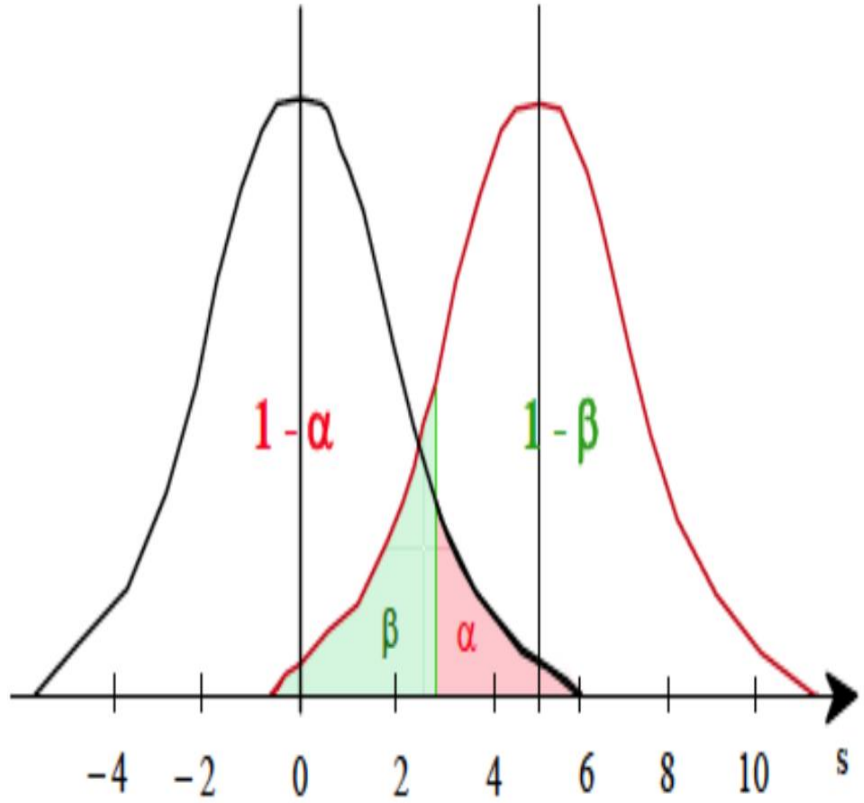
Elle augmente avec taille de l'échantillon N étudié, et diminue lorsque α diminue.

0. Généralités sur les tests

0.4. Puissance d'un test

Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans le tableau suivant :

Décision Réalité	H_0 vraie	H_1 vraie
H_0 acceptée	correct	manque de puissance risque de seconde espèce β
H_1 acceptée	rejet à tort risque de premières espèce α	puissance du test $1-\beta$



Méthodes statistiques relatives à une ou plusieurs moyennes

1. **Test de conformité d'une moyenne**

Cette partie de la statistique, qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations mais extrapole des observations faites sur un ensemble limité à un ensemble plus large, permet de tester des hypothèses sur cet ensemble et de prendre des décisions à leur sujet.

Méthodes statistiques relatives à une ou plusieurs moyennes

1. Test de conformité d'une moyenne

1.1. Méthodes statistiques relatives aux valeurs moyennes

Toutes les méthodes que nous allons présenter en relation avec les moyens supposent que les conditions suivantes soient remplies :

- Normalité des populations initiales,
- Aléa et simplicité des échantillons tirés,
- Pour certains tests, en plus, égalité des variances des populations.

La première condition n'est pas indispensable pour les grands nombres ($N > 30$).

1. Test de conformité d'une moyenne

1.2. Intervalle de confiance et test de conformité d'une moyenne

1.2.1. Intervalle de confiance

- *Dans le cas où la* ***variance de la population parente est connue***, les limites de confiance $\bar{X}$ la moyenne $\bar{X} = \bar{x} \pm \mu_{1-\alpha/2} \frac{s}{\sqrt{n}}$

α peut prendre habituellement la valeur (0.05, 0.01 ou 0.001). $\mu_{1-\alpha/2}$ est tirée à partir de la table de la distribution normale réduite.

- *Dans le* ***cas où la variance de la population parente est inconnue***, alors il faut l'estimer et ceci provoque l'élargissement de l'intervalle de confiance. Cet élargissement est obtenu pour un degré de confiance (1- α) en remplaçant la valeur $\mu_{1-\alpha/2}$ de la distribution normale par la valeur $t_{1-\alpha/2}$ de la distribution de Student à (n – 1) degrés de liberté. Les limites de confiance de la moyenne $\bar{x}$ estimée sont ainsi : $\bar{X} = \bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}$

1. **Test de conformité d'une moyenne**

1.2. Intervalle de confiance et test de conformité d'une moyenne

1.2.1. Intervalle de confiance

S représente l'écart type estimé et $t_{1-\alpha/2}$ est tirée à partir de la table de la distribution de Student pour (n – 1) ddl.

En pratique cette formule est remplacée par : $\bar{X} = \bar{x} \pm t_{1-\alpha/2}\sqrt{\frac{SCE}{n(n-1)}}$

Lorsque n est supérieur à 30 et α = 0.05 $\bar{X} = \bar{x} \pm 2\sqrt{\frac{SCE}{n(n-1)}}$

1. Test de conformité d'une moyenne

1.2. Intervalle de confiance et test de conformité d'une moyenne

1.2.2. Test de conformité

Le test de conformité d'une moyenne a pour but de vérifier si la moyenne d'une population est ou n'est pas égale à une valeur donnée $\overline{X}_0$

On rejette l'hypothèse d'égalité $\overline{X} = \overline{X}_0$

lorsque la moyenne observée est trop différente de la valeur théorique.

Le test est réalisé en calculant la valeur suivante : $$t_{obs.} = \frac{|\overline{X} - \overline{X}_0|}{\sqrt{\frac{SCE}{n(n-1)}}}$$

On rejette l'hypothèse H0 : $\overline{X} = \overline{X}_0$ si $t_{obs.} \geq t_{1-\alpha/2}$ pour (n-1) ddl. et l'accepter si $t_{obs.} < t_{1-\alpha/2}$ pour (n-1) ddl.

1. **Test de conformité d'une moyenne**

1.3. Exercice

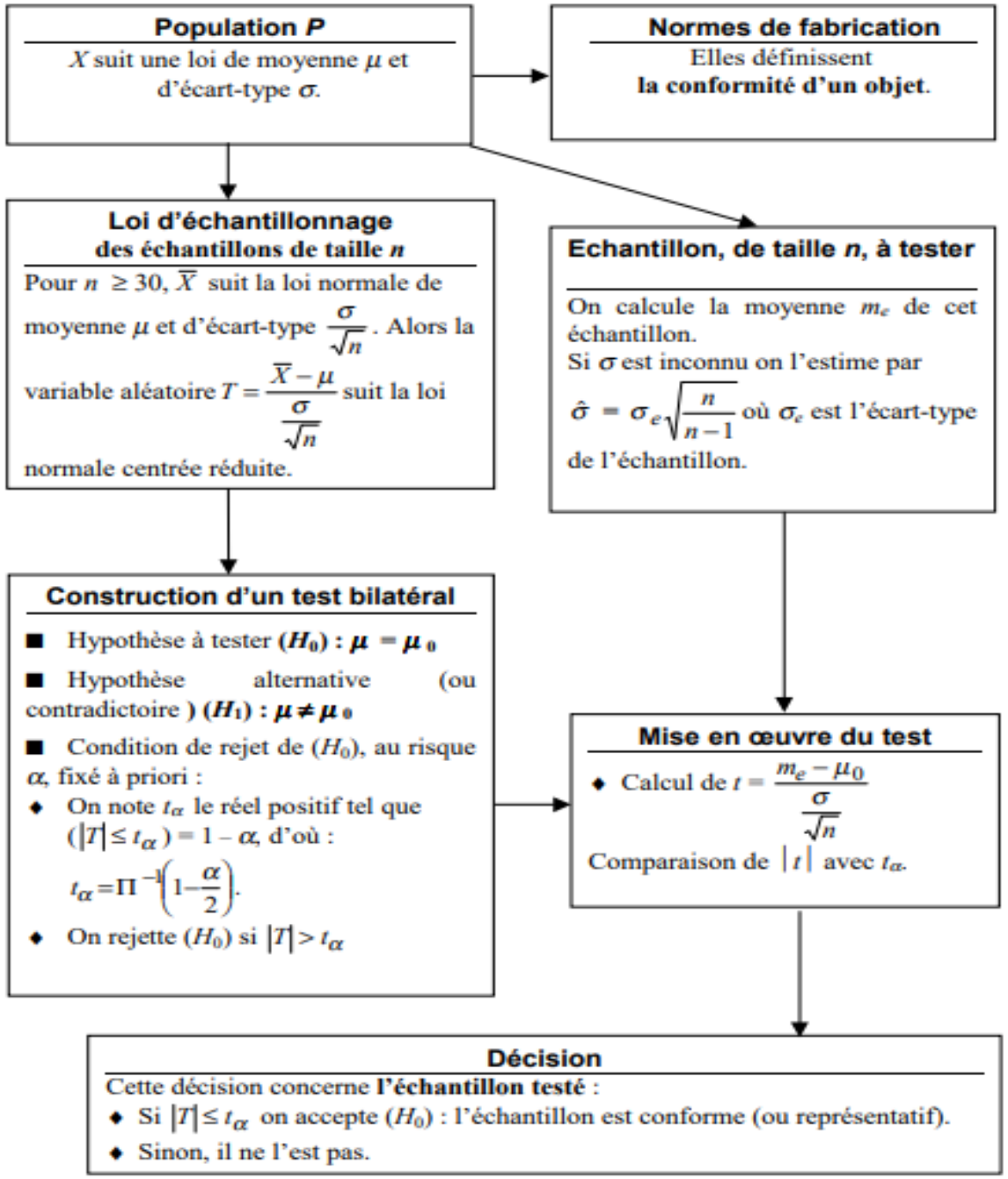
Exercice 1

On extrait d'une population, dont le caractère masse est distribué selon le modèle normal, un échantillon aléatoire simple de 7 éléments. La moyenne de cet échantillon est 257,9. L'écart-type de la population est supposé connu et égal à 14. Est-il possible, au seuil de 5%, de conclure que la moyenne de la population est supérieure à 250 ?

Méthodes statistiques relatives à une ou plusieurs moyennes

1. Test de conformité d'une moyenne

1.4. Synthèse



2. Test de conformité de la variance

2.1. Moyenne connue

On suppose que l'on a un échantillon qui suit une loi normale N (μ, σ^2) où la moyenne est connue. On veut tester $H_0 : \sigma^2 = \sigma^2_0$ contre $H_1 : \sigma^2 \neq \sigma^2_0$.

Sous l'hypothèse H_0 la statistique. $$V = \frac{n\overline{S_n}^2}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma_0} \right)^2$$

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

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\alpha/2}(n) \leq \frac{n\overline{S_n}^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}(n) \right) = 1 - \alpha$$

avec $\chi^2_{\alpha/2}$ (n) et $\chi^2_{1-\alpha/2}$ (n) les quantiles d'ordre α/2 et 1 − α/2 de la loi χ^2 (n).

