Inférence statistique

Seconde partie

Frédéric SUR
frederic.sur@univ-lorraine.fr
https://members.loria.fr/FSur/

Tronc commun scientifique FICM 1A
École des Mines de Nancy

2026-2027

Table des matières

À l'attention des étudiants FICM 1A **5**

1 Tests statistiques d'hypothèse (1) **7**
1.1 Tests statistiques : généralités 7
1.2 Quelques tests paramétriques basiques 16
1.3 Exercices 19

2 Tests statistiques d'hypothèse (2) **23**
2.1 Tests de comparaison de deux échantillons 23
2.2 Tests d'ajustement à une loi 27
2.3 Tests d'indépendance 29
2.4 Limite des tests d'hypothèse à l'heure des données massives 32
2.5 Exercices 33

3 Régression linéaire **37**
3.1 Introduction 37
3.2 La régression linéaire simple 38
3.3 Inférence statistique sous hypothèse de normalité 44
3.4 Discussion d'un exemple 48
3.5 Qualité de l'ajustement et diagnostics avancés 49
3.6 Exercices 50

4 Analyse de la variance **53**
4.1 Introduction 53
4.2 ANOVA à un facteur 54
4.3 ANOVA à deux facteurs 58
4.4 Exercices 65

A Tables statistiques **69**
A.1 Loi normale centrée réduite 69
A.2 Loi de Student 71
A.3 Loi du χ^2 72
A.4 Loi de Fisher-Snedecor 74

À l'attention des étudiants FICM 1A
(à lire attentivement avant la première séance)

Ce document constitue le support de la seconde partie du cours d'Inférence statistique, à la suite de la première partie animée par M. Frénais.

Les notations peuvent légèrement varier entre ces deux parties.

Il est attendu de votre part un travail de lecture active de chaque chapitre *avant* la séance correspondante. Il faut compter approximativement une heure. L'objectif de la séance en amphithéâtre est d'éclairer les concepts sous un angle différent, et pas de revenir sur les définitions et théorèmes. Il est indispensable d'avoir travaillé le chapitre avant la séance de travaux dirigés. Même si la lecture peut sembler ardue, tout cela s'éclairera avec la pratique. Les exercices qui seront traités en travaux dirigés figurent dans la dernière section de chaque chapitre : il n'est pas attendu que vous les travailliez en autonomie avant la séance.

Les remarques et notes de bas de page sont des compléments d'information facultatifs.

Tout le matériel du cours sera disponible sur la page Arche.

Avant la première séance de cette seconde partie, il faudra donc lire le chapitre 1 des présentes notes.

Frédéric Sur, le 4 novembre 2026.

Chapitre 1

Tests statistiques d'hypothèse (1)

Dans cette seconde partie du cours de statistique, nous introduisons une procédure de décision qui permet d'accepter ou de rejeter une affirmation sur une population à partir de l'analyse de données. Il s'agit des **tests statistiques d'hypothèse**.

1.1 Tests statistiques : généralités

Imaginons une situation