Donc la région de rejet est : $[0, \chi^2_{\alpha/2}(n)[\cup]\chi^2_{1-\alpha/2}(n), +\infty[$

2. Test de conformité de la variance

2.1. Moyenne connue

On calcule alors pour les valeurs de l'échantillon, V, et on accepte ou on rejette au risque α H0 suivant la valeur trouvée.

Si on a une hypothèse alternative H1 : $\sigma^2 > \sigma^2_0$ on fera un test unilatéral, et obtient au risque α

$$\mathbb{P}\left(\frac{n\overline{S_n}^2}{\sigma^2} \leq \chi^2_{1-\alpha}(n)\right) = 1 - \alpha$$

avec $\chi^2_{1-\alpha}(n)$ le quantile d'ordre $1 - \alpha$ de la loi χ^2 (n). Donc la région de rejet est] $\chi^2_{1-\alpha}(n)$, $+\infty$[

2. Test de conformité de la variance

2.2. Moyenne inconnue

- On suppose que l'on a un échantillon qui suit une loi normale N (μ, σ^2) où la moyenne est inconnue.
- On veut tester $H_0 : \sigma^2 = \sigma^2_0$ contre H1 : $\sigma^2 \neq \sigma^2_0$.
- Sous l'hypothèse H0 la statistique

$$V = \frac{n\overline{S'_n}^2}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \overline{X_n}}{\sigma_0} \right)^2$$

suit une loi du χ^2 à $n - 1$ degrés de libertés.

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\alpha/2}(n-1) \leq \frac{n\overline{S'_n}^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}(n-1) \right) = 1 - \alpha$$

avec $\chi^2_{\alpha/2}(n-1)$ et $\chi^2_{1-\alpha/2}(n-1)$ les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de la loi $\chi^2(n-1)$.

Donc la région de rejet est :

$$[0, \chi^2_{\alpha/2}(n-1)[\cup]\chi^2_{1-\alpha/2}(n-1), +\infty[$$

On calcule alors pour les valeurs de l'échantillon, V , et on accepte ou on rejette au risque α H_0 suivant la valeur trouvée.

2. Test de conformité de la variance

2.3. Exercice

Exercice

On extrait d'une fabrication, dont le caractère masse est distribué selon le modèle normal, un échantillon aléatoire simple de taille 18. Les masses, exprimées en grammes, des 18 éléments de cet échantillon sont les suivantes : 304 334 307 309 330 314 310 316 309 314 299 311 348 290 311 309 326 278. La moyenne de la fabrication est supposée inconnue. Est-il possible, au seuil de 5% de conclure que la variance de la fabrication est différente de 605 ?

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

Si les deux échantillons on la même taille n1 = n2 = n. Le test se ramène à un test à une moyenne nulle de l'échantillon $(Z_1, ..., Z_n)$, avec $Z_i = X_i - Y_i$

3.1. Variance connue

On suppose que l'on a deux échantillons $(X_1, \ldots, X_{n1})$ et (Y1, . . . , Yn2)qui suivent une loi normale N (μ1, σ2 1) et N (μ2, σ2 2) où les variances sont connues.

On veut tester H0 : μ1 = μ2 contre H1 : μ1 ≠ μ2, c'est le cas bilatéral.

Sous l'hypothèse H0 la variable aléatoire $\overline{X_{n_1}} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k$ suit une loi N (μ_1, σ^2_1/n1)

et la variable aléatoire $\overline{Y_{n_2}} = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k$ uit une loi N (μ_2, σ^2_2/n2),

par conséquent la statistique $U = \dfrac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ une loi normale centrée réduite.

3. Test de comparaison de deux moyennes (Echantillons indépendants)

3.1. Variance connue

Pour un risque d'erreur α fixé on a donc : $\mathbb{P}(|U| \leq q_{1-\alpha/2}) = 1 - \alpha$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi N (0, 1); et donc la région de rejet est : $]-\infty, q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[$

- On calcule alors pour les valeurs de l'échantillon U et on accepte ou on rejette H0 suivant la valeur trouvée, au risque α. $\mathbb{P}(Z \leq q_{1-\alpha}) = 1 - \alpha$
- Si on considère un test unilatéral et une hypothèse alternative H1 : $\mu_1 > \mu_2$ par exemple, on obtient pour un risque d'erreur α.

avec $q_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi N (0, 1); et donc la région de rejet est $]q_{1-\alpha}, +\infty[$

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

3.2. Variance inconnue

On suppose que l'on a que l'on a deux échantillons $(X_1, \ldots, X_{n1})$ et $(Y_1, \ldots, Y_{n2})$ qui suivent une loi normale $N(\mu_1, \sigma^2_1)$ et $N(\mu_2, \sigma^2_2)$ où les variances sont inconnues.

Cas 1 : n1 et n2 supérieurs à 30.

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ une loi $N(0, \sigma^2_1/n_1 + \sigma^2_2/n_2)$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\overline{S'_{n_1}}^2 + \overline{S'_{n_2}}^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2$$

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

3.2. Variance inconnue.

- Alors la variable aléatoire $N = \dfrac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2}}$ peut être approximé par une loi normale centrée réduite.

- Pour un risque d'erreur α fixé on a donc $P(|N| \leq t_{1-\alpha/2}) = 1 - \alpha$
- avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite ; et donc la région de rejet est $]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$.

- On calcule alors pour les valeurs de l'échantillon N et on accepte ou on rejette H0 suivant la valeur trouvée, au risque α.

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

3.2. Variance inconnue.

Cas 2 : n1 ou n2 inférieur à 30 et σ1 = σ2

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ suit une loi $N(0, \sigma^2_1/n_1 + \sigma^2_2/n_2)$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée :

$$\overline{S'_{n_1,n_2}}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2\right)$$

Alors la variable aléatoire $T = \dfrac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1,n_2}}^2(1/n_1 + 1/n_2)}}$ suit une de Student à $n_1 + n_2 - 2$ degrés de liberté.

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

3.2. Variance inconnue.

Pour un risque d'erreur α fixé on a donc :

$$\mathbf{P(|T| \leq t_{1-\alpha/2}) = 1 - \alpha}$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à n_1+n_2-2 degrés de liberté ; et donc la région de rejet est $]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H0 suivant la valeur trouvée, au risque α

3. Test de comparaison de deux variances (Echantillons indépendants)

3.2. Variance inconnue.

Cas 3 : n_1 ou n_2 inférieur à 30 et $\sigma_1 \neq \sigma_2$

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, dans le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_{n_1}} - \overline{Y_{n_2}}$ suit une loi N (0, σ^2_1/n1 + σ^2_2/n2). Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\overline{S'_{n_1}}^2 + \overline{S'_{n_2}}^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2 + \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2$$

Alors la variable aléatoire $T = \dfrac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2}}$ suit une de Student à ν degrés de liberté, où ν est l'entier le plus proche de $\dfrac{(\overline{S'_{n_1}}^2/n_1 + \overline{S'_{n_2}}^2/n_2)^2}{(n_1 - 1)\overline{S'_{n_1}}^4/n_1^4 + (n_2 - 1)\overline{S'_{n_2}}^4/n_2^4}$

3. **Test de comparaison de deux moyennes (Echantillons indépendants)**

3.2. Variance inconnue

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|T| \leq t_{1-\alpha/2}) = 1 - \alpha$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student précédente ; et donc la région de rejet est $]-\infty, t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H0 suivant la valeur trouvée, au risque α

Conditions de validité conseillées:

– n > 30 ou distribution normale, en fait ces deux conditions sont à considérer ensemble

– Variances égales (sinon approximation de Welch)

4. **Test de comparaison de deux moyennes (Echantillons associés / pariés)**

- Un autre cas important de comparaison de moyennes est relatif aux échantillons dont les individus sont associés par paires ou par couples.
- Ce cas se présente, par exemple, on compare deux méthodes de mesure en soumettant à ces deux méthodes les mêmes individus, tirés d'une population donnée.
- A chacune des méthodes correspond alors une population de mesures, mais les populations et les échantillons extraits ne sont pas indépendants.

4. **Test de comparaison de deux moyennes (Echantillons associés)**

Pour tester l'égalité des moyennes, on doit alors considérer la population des différences et vérifier la nullité de sa moyenne.

On remplace alors le test d'égalité de deux moyennes par un test de conformité d'une moyenne.

Les conditions d'application du test sont :

- caractère aléatoire et simple de l'échantillon
- normalité de la population des différences.

Le test se réalise comme suit :

On pose l'hypothèse nulle : (H0) : = 0

On calcule les différences : $\bar{d}$

E_1	E_2	Différences (d_i)
x_{11}	x_{21}	$x_{11} - x_{21} = d_1$
x_{12}	x_{22}	$x_{12} - x_{22} = d_2$
…	…	…
x_{1n}	x_{2n}	$x_{1n} - x_{2n} = d_n$

4. **Test de comparaison de deux moyennes (Echantillons associés)**

On calcule la quantité suivante : $$t_{obs.} = \frac{|\bar{d}|}{\sqrt{\frac{SCEd}{n\,(n-1)}}}$$

- On rejette l'hypothèse $H0 : (\bar{d} = 0)$ si $t_{obs.} \geq t_{1-\alpha/2}$ pour (n - 1) ddl
- et l'accepter si tobs.< $t_{1-\alpha/2}$ pour (n - 1) ddl.

Ce test est appelé **test de Student par couples**.

Quand on rejette l'hypothèse de nullité de la moyenne des différences, nous pouvons calculer l'intervalle de confiance de la différence.

$$\bar{d} \pm t_{1-\alpha/2} \sqrt{\frac{SCEd}{n\,(n-1)}}$$

4. **Test de comparaison de deux moyennes (Echantillons associés)**

Dans le cas où les conditions d'application ne sont pas satisfaites, la normalisation de la distribution par une transformation de la variable mesurée est indispensable.

Si la transformation ne donne pas un résultat, on s'oriente alors vers les méthodes de la statistique non paramétrique :

test de Wilcoxon pour échantillons associés, test de Page, test des signes etc…

5. Test de comparaison de deux variances

Avec les mêmes notations que précédemment on teste

$H_0 : \sigma^2_1 = \sigma^2_2$ contre $H_1 : \sigma^2_1 \neq \sigma^2_2$

On considère $\overline{S'^2_{n_1}} = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \overline{X_{n_1}})^2$ et $\overline{S'^2_{n_2}} = \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \overline{Y_{n_2}})^2$

ainsi que la statistique $Z = \frac{\overline{S'^2_{n_1}}}{\overline{S'^2_{n_2}}}$

Sous l'hypothèse H0 la statistique Z suit une **loi de Fisher-Snedecor** $F(n_1 - 1, n_2 - 1)$ à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

Pour un risque d'erreur α fixé on a une région de rejet $\mathbf{[0, F_{\alpha/2}(n)[\cup]F_{1-\alpha/2}(n), +\infty[}$ où les quantiles sont déterminées à l'aide de la loi précédente.

5. **Test de comparaison de deux variances**

Quelques précisions sur la loi de Fisher Snedecor :

Définition

Si U et V sont deux variables aléatoires indépendantes distribuées respectivement selon les lois de $\chi^2(\nu_1)$ et de $\chi^2(\nu_2)$,

alors la loi de la variable aléatoire : $W = \dfrac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}}$

est distribuée selon la loi de **Fisher-Snedecor** à $(\nu_1 ; \nu_2)$ degrés de liberté (ν_1 degrés de liberté au numérateur et ν_2 degrés de liberté au dénominateur), notée $F(\nu_1 ; \nu_2)$.

5. **Test de comparaison de deux variances**

Conséquence immédiate :

Si W est une variable aléatoire distribuée selon la loi F(ν1 ; ν2), alors $\frac{1}{W}$

est distribuée selon la loi F(ν_2 ; ν_1)

Propriété

Le quantile d'ordre α de la loi F(ν_1 ; ν_2), est l'inverse du quantile d'ordre 1 − α de la loi F(ν_2 ; ν_1). On a donc la relation :

$$f_\alpha(\nu_1 ; \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2 ; \nu_1)}$$

On note $f_\alpha(\nu_1 ; \nu_2)$ le nombre réel tel que : $P\left(F < f_\alpha(\nu_1 ; \nu_2)\right) = \alpha$

5. Test de comparaison de deux variances

Cas d'échantillons issus de populations gaussiennes

Or, si on dispose de deux échantillons aléatoires simples et indépendants de tailles respectives n_1 et n_2 issus de deux lois mères gaussiennes, alors les variables aléatoires

$\frac{n_1 S_1^2}{\sigma_1^2}$ et $\frac{n_2 S_2^2}{\sigma_2^2}$ sont indépendantes et distribuées respectivement selon les lois

du $\chi^2 (n_1 - 1)$ et du $\chi^2 (n_2 - 1)$.

Sous l'***hypothèse d'égalité des variances*** ($\sigma^2_1 = \sigma^2_2$),

on en déduit que la variable aléatoire $\dfrac{\frac{n_1 S_1^2}{n_1 - 1}}{\frac{n_2 S_2^2}{n_2 - 1}}$ est distribuée selon la loi de $F(n_1 - 1 ; n_2 - 1)$

5. **Test de comparaison de deux variances**

Test bilatéral :

Hypothèses : H_0 : "$\sigma^2_1 = \sigma^2_2$ " et H_1 : "$\sigma^2_1 \neq \sigma^2_2$ " Sous l'hypothèse H_0, la variable

$F = \dfrac{\dfrac{n_1 S_1^2}{n_1 - 1}}{\dfrac{n_2 S_2^2}{n_2 - 1}}$ est distribuée selon la loi de Fisher Snedecor $F(n_1 - 1 ; n_2 - 1)$

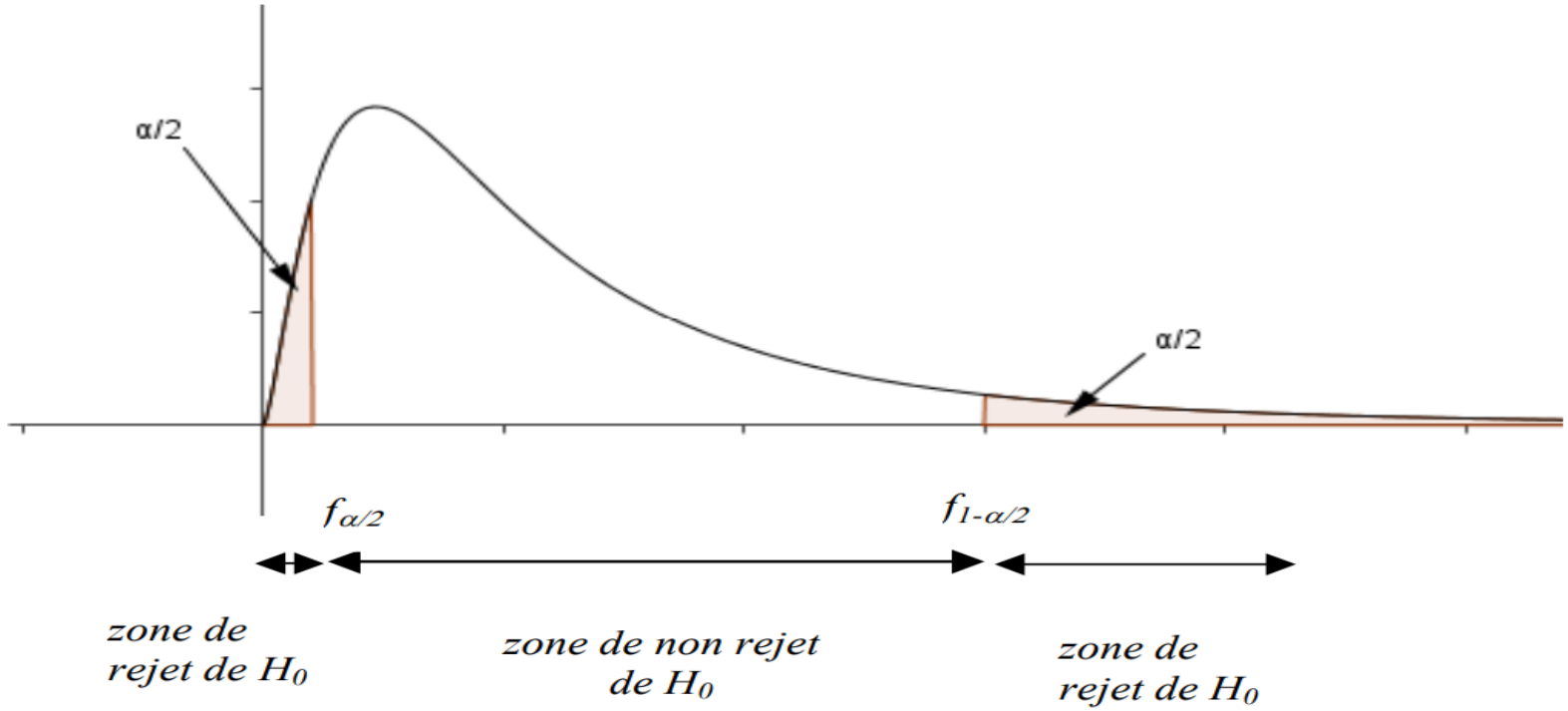
Dans ce cas, le test de comparaison étant bilatéral, on rejette H_0 au seuil de risque α dans les deux cas suivants : $f_{obs} \leq f_{\alpha/2}(\nu_1 ; \nu_2)$ ou $f_{obs} \geq f_{1-\alpha/2}(\nu_1 ; \nu_2)$

On a la relation suivante : $$P\left(f_{\frac{\alpha}{2}}(n_1 - 1 ; n_2 - 1) < F < f_{1-\frac{\alpha}{2}}(n_1 - 1 ; n_2 - 1)\right) = 1 - \alpha.$$

5. **Test de comparaison de deux variances**

D'où la règle de décision : Si $f_{obs} = \dfrac{\dfrac{n_1 s_1^2}{n_1 - 1}}{\dfrac{n_2 s_2^2}{n_2 - 1}} \notin \left[f_{\frac{\alpha}{2}}(n_1 - 1 ; n_2 - 1) ; f_{1-\frac{\alpha}{2}}(n_1 - 1 ; n_2 - 1) \right]$

on rejette H_0 au seuil de risque α. Dans le cas contraire, on n'est pas en situation de rejeter H0.



5. **Test de comparaison de deux variances**

Exercice

Une expérimentation a été menée dans le but de résoudre des problèmes liés à l'intensification de l'agriculture et particulièrement à une nouvelle méthode d'engraissement des bovins. La race traditionnelle *A. Angus* n'étant pas adaptée à ce système d'élevage, un croisement : *A.Angus x Charolaise* a été créé.

L'objectif est d'obtenir des animaux mieux adaptés à ces nouvelles pratiques tout en maintenant une homogénéité comparable à celle de race traditionnelle.

Le caractère étudié est le GMQ (Gain Moyen Quotidien) exprimé en kg. Les résultats observés sur deux lots (ici au sens " échantillons") sont les suivants :

- pour le lot 1 : race pure, taille de l'échantillon : 16, variance : 0,26.
- pour le lot 2 : race croisée, taille de l'échantillon : 21, variance : 0,37.

Peut-on considérer que le GMQ du croisement *A .Angus* x *Charolaise* donne des résultats aussi homogènes que celui de la race pure ? (on prendra un seuil de risque de 0,05).

5. **Test de comparaison de deux variances**

Exercice

Notations

On définit la variable aléatoire X_1 (respectivement X_2) qui, à chaque vache de race *A.Angus* (respectivement *A.Angus* x Charolaise) prélevée au hasard, associe son GMQ. Ces deux variables sont supposées distribuées normalement de variances respectives σ^2_1 et σ^2_2.

Soit $S^2 1$ (respectivement $S^2 2$) la variable aléatoire qui à chaque échantillon de n1 vaches *A.Angus* (respectivement n2 vaches *A.Angus x Charolaise*) associe sa variance notée $s^2 1$ (respectivement $s^2 2$).

Les échantillons sont supposés aléatoires simples et indépendants.

5. Test de comparaison de deux variances

***Exemple 1* : Eléments de correction :**

Hypothèses : H_0 : "$\sigma_1 = \sigma_2^2$" et H_1 : "$\sigma_1^2 \neq \sigma_2^2$"

La variable aléatoire $F = \dfrac{\dfrac{n_1 S_1^2}{n_1 - 1}}{\dfrac{n_2 S_2^2}{n_2 - 1}}$ est distribuée selon la loi de F(15 ; 20).

Calcul de la valeur observée : $f_{obs} = \dfrac{\dfrac{n_1 s_1^2}{n_1 - 1}}{\dfrac{n_2 s_2^2}{n_2 - 1}} = \dfrac{\dfrac{16 \times 0{,}26}{15}}{\dfrac{21 \times 0{,}37}{20}} \simeq 0{,}71.$

5. Test de comparaison de deux variances

```
> qf(0.025, 15,20)
[1] 0.3628576
```

On détermine les valeurs critiques au risque de 0,05 :

$f_{0,025}(15\,;20) = \frac{1}{f_{0,975}(20\,;15)} \simeq \frac{1}{2,76} \simeq 0,36$ et $f_{0,975}(15\,;20) \simeq 2,57$.

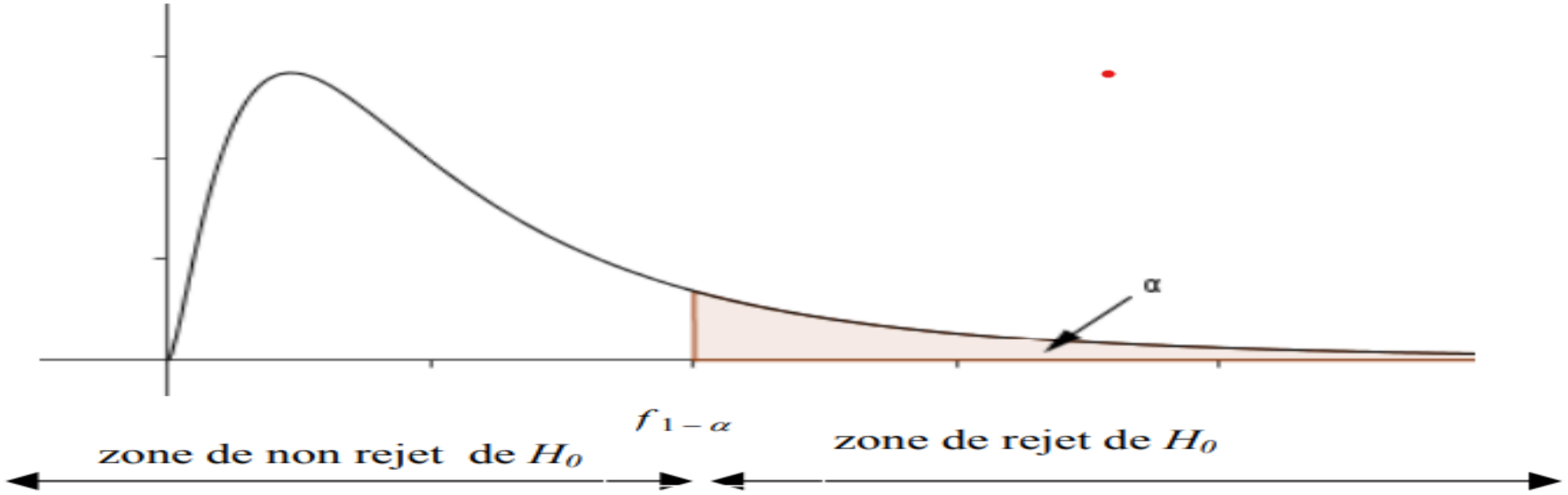
```
> qf(0.975, 15,20)
[1] 2.573096
```

$0,71 \in [0,36\,;2,57]$. On ne peut pas en déduire que les variances du caractère GMQ des deux populations sont différentes, donc on accepte H_0.

Test unilatéral

Hypothèses : H_0 : "$\sigma_1 = \sigma_2^2$" et H_1 : "$\sigma_1^2 > \sigma_2^2$"

Règle de décision : si $f_{obs} = \frac{\frac{n_1 s_1^2}{n_1 - 1}}{\frac{n_2 s_2^2}{n_2 - 1}} > f_{1-\alpha}(n_1 - 1\,;n_2 - 1)$, on rejette H_0 au seuil de risque α.



6. **Analyse de la variance à un facteur**

6.1. Objectif

Dans ce chapitre, nous allons étudier un test statistique (nous renvoyons au cours sur les tests pour toutes les définitions sur ce sujet) permettant de comparer les moyennes de plusieurs variables aléatoires indépendantes gaussiennes de même variance.

L'analyse de la variance est l'une des procédures les plus utilisées dans les applications de la statistique ainsi que dans les méthodes d'analyse de données.

6. **Analyse de la variance à un facteur**

6.2. Exemple introductif

Une étude de reproductibilité a été menée pour étudier les performances de trois laboratoires relativement à la détermination de la quantité de sodium de lasalocide dans de la nourriture pour de la volaille.

Une portion de nourriture contenant la dose nominale de 85 mg kg−1 de sodium de lasalocide a été envoyée à chacun des laboratoires à qui il a été demandé de procéder à 10 réplications de l'analyse.
Les mesures de sodium de lasalocide obtenues sont exprimées en mg kg−1. Elles ont été reproduites sur le transparent suivant.

La reproductibilité d'une expérience scientifique est une des conditions qui permettent d'inclure les observations réalisées durant cette expérience dans le processus d'amélioration perpétuelle des connaissances scientifiques. Cette condition part du principe qu'on ne peut tirer de conclusions que d'un événement bien décrit, qui est apparu plusieurs fois, provoqué par des personnes différentes. Cette condition permet de s'affranchir d'effets aléatoires venant fausser les résultats ainsi que des erreurs de jugement ou des manipulations de la part des scientifiques.

6. **Analyse de la variance à un facteur**

6.2. Exemple introductif

TABLE: Source : Analytical Methods Committee, Analyst, 1995

	Laboratoire		
	A	*B*	*C*
1	87	88	85
2	88	93	84
3	84	88	79
4	84	89	86
5	87	85	81
6	81	87	86
7	86	86	88
8	84	89	83
9	88	88	83
10	86	93	83

Remarque

Cette écriture du tableau est dite « désempilée ».

Nous pouvons l'écrire sous forme standard (« empilée »), c'est-à-dire avec deux colonnes, une pour le laboratoire et une pour la valeur de sodium de lasalocide mesurée, et trente lignes, une pour chacune des observations réalisées.

6. **Analyse de la variance à un facteur**

6.2. Exemple introductif

Tableau empilé de l'exemple des laboratoires

Essai	Laboratoire	Lasalocide
1	Laboratoire *A*	87
2	Laboratoire *A*	88
3	Laboratoire *A*	84
4	Laboratoire *A*	84
5	Laboratoire *A*	87
6	Laboratoire *A*	81
7	Laboratoire *A*	86
8	Laboratoire *A*	84
9	Laboratoire *A*	88
10	Laboratoire *A*	86

Essai	Laboratoire	Lasalocide
11	Laboratoire *B*	88
12	Laboratoire *B*	93
13	Laboratoire *B*	88
14	Laboratoire *B*	89
15	Laboratoire *B*	85
16	Laboratoire *B*	87
17	Laboratoire *B*	86
18	Laboratoire *B*	89
19	Laboratoire *B*	88
20	Laboratoire *B*	93

Essai	Laboratoire	Lasalocide
21	Laboratoire *C*	85
22	Laboratoire *C*	84
23	Laboratoire *C*	79
24	Laboratoire *C*	86
25	Laboratoire *C*	81
26	Laboratoire *C*	86
27	Laboratoire *C*	88
28	Laboratoire *C*	83
29	Laboratoire *C*	83
30	Laboratoire *C*	83

6. **Analyse de la variance à un facteur**

6.2. Exemple introductif

Définitions

Pour ce ***exemple***, sur **chaque essai**, nous observons **deux variables**.

1. *Le laboratoire*. Il est totalement contrôlé. La variable « Laboratoire » est considérée comme qualitative avec trois modalités bien déterminées. Nous l'appelons le **facteur**. Ici le facteur « Laboratoire » est à **effets fixes**.

2. *La quantité de Lasalocide*. La variable « Lasalocide » est considérée comme quantitative comme généralement tous les résultats obtenus par une mesure. Nous l'appelons **la réponse**.

6. **Analyse de la variance à un facteur**

6.2. Exemple introductif

Notations

La variable mesurée dans un tel schéma expérimental sera notée Y. Pour les observations nous utilisons deux indices :

- le premier indice indique le numéro du groupe dans la population (« Laboratoire »),
- le second indice indique le numéro de l'observation dans l'échantillon (« Essai »).

Signification des indices

Pour le **premier indice**, nous utilisons i (ou encore i', i'', i_1, i_2).
Pour le **second indice**, nous utilisons j (ou encore j' , j'', j_1, j_2).

6. **Analyse de la variance à un facteur**

6.3. Notations et définitions

Notation

Ainsi les observations sont en général notées par :

y_{ij}, $\quad i = 1, \ldots, I \quad j = 1, \ldots, J(i)$.

Définition

Lorsque les échantillons sont de même taille, à savoir J(i) = J et ce quelque soit i, nous disons que **l'expérience est équilibrée**.

Remarque

Si les tailles des échantillons sont différentes, alors elles sont notées par : ni , où i = 1, . . . , I.
Mais ce plan expérimental est à éviter parce que les différences qu'il est alors possible de détecter sont supérieures à celles du schéma équilibré.

6. **Analyse de la variance à un facteur**

6.3. Notations et définitions

Définition

En se plaçant dans le cas équilibré nous notons les moyennes de chaque échantillon par :

$$\overline{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij}, \quad i = 1, \ldots, I,$$

et les variances de chaque échantillon par : $s_i^2(y) = \frac{1}{J}\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2, \quad i = 1, \ldots, I.$

Remarque

Cette dernière formule exprime la variance non corrigée. Très souvent, dans les ouvrages ou les logiciels, c'est la variance corrigée qui est utilisée : au lieu d'être divisée par J, la somme est divisée par J – 1

6. **Analyse de la variance à un facteur**

Application sur l'exemple «*Les laboratoires*»

Nous allons d'abord importer les données sous R, en utilisant les lignes de commande suivantes :

```
> laboratoire<-rep(1:3,c(10,10,10))
>  quantite<-c(87,88,84,84,87,81,86,84,88,86,  88,93,88,89,85,87,86,89,88,93,85,84,79,86,81,86,88,83,83,83)
> jeutotal<-data.frame(laboratoire,quantite)
> moy<-tapply(jeutotal$quantite, jeutotal$laboratoire,mean)
> moy
> sd<-tapply(jeutotal$quantite, jeutotal$laboratoire,sd)
> sd
```

Nous obtenons donc : $\bar{y}_1 = 85,500 \quad \bar{y}_2 = 88,600$ $\bar{y}_3 = 83,800.$ et $s_{1,c}(y) = 2,224 \quad s_{2,c}(y) = 2,633$ $s_{3,c}(y) = 2,616.$

Le nombre total d'observations est égal à : n = IJ = 3 × 10 = 30.

6. Analyse de la variance à un facteur

6.4. Conditions fondamentales de l'ANOVA

Les résidus $\{e_{bij}\}$ sont associés, sans en être des réalisations, aux variables erreurs $\{\varepsilon ij\}$ qui sont inobservables et satisfont aux trois conditions suivantes :

- Elles sont **indépendantes**.
- Elles sont de **loi gaussienne**.
- Elles ont **même variance** σ^2 inconnue. C'est la condition d'**homogénéité** ou d'**homoscédasticité**.

Remarque

Par conséquent ces trois conditions se transfèrent sur les variables aléatoires $\{Y_{ij}\}$.

Modèle statistique

Nous pouvons donc écrire le modèle : $$\mathcal{L}(Y_{ij}) = N(\mu_i ; \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

Ainsi nous constatons que, si les lois $L(Y_{ij})$ sont différentes, elles ne peuvent différer que par leur moyenne théorique. Il y a donc un simple décalage entre elles.

6. **Analyse de la variance à un facteur**

6.4. Conditions fondamentales de l'ANOVA

Remarque

Parfois, le modèle statistique est écrit de la façon suivante : $Y_{ii} = \mu + \alpha i + \varepsilon_{ij}$

Où $\sum_{i=1}^{I} \alpha_i = 0$ et $\mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0; \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$

Nous avons donc la correspondance suivante : $\mu_i = \mu + \alpha_i \qquad i = 1, \ldots, I.$

Les deux modèles sont donc statistiquement équivalents.

Mise en place du test de comparaison des moyennes

Nous nous proposons de tester l'hypothèse nulle

H0 : $\mu 1 = \mu 2 = \cdots = \mu I$

contre l'hypothèse alternative H1 : Les moyennes μi ne sont pas toutes égales.

La méthode statistique qui permet d'effectuer ce test est appelée l'**analyse de la variance à un facteur**.

6. **Analyse de la variance à un facteur**

6.5.Tableau de l’analyse de la variance

Deux propriétés fondamentales

Le test est fondé sur deux propriétés des moyennes et des variances.

Première propriété

La moyenne de toutes les observations est la moyenne des moyennes de chaque échantillon. Ceci s’écrit :

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} = \frac{1}{I}\sum_{i=1}^{I} \bar{y}_i$$

6. Analyse de la variance à un facteur

6.5.Tableau de l'analyse de la variance

Application sur l'exemple 1 «*Les laboratoires*»

Pour cet exemple, nous constatons cette propriété.
En effet, nous avons avec le logiciel **R** :

$\overline{y}$ = 1/30 × 2579 = 1/3 (85, 500 + 88, 600 + 83, 800) = 1/3 × 257, 900 = 85, 967,
Puisque n = 30 = I × J = 3 × 10.

Deuxième propriété
La variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances. Ceci s'écrit :

$$s^2(y) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2 = \frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2 + \frac{1}{I}\sum_{i=1}^{I} s_i^2(y). \quad (1)$$

6. **Analyse de la variance à un facteur**

6.5.Tableau de l'analyse de la variance

Exemple sur les laboratoires

Un calcul avec **R** donne : $s^2(y) = 9, 566$

D'autre part, nous constatons que la variance des moyennes est égale à :

$$\frac{1}{I}\sum_{i=1}^{I} s_i^2(y) = \frac{1}{3}(4,450 + 6,240 + 6,160) = 5,617$$

Nous constatons également que la moyenne des variances est égale à :

$$\begin{aligned}\frac{1}{I}\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2 &= \frac{1}{3}\Big((85,500 - 85,967)^2 + (88,600 - \\ &\quad 85,967)^2 + (83,800 - 85,967)^2\Big) \\ &= 3,949.\end{aligned}$$

En faisant la somme des deux derniers résultats, nous retrouvons bien la valeur de 9, 566 que nous avons obtenue par le calcul simple. Donc la relation (1) est bien vérifiée.

6. **Analyse de la variance à un facteur**

6.6. Résultat fondamental de l'ANOVA

En multipliant les deux membres par n de l'équation (1), nous obtenons :

ou encore ce qui s'écrit :

$sc_{tot} = sc_F + sc_{res}$ (2)

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\overline{y})^2 = J\sum_{i=1}^{I}(\overline{y}_i-\overline{y})^2 + \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij}-\overline{y}_i)^2\right)$$

Application sur l'exemple «*Les laboratoires*»

Avec le logiciel R, nous avons d'une part : $sc_{tot} = 286, 967$ et d'autre part :
$\mathbf{sc_F} = 118, 467$ et $\mathbf{sc_{res}} = 168, 500$.

Donc lorsque nous faisons la somme des deux derniers résultats nous retrouvons bien la valeur du premier résultat. Donc la relation (2) est bien vérifiée.

6. **Analyse de la variance à un facteur**

6.6. Résultat fondamental de l'ANOVA

Définition

Nous appelons **variation totale (total variation)** le terme : $$SC_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y})^2$$

Elle indique la dispersion des données autour de la moyenne générale.

Nous appelons **variation due au facteur (variation between)** le terme : $$SC_F = J\sum_{i=1}^{I}(\overline{y}_i - \overline{y})^2$$

Elle indique la dispersion des moyennes autour de la moyenne générale.

Nous appelons variation résiduelle (variation within) le terme : $$SC_{res} = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij} - \overline{y}_i)^2\right)$$

Elle indique la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

6. **Analyse de la variance à un facteur**

6.6. Résultat fondamental de l'ANOVA

Principe du test : Si l'hypothèse nulle H_0 est vraie alors la quantité SCF doit être petite par rapport à la quantité SCres.

Par contre, si l'hypothèse alternative H_1 est vraie alors la quantité SC_F doit être grande par rapport à la quantité SCres.

Pour comparer ces quantités, R.A. Fisher, après les avoir « corrigées » par leurs degrés de liberté (ddl), a considéré leur rapport.

Définition

Nous appelons **carré moyen associé au facteur** le terme: $CM_F = \frac{SC_F}{I-1}$

et **carré moyen résiduel** le terme: $CM_{res} = \frac{SC_{res}}{n-I}$

6. Analyse de la variance à un facteur

6.6. Résultat fondamental de l'ANOVA

Propriété

Le **carré moyen résiduel** est un estimateur sans biais de la variance des erreurs σ^2.

C'est pourquoi il est souvent également appelé **variance résiduelle** et presque systématiquement noté S^2_{res} lorsqu'il sert à estimer la variance des erreurs.

Sa valeur observée sur l'échantillon est ainsi notée cm_{res} ou s^2_{res}.

Si les **trois conditions** sont satisfaites et si l'hypothèse nulle H0 est vraie alors

$$F_{obs} = \frac{cm_F}{cm_{res}}$$

est une réalisation d'une variable aléatoire F qui suit une loi de Fisher à $I - 1$ degrés de liberté au numérateur et $n - I$ degrés de liberté au dénominateur.

Cette loi est notée $F_{I-1,n-I}$.

6. **Analyse de la variance à un facteur**

6.6. Résultat fondamental de l'ANOVA

Décision et conclusion du test

Pour un seuil donné α (=5%=0,05 en général), les tables de Fisher nous fournissent une valeur critique $c\alpha$ telle que $P_{H0}(F \leq c_\alpha) = 1 - \alpha$.

Si la valeur de la statistique calculée sur l'échantillon, notée F_{obs}, est supérieure ou égale à c_α, alors le test est significatif.

Vous rejetez H_0 et vous décidez que H_1 est vraie avec un risque d'erreur de première espèce **alpha** = 5%.

Si la valeur de la statistique calculée sur l'échantillon, notée F_{obs}, est strictement inférieure à c_α, alors le test n'est pas significatif.

Vous conservez H_0 avec un risque d'erreur de deuxième espèce β qu'il faut évaluer.

6. **Analyse de la variance à un facteur**

6.6. Résultat fondamental de l'ANOVA

Tableau de l'ANOVA

L'ensemble de la procédure est résumé par un tableau, appelé **tableau de l'analyse de la variance**, du type suivant :

Variation	SC	ddl	CM	F_{obs}	F_c
Due au facteur	sc_F	$l-1$	cm_F	$\frac{cm_F}{cm_{res}}$	c
Résiduelle	sc_{res}	$n-l$	cm_{res}		
Totale	sc_{tot}	$n-1$			

6. **Analyse de la variance à un facteur**

Application sur l'exemple «*Les laboratoires*»

Pour les données de l'exemple des laboratoires, le tableau de l'analyse de la variance s'écrit :

Variation	*SC*	*ddl*	*CM*	F_{obs}	F_c
Due au facteur	118,467	2	59,233	9,49	3,35
Résiduelle	168,500	27	6,241		
Totale	286,967	29			

6. **Analyse de la variance à un facteur**

Application sur l'exemple «*Les laboratoires*»

Décision et conclusion du test de Fisher pour les laboratoires

Pour un seuil $\alpha = 5\%$, les tables de Fisher nous fournissent la valeur critique $\mathbf{F_c = 3, 35}$.

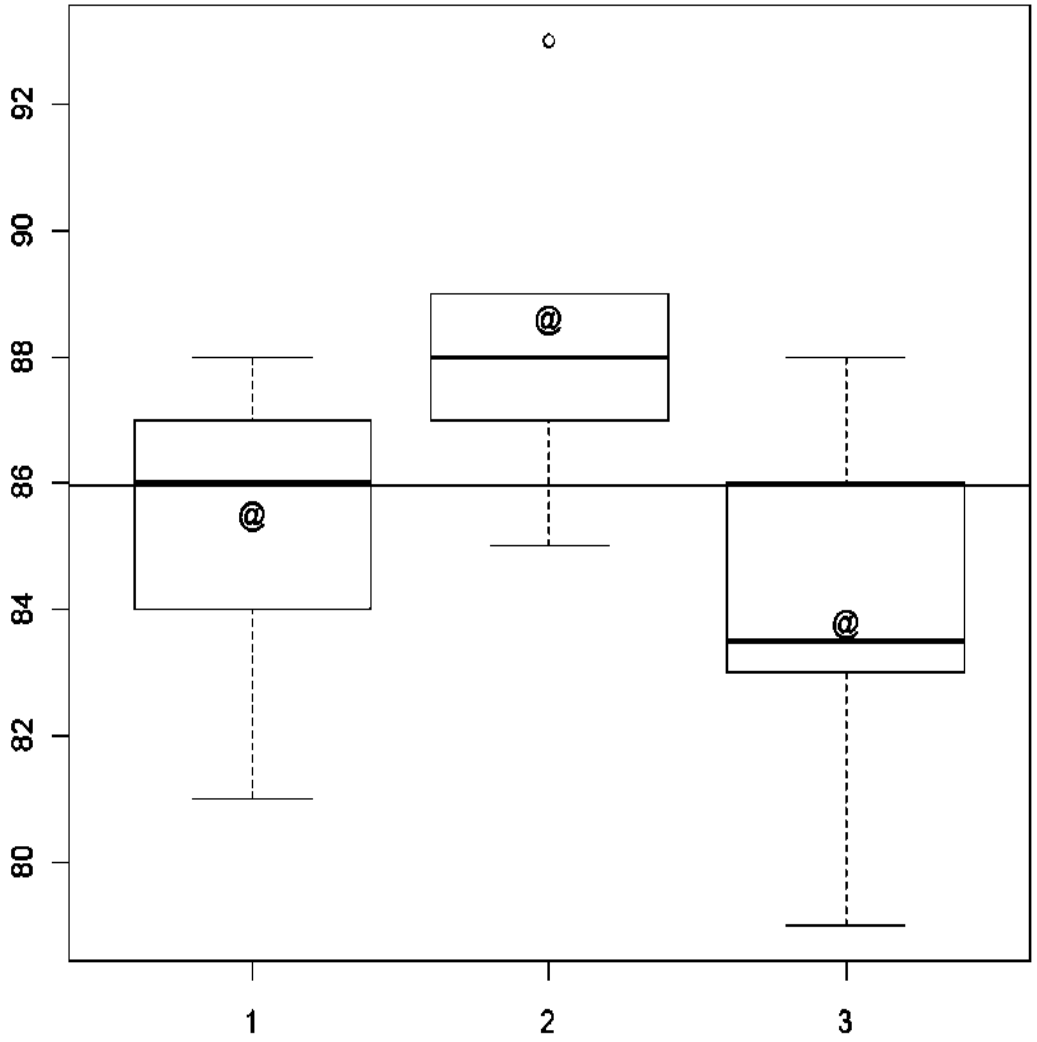
Le test est significatif puisque $\mathbf{9, 49 \geq 3, 35}$.
Nous décidons donc de rejeter l'hypothèse nulle H_0 et de décider que l'hypothèse alternative H_1 est vraie : il y a une différence entre les moyennes théoriques des quantités de lasalocide entre les laboratoires.

Le risque associé à cette décision est un risque de première espèce qui vaut $\alpha = 5\%$.

Nous en concluons que la quantité de lasalocide mesurée varie significativement d'un laboratoire à l'autre.

6. **Analyse de la variance à un facteur**

Application sur l'exemple «*Les laboratoires*»



Remarques

- Nous avons décidé que les moyennes théoriques sont différentes dans leur ensemble, mais nous aurions très bien pu trouver le contraire.

- Comme nous avons décidé que **les moyennes théoriques** sont **différentes** dans leur ensemble que le facteur étudié est à **effets fixes** et qu'il a **plus de trois modalités**, nous pourrions essayer de déterminer là où résident les différences avec un des tests de **comparaisons multiples**

6. **Analyse de la variance à un facteur**

6.7.Vérification des trois conditions

- **Indépendance**
- **Normalité**
- **Homogénéité**

Nous étudions les possibilités d'évaluer la validité **des trois conditions** que nous avons supposées satisfaites.

Condition d'indépendance

Il n'existe pas, dans un contexte général, **de test statistique simple permettant d'étudier l'indépendance**.

Ce sont les conditions de l'expérience qui nous permettront d'affirmer que nous sommes dans le cas de l'indépendance.

Condition de normalité

Nous ne pouvons pas, en général, la tester pour chaque échantillon. En effet le nombre d'observations est souvent très limité pour chaque échantillon.

6. **Analyse de la variance à un facteur**

6.7.Vérification des trois conditions

Remarque

Remarquons que si les conditions sont satisfaites et si nous notons : $\mathcal{E}_{ij} = Y_{ij} - \mu_i$

alors $\mathcal{L}(\mathcal{E}_{ij}) = \mathcal{N}(0\,;\,\sigma^2)$

alors c'est la même loi pour l'ensemble des unités. Les moyennes μi étant inconnues, nous les estimons par les estimateurs de la moyenne : les $\overline{Y}_i$ où ils sont définis par : $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$

Nous obtenons alors les estimations y_i. Les quantités obtenues s'appellent les **résidus** et sont notées $\hat{e}_{bij}$. Les résidus s'expriment par : $\hat{e}_{ij} = y_{ij} - \overline{y}_{i},\quad i = 1,\ldots,I,\quad j = 1,\ldots,J$

Les résidus peuvent s'interpréter comme des estimations des erreurs de mesure.

6. Analyse de la variance à un facteur

6.7.Vérification des trois conditions

A-Tests utilisés pour tester la normalité

Nous pouvons alors tester la normalité, avec le **test de Shapiro-Wilk** sur l'ensemble des résidus.

Hypothèses

Nous notons $\widehat{\mathcal{E}_{ij}}$ la variable aléatoire dont le résidu $\hat{e}_{ij}$ est la réalisation.

L'hypothèse nulle $\mathcal{H}_0 : \mathcal{L}(\widehat{\mathcal{E}_{ij}}) = \mathcal{N}$

contre l'hypothèse alternative $\mathcal{H}_1 : \mathcal{L}(\widehat{\mathcal{E}_{ij}}) \neq \mathcal{N}$

6. **Analyse de la variance à un facteur**

6.7.Vérification des trois conditions

Retour à l'exemple : le test de Shapiro-Wilk

Avec le logiciel **R**, nous avons

```
> shapiro.test(residuals(modele))
Shapiro-Wilk normality test
data: residuals(modele)
W = 0.9737, p-value = 0.6431
```

Comme la p−valeur (0,6431) est supérieure à 0,05, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle H_0.

Le risque d'erreur associé à cette décision est un risque de deuxième espèce β que nous ne pouvons pas évaluer.

Nous décidons que l'hypothèse de normalité est satisfaite.

6. Analyse de la variance à un facteur

6.7.Vérification des trois conditions

B- Condition d'homogénéité

Plusieurs tests permettent de tester l'égalité de plusieurs variances. Parmi ceux-ci, le test le plus utilisé est le **test de Bartlett** dont le protocole est le suivant :

Hypothèses

L'hypothèse nulle $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2$ contre l'hypothèse alternative $\mathcal{H}_1$: Les variances σ_i^2 ne sont pas toutes égales

Statistique

$$B_{obs} = \frac{1}{C_1}\left[(n-I)\ln(s_R^2) - \sum_{i=1}^{I}(n_i-1)\ln(s_{c,i}^2)\right] \qquad (3)$$ où

- la quantité C1 est définie par : $C_1 = 1 + \frac{1}{3(I-1)}\left(\left(\sum_{i=1}^{I}\frac{1}{n_i-1}\right) - \frac{1}{n-I}\right)$

- s^2_R la variance résiduelle, S^2_c la variance corrigée des observations de l'échantillon d'ordre i, (i = 1, . . . , I).

6. **Analyse de la variance à un facteur**

6.7.Vérification des trois conditions

Propriété

Sous l'hypothèse nulle H_0 le nombre B_{obs} est la réalisation d'une variable aléatoire B qui suit asymptotiquement une loi du khi-deux à $I - 1$ degrés de liberté.

En pratique, nous pouvons l'appliquer lorsque les effectifs n_i des I échantillons sont tous au moins égaux à 3.

Remarque

Ce test dépend de la normalité des résidus. Il se fait donc après avoir vérifié la normalité des résidus.

6. **Analyse de la variance à un facteur**

6.8. Décision et conclusion du test

Pour un seuil donné α (= 5% en général), les tables du khi-deux nous fournissent une valeur critique $c\alpha$ telle que $P_{H0}\ (B \leq c_\alpha) = 1 - \alpha$.

Si la valeur de la statistique calculée sur l'échantillon, notée B_{obs}, est supérieure ou égale à $c\alpha$, alors le test est significatif.

Vous rejetez H0 et vous décidez que H_1 est vraie avec un risque d'erreur de première espèce $\alpha = 5\%$.

Si la valeur de la statistique calculée sur l'échantillon, notée B_{obs}, est strictement inférieure à c_α, alors le test n'est pas significatif.

Vous conservez H_0 avec un risque d'erreur de deuxième espèce β que vous ne pouvez pas évaluer.

6. **Analyse de la variance à un facteur**

Retour à l'exemple

Nous obtenons avec le logiciel **R** et la commande bartlett.test :

```
> bartlett.test(residuals(modele)~laboratoire, data=analyse)

Bartlett test of homogeneity of variances
data: residuals(modele) by laboratoire
Bartlett's K-squared = 0.3024, df = 2, p-value = 0.8597
```

En se souvenant que les n_i sont tous égaux, nous lisons, avec le logiciel **R** :

$B_{obs} = 0, 3024.$

Pour un seuil $\alpha = 5\%$ la valeur critique d'un khi-deux à 2 degrés de liberté, est $c = 5, 991$. Comme $B_{obs} < c$, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle H_0.

Le risque d'erreur associé à cette décision est un risque de deuxième espèce β que nous ne pouvons pas évaluer.

Nous décidons que l'hypothèse d'homogénéité des variances est vérifiée.

6. **Analyse de la variance à un facteur**

6.9. **Test non paramétrique**

Si le test de Kruskal-Wallis indique une hétérogénéité au sein du groupe d'échantillons analysés, on se demande quels sont les échantillons qui différents les uns des autres ou quels groupes d'échantillons se révèlent significativement différents des autres.

Le test non paramétrique de comparaisons multiples nous permet de répondre à cette question.

La réalisation du test se fait de la même manière que le test SNK selon les étapes suivantes :

- Ordonner les sommes des rangs des différents échantillons par ordre croissant,

- Faire une série de comparaisons en commençant avec la plus grande différence entre les sommes des rangs prises deux à deux

6. **Analyse de la variance à un facteur**

6.9. **Test non paramétrique**

- Calculer : $q_{KW} = \frac{Y.max - Y.min}{SR}$

où ($Y_{.max} - Y_{.min}$) correspond à la différence entre les sommes des rangs et SR à l'erreur type donnée par la formule suivante $S_R = \sqrt{\frac{n(n\,p)\,(n\,p+1)}{12}}$

où n représente l'effectif de l'échantillon, qui doit être constant d'un échantillon à l'autre et p = 2 + (le nombre d'échantillons dont la valeur de Y. est comprise entre $Y_{.max} - Y_{.min}$ considérés). Au premier pas de la démarche (P = k), au deuxième pas à (k-1) au troisième pas à (k-2) et ainsi de suite.

- Comparer la valeur calculée qKW à la valeur critique qα fournie par la table des valeurs critiques de l'étendue de Student pour α choisi en fonction de la valeur de k et du ddl = (∞).
Si q_{kw} > qα H0 est rejetée et les deux sommes des rangs comparées sont significativement différentes au seuil considéré.

7. **Analyse de la variance à deux facteurs**

7.1. **Définitions**

L'**analyse de variance à deux** facteurs est une **généralisation de l'analyse de variance à un facteur**, qui permet de tenir compte **simultanément de deux facteurs**.

Les deux facteurs peuvent être placés soit sur un pied d'égalité, soit subordonnés l'un à l'autre. Dans le premier cas, les modèles d'analyse de variance sont dits croisés, et, dans le second cas, ils sont appelés hiérarchisés ou multi-niveaux.

Là encore, on distinguera entre **modèles fixes**, **modèles aléatoires** et **modèles mixtes**. Une distinction importante sera faite entre le cas des **effectifs égaux, souvent qualifié de plan équilibré ou orthogonal**, et le cas des **effectifs inégaux, souvent qualifié de plan non équilibré ou non orthogonal**.

Globalement, les **conditions d'application** de l'analyse de variance à deux facteurs sont de la même nature que pour un seul facteur : **populations normales, de même variance, et échantillons simples et indépendants**.

7. **Analyse de la variance à deux facteurs**

7.2. Contexte

Dans cette étude nous évaluons les effets simultanés d'un premier facteur à ***I*** modalités et d'un second facteur à ***J*** modalités sur une variable quantitative **Y.**

Supposons que **Y** suive des lois normales, a priori différentes dans les ***IJ*** populations disjointes.

Supposons que, dans la population correspondant à la modalité d'ordre i du premier facteur et à la modalité d'ordre j du deuxième facteur, nous ayons :

pour tout i = 1, . . . , I, j = 1, . . . , J et k = 1 . . . , K. $\mathcal{L}(Y_{ijk}) = \mathcal{N}(\mu_{ij}; \sigma^2)$

7. **Analyse de la variance à deux facteurs**

7.2. Contexte

Pour mettre en évidence les éventuelles différences entre le comportement de la variable **Y** dans les ***I*** modalités du premier facteur, dans les J modalités du deuxième facteur, ou encore dans l'interaction entre les deux facteurs, nous considérons des échantillons indépendants de même taille **K** de la variable Y dans chacune des ***IJ*** populations , soit au total un n-échantillon avec $n = I \times J \times K$.

Le modèle statistique

Pour la variable d'ordre k de la population d'indice (i, j), notée Y_{ijk} , nous posons :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

pour tout i = 1, . . . , I, j = 1, . . . , J et k = 1, . . . , K, avec, pour éviter une surparamétrisation, les contraintes

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij_0} = \sum_{j=1}^{J} (\alpha\beta)_{i_0 j} = 0$$ pour $i_0 = 1, \ldots, I$ et $j0 = 1, \ldots, J$.

7. **Analyse de la variance à deux facteurs**

7.3. Hypothèses du modèle

Les variables erreurs $\mathcal{E}_{ijk}$

sont supposées indépendantes et suivre une loi normale N (0; σ^2). Leurs réalisations, notées $\hat{e}_{ijk}$, sont considérées comme les erreurs de mesure, elles sont inconnues et vérifient : $y_{ijk} = \mu_{ij} + \hat{e}_{ijk}$, pour i = 1, . . . , I, j = 1, . . . , J, k = 1, . . . , K.

Les trois tests

L'analyse de la variance à deux facteurs avec répétitions permet trois tests de Fisher.

Le premier test

Nous testons : l'effet du premier facteur F_α : Nous testons l'égalité des **I** paramètres α_i correspondant aux I modalités du premier facteur

$$\begin{cases} (\mathcal{H}_0): & \text{les } I \text{ paramètres } \alpha_i \text{ sont tous nuls} \\ \text{contre} & \\ (\mathcal{H}_1): & \text{les } I \text{ paramètres } \alpha_i \text{ ne sont pas tous nuls} \end{cases}$$

7. **Analyse de la variance à deux facteurs**

7.3. Hypothèses du modèle

Le deuxième test

Nous testons : l'effet du deuxième facteur F_β. Il consiste à tester l'égalité des J paramètres β_j correspondant aux J modalités du deuxième facteur

$$\begin{cases} (\mathcal{H}_0): & \text{les } J \text{ paramètres } \beta_j \text{ sont tous nuls} \\ \text{contre} & \\ (\mathcal{H}_1): & \text{les } J \text{ paramètres } \beta_j \text{ ne sont pas tous nuls.} \end{cases}$$

Le troisième test

Nous testons : l'effet de l'interaction entre les deux facteurs F_α et F_β. Il consiste à comparer

$$\begin{cases} (\mathcal{H}_0): & \text{les } IJ \text{ paramètres } (\alpha\beta)_{ij} \text{ sont tous nuls} \\ \text{contre} & \\ (\mathcal{H}_1): & \text{les } IJ \text{ paramètres } (\alpha\beta)_{ij} \text{ ne sont pas tous nuls} \end{cases}$$

7. Analyse de la variance à deux facteurs

7.4. Notations

Nous posons

$$\overline{Y} = \frac{1}{n}\sum_{i,j,k} Y_{ijk}\, \overline{Y}_{ij\bullet} = \frac{1}{K}\sum_{k} Y_{ijk}, \quad \overline{Y}_{i\bullet\bullet} = \frac{1}{JK}\sum_{j,k} Y_{ijk}, \quad \overline{Y}_{\bullet j\bullet} = \frac{1}{IK}\sum_{i,k} Y_{ijk}$$

$$SC_T = \sum_{i,j,k}(Y_{ijk} - \overline{Y})^2, \quad SC_R = \sum_{i,j,k}(Y_{ijk} - \overline{Y}_{ij\bullet})^2,$$

$$SC_\alpha = \sum_{i,j,k}(\overline{Y}_{i\bullet\bullet} - \overline{Y})^2, \quad SC_\beta = \sum_{i,j,k}(\overline{Y}_{\bullet j\bullet} - \overline{Y})^2$$

$$SC_{\alpha\beta} = \sum_{i,j,k}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y})^2$$

7. **Analyse de la variance à deux facteurs**

7.5. Equation de l'ANOVA

L'équation de l'analyse de la variance devient pour ce modèle :

$\mathbf{SC_{Tot} = SC_R + SC_\alpha + SC_\beta + SC_{\alpha\beta}}$. où

- la somme SC_{Tot} , **la somme totale**, mesure la somme des carrés des écarts à la moyenne globale, toutes causes confondues,
- la somme SCR, **la somme résiduelle**, cumule les carrés des écarts des différentes observations à la moyenne de l'échantillon dont elles font partie. Dans la somme totale elle représente la part de la dispersion due aux **fluctuations individuelles**.
- la somme SC_α, **la somme due au premier facteur**, ou **somme entre modalités du facteur F_α**, mesure l'effet du premier facteur.
- la somme SC_β, ou **somme due au deuxième facteur**, ou **somme entre modalités du facteur F_β**, mesure l'effet du deuxième facteur.
- la somme $SC_{\alpha\beta}$ mesure l'effet de **l'interaction entre les deux facteurs**.

7. **Analyse de la variance à deux facteurs**

7.6. Propriété

Sous les différentes hypothèses nulles (H_0) d'égalité des paramètres de la décomposition des μ_{ij}, nous pouvons préciser les lois respectives des variables précédentes. Elles suivent des lois du χ^2:

$$\mathcal{L}_{\mathcal{H}_0}\left(\frac{1}{\sigma^2}SC_{Tot}\right) = \chi^2_{n-1}, \quad \mathcal{L}_{\mathcal{H}_0}\left(\frac{1}{\sigma^2}SC_R\right) = \chi^2_{n-IJ}$$

$$\mathcal{L}_{\mathcal{H}_0}\left(\frac{1}{\sigma^2}SC_\alpha\right) = \chi^2_{I-1}, \quad \mathcal{L}_{\mathcal{H}_0}\left(\frac{1}{\sigma^2}SC_\beta\right) = \chi^2_{J-1}$$

$$\mathcal{L}_{\mathcal{H}_0}\left(\frac{1}{\sigma^2}SC_{\alpha\beta}\right) = \chi^2_{(I-1)(J-1)}$$

7. **Analyse de la variance à deux facteurs**

7.6. Propriété

De plus, les variables SC_R et SC_α, SC_R et SC_β, SC_R et $SC_{\alpha\beta}$ sont indépendantes, de sorte que :

$$\mathcal{L}_{\mathcal{H}_0}\left(\frac{\frac{SC_\alpha}{I-1}}{\frac{SC_R}{IJ(K-1)}}\right) = \mathcal{F}_{(I-1),IJ(K-1)} \qquad \mathcal{L}_{\mathcal{H}_0}\left(\frac{\frac{SC_\beta}{J-1}}{\frac{SC_R}{IJ(K-1)}}\right) = \mathcal{F}_{(J-1),IJ(K-1)} \qquad \mathcal{L}_{\mathcal{H}_0}\left(\frac{\frac{SC_{\alpha\beta}}{(I-1)(J-1)}}{\frac{SC_R}{IJ(K-1)}}\right) = \mathcal{F}_{(I-1)(J-1),IJ(K-1)}$$

Les tests

Les tests sont réalisés à l'aide des valeurs numériques suivantes :

$$\bar{y} = \frac{1}{IJK}\sum_{i,j,k} y_{ijk}, \quad \bar{y}_{ij\bullet} = \frac{1}{K}\sum_{k} y_{ijk} \qquad \bar{y}_{i\bullet\bullet} = \frac{1}{JK}\sum_{j,k} y_{ijk}, \quad \bar{y}_{\bullet j\bullet} = \frac{1}{IK}\sum_{i,k} y_{ijk}$$

7. Analyse de la variance à deux facteurs

7.7. Décision

Pour un seuil α(= 5% = 0, 05 en général), les tables de la loi de Fisher notée F nous fournissent pour chacun des trois tests une valeur critique c telle que P_{H0} (F < c) = 1 − α. Alors nous décidons :

$$\begin{cases} \text{si } F_{obs} < c & (\mathcal{H}_0) \text{ est vraie ,} \\ \text{si } c \leqslant F_{obs} & (\mathcal{H}_1) \text{ est vraie.} \end{cases}$$

Les résultats des calculs sont généralement présentés sous la forme d'un tableau.

Tableau de l'ANOVA

Variation	SC	ddl	s^2	F_{obs}	F_c
Due au facteur α	SC_α	$I-1$	s^2_α	$\frac{s^2_\alpha}{s^2_R}$	c
Due au facteur β	SC_β	$J-1$	s^2_β	$\frac{s^2_\beta}{s^2_R}$	c
Interaction	$SC_{\alpha\beta}$	$(I-1)(J-1)$	$s^2_{\alpha\beta}$	$\frac{s^2_{\alpha\beta}}{s^2_R}$	c
Résiduelle	SC_R	$IJ(K-1)$	s^2_R		
Totale	SC_{Tot}	$n-1$			

7. Analyse de la variance à deux facteurs

$$SC_{Tot} = \sum_{i,j,k}(y_{ijk} - \overline{y})^2 = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2\right) - IJK\overline{y}^2,$$

$$SC_R = \sum_{i,j,k}(y_{ijk} - \overline{y}_{ij\bullet})^2 = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2\right) - K\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{y}_{ij\bullet}^2,$$

$$SC_\alpha = \sum_{i,j,k}(\overline{y}_{i\bullet\bullet} - \overline{y})^2 = JK\sum_{i=1}^{I}\overline{y}_{i\bullet\bullet}^2 - IJK\overline{y}^2,$$

$$SC_\beta = \sum_{i,j,k}(\overline{y}_{\bullet j\bullet} - \overline{y})^2 = IK\sum_{j=1}^{J}\overline{y}_{\bullet j\bullet}^2 - IJK\overline{y}^2,$$

$$SC_{\alpha\beta} = \sum_{i,j,k}(\overline{y}_{ij\bullet} - \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet j\bullet} + \overline{y})^2$$

$$= K\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{y}_{ij\bullet}^2 - JK\sum_{i=1}^{I}\overline{y}_{i\bullet\bullet}^2 - IK\sum_{j=1}^{J}\overline{y}_{\bullet j\bullet}^2 + IJK\overline{y}^2$$

7. Analyse de la variance à deux facteurs

7.8. Conclusion

Supposons provisoirement que les conditions du modèle d'anova à 2 facteurs avec répétitions soient vérifiées, nous pouvons

1. décider d'accepter H_1. Il y a effet du premier facteur.
2. décider d'accepter H_1. Il y a effet du deuxième facteur.
3. décider d'accepter H_0. Il n'y a pas d'effet de l'interaction entre les deux facteurs.

7. Analyse de la variance à deux facteurs

7.9. Vérification des conditions

Pour ce modèle, l'estimation des moyennes théoriques μij se fait par les moyennes observées

$\overline{y}_{ij\bullet}$ (« valeurs ajustées »). Les résidus sont alors donnés par l'expression :

$$\hat{e}_{ijk} = y_{ijk} - \overline{y}_{ij\bullet}, \quad i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K.$$

Leur normalité et l'homogénéité des variances se vérifient par les mêmes méthodes que pour une analyse de la variance à un facteur.

7. Analyse de la variance à deux facteurs

7.10. ANOVA à deux facteurs sans répétition

- L'idée générale

Dans le cas où nous étudions l'effet simultané de deux facteurs à, respectivement, I et J modalités et que nous disposons d'une seule observation pour chaque population, c'est à dire K = 1, les résultats du paragraphe précédent ne sont plus valables. Nous devons supposer que l'interaction entre les deux facteurs est nulle. Partant du même modèle, nous écrivons plus simplement : $Y_{ij} = \mu + \alpha_i + \beta_j + \mathcal{E}_{ij}$

avec les contraintes $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$

- Notations

Nous avons les notations analogues :

$$\overline{y} = \frac{1}{IJ}\sum_{i,j} y_{ij}, \quad \overline{y}_{i\bullet} = \frac{1}{J}\sum_{j} y_{ij}, \quad \overline{y}_{\bullet j} = \frac{1}{I}\sum_{i} y_{ij}$$

$$SC_{Tot} = \sum_{i,j}(y_{ij} - \overline{y})^2 = \sum_{i,j} y_{ij}^2 - IJ\overline{y}^2,$$

$$SC_R = \sum_{i,j}(y_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y})^2,$$

7. **Analyse de la variance à deux facteurs**

- Notations

$$SC_\alpha = \sum_{i,j}(\overline{y}_{i\bullet} - \overline{y})^2 = J\sum_i \overline{y}_{i\bullet}^2 - IJ\overline{y}^2,$$

$$SC_\beta = \sum_{i,j}(\overline{y}_{\bullet j} - \overline{y})^2 = I\sum_j \overline{y}_{\bullet j}^2 - IJ\overline{y}^2.$$

- Remarque importante

Remarquons que l'expression définissant, dans le cas avec répétitions, la somme des carrés associée à l'interaction, est associée ici à la somme des carrés de la résiduelle.

- Tableau de l'ANOVA

Nous avons alors le tableau de l'analyse de la variance suivant :

Variation	SC	ddl	F_{obs}	F_c
Due au facteur α	SC_α	$I-1$	$\frac{s_\alpha^2}{s_R^2}$	c
Due au facteur β	SC_β	$J-1$	$\frac{s_\beta^2}{s_R^2}$	c
Résiduelle	SC_R	$(I-1)(J-1)$		
Totale	SC_T	$IJ-1$		

7. **Analyse de la variance à deux facteurs**

- Idée générale

La démarche est alors analogue à celle de l'analyse de la variance à deux facteurs avec répétitions. Notons que dans ce cas les valeurs ajustées sont données par : $\widehat{\mu_{ij}} = \bar{y}_{i\bullet} + \bar{y}_{\bullet j} - \bar{y}$

et les résidus par l'expression : $\hat{e}_{ij} = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}, \quad i = 1,\ldots,I; j = 1,\ldots,J$

7. **Analyse de la variance à deux facteurs**

Exercice

Considérez la base de données nommée " anovaplus « ,Les données sont relatives aux paramètres nutritionnels du légume *Solanum microcarpum*, suivant trois pratiques culturales (P1,P2,P3) et deux tours de coupe (C1 et C2).
L'objectif est de tester si la pratique culturale et/ou le nombre de coupe influencent significativement les paramètres nutritionnels. Prenons comme exemple le paramètre "N« ,
Réalisez le test statistique approprié après avoir vérifié ses conditions d’application

7. **Analyse de la variance à deux facteurs**

Les limites de l'analyse de la variance

Bien que l'ANOVA aide à analyser la différence de moyenne entre deux variables indépendantes, elle ne dira pas quels groupes statistiques sont différents les uns des autres.

Si le test renvoie une statistique significative (valeur obtenue lors de l'exécution du test), il faudra effectuer un test complémentaire pour dire exactement quels groupes ont une différence.

8. **Comparaison multiple des moyennes**

8.1. But

L'analyse de la variance (ANOVA) constitue la première étape d'une comparaison des moyennes de plusieurs échantillons indépendants.

Dans le cas du rejet de l'hypothèse d'homogénéité (Hypothèse nulle), une question supplémentaire se pose.

En effet, il est intéressant de savoir quelles sont les moyennes qui différent significativement entre elles.

Autrement, il faut poursuivre l'analyse par un test de comparaison multiple des moyennes pour rechercher les groupes homogènes éventuellement.

Plusieurs tests nous permettent de répondre à cette question : **test LSD** (Least Significant Difference), **test de Duncan** (Duncan's multiple range test), **test SNK** (méthode de Student-Newman-Keuls), **test HSD de Tukey** (Honestly Significant Difference) etc.

8. Comparaisons multiples des moyennes

8.2. Statistique de rang studentisée

Adaptation du test t pour comparer deux moyennes a posteriori (n_i supposés égaux).

- Ordonner les laboratoires en fonction des moyennes observées : $\bar{y}_{(1)} \leqslant \bar{y}_{(2)} \leqslant \bar{y}_{(3)} \leqslant \cdots \leqslant \bar{y}_{(I)}$
- Puis appliquer la procédure du test qui va suivre.

Test basé sur la statistique de rang studentisée

Objectif : comparer le laboratoire i au laboratoire j, où i < j

- Hypothèses : $(H_0) : \mu_i = \mu j$ contre $(H_1) : \mu_i < \mu_j$
- Calcul de la statistique du test : $q_{r,obs} = \dfrac{\bar{y}_{(j)} - \bar{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{J}}}$ avec $r = j - i + 1$
- Règle de décision : si $q_{r,obs} < q_{r,n-I}$, alors le test n'est pas significatif. Si $q_{r,obs} \geq q_{r,n-I}$, alors le test est significatif.

8. Comparaisons multiples des moyennes

8.2. Statistique de rang studentisée

Remarque

Le seuil critique dépend du nombre de traitements entre i et j et du type d'approche.

Notion de « plus petite différence significative »

- Si nous désirons comparer par une statistique de rang studentisée deux moyennes µi et µj , nous calculerons la quantité suivante : $q_{r,obs} = \dfrac{\bar{y}_{(j)} - \bar{y}_{(i)}}{\sqrt{\dfrac{s_R^2}{n}}}$ avec r = j – i + 1

Quelle est la plus petite valeur de $\bar{y}_{(j)} - \bar{y}_{(i)}$ à partir de laquelle le test sera rejeté ?

Réponse : la plus petite valeur de la différence entre les moyennes, à partir de laquelle le test sera rejeté, est égale à :

$$\bar{y}_{(j)} - \bar{y}_{(i)} \geqslant \sqrt{\frac{s_R^2}{n}} \times q_{r,n-l} = W_r$$

8. Comparaisons multiples des moyennes

8.3. Test de Newman Keuls

Objectif : classer les traitements par groupes qui sont significativement différents.

La méthode est la suivante :

Étape 1 : ordonner les moyennes et calculer toutes les différences deux à deux entre moyennes.
Étape 2 : calculer pour r = 2 à *I* les différences minimum significatives Wr.
Étape 3 : dans le tableau des différences, rechercher toutes les différences significatives en fonction de leur « distance » r.
Étape 4 : classer les traitements par groupes significativement différents.

8. Comparaisons multiples des moyennes

8.4. Test de Tukey

But : comme pour le test de Newman Keuls, classer les traitements par groupes qui sont significativement différents.

Méthode : elle est identique à celle du test de Newman-Keuls mais nous prendrons comme différence minimum significative W_k pour toutes les différences. W_k est ici alors noté « HSD » (Honestly Significant Difference)

Comparaison des deux méthodes : la méthode de Tukey trouvera moins de différences significatives que la méthode de Newman Keuls (erreur de type I globale plus faible mais moins de puissance que la méthode de Newman Keuls).

Contexte du test de Tukey

Les moyennes observées $\bar{y}_i$ sont rangées par ordre croissant. Nous rappelons que nous les notons par : $\bar{y}_{(1)}, \bar{y}_{(2)}, \cdots, \bar{y}_{(I)}$

et les moyennes théoriques associées par : μ(1) , μ(2) , . . . , μ(I) .

8. Comparaisons multiples des moyennes

8.4. Test de Tukey

La procédure du test de Tukey est la suivante :

Pour chaque i < i', nous considérons l'**hypothèse nulle** $H_0 : \mu(i) = \mu(i')$

contre l'**hypothèse alternative** $H1 : \mu(i\ 0) > \mu(i)$.

Statistique

Nous considérons le rapport :

$$t_{i',i,obs} = \frac{\overline{y}_{(i')} - \overline{y}_{(i)}}{\sqrt{\frac{s_R^2}{2}\left(\frac{1}{n_{i'}} + \frac{1}{n_i}\right)}} \quad (4)$$

Propriété

Le rapport $t_{i',i,obs}$ défini par (4) est la réalisation d'une variable aléatoire T qui, si l'hypothèse nulle H_0 est vraie, suit une loi appelée **étendue studentisée (studentized range)** et que nous notons: $\widetilde{T}_{n-l,l}$

8. Comparaisons multiples des moyennes

8.4.Test de Tukey

Décision et conclusion du test

Pour un seuil donné α (= 5% en général), les tables de l'étendue studentisée nous fournissent une valeur critique c telle que PH0 (T ≤ c) = 1 − α. **Si la valeur de la statistique calculée sur l'échantillon, notée 6 $t_{i',i,obs}$, est supérieure ou égale à c, alors le test est significatif**. **Vous rejetez H_0 et vous décidez que H1 est vraie avec un risque d'erreur de première espèce α** = 5%. Si la valeur de la statistique calculée sur l'échantillon, notée 6 $t_{i',i,obs}$, est inférieure à c, alors le test n'est pas significatif. Vous conservez H_0 avec un risque d'erreur de deuxième espèce *β* qu'il faut évaluer.

Remarque

La valeur critique c ne dépend que des indices n − I, degrés de liberté de la somme des carrés résiduelle, et de I, nombre des moyennes comparées. De plus, les moyennes théoriques, dont les moyennes observées sont comprises entre deux moyennes observées, dont les moyennes théoriques correspondantes sont déclarées égales, sont déclarées égales avec ces dernières.

8. Comparaisons multiples des moyennes

8.4. Test de Tukey

Exemple sur les laboratoires

```
> modele1<-aov(lasalocide laboratoires, +data=analyse)
> TukeyHSD(modele1)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = lasalocide ~ laboratoires, data = analyse)
$laboratoires
diff lwr upr p adj
2-1 3.1 0.3299809 5.870019 0.0259501
3-1 -1.7 -4.4700191 1.070019 0.2969093
3-2 -4.8 -7.5700191 -2.029981 0.0005724
```

8. Comparaisons multiples des moyennes

8.4. Test de Tukey

Groupement avec la méthode de Tukey

Laboratoire	Taille	Moyenne	Groupement
A	10	85,5	*B*
B	10	88,6	*A*
C	10	83,8	*B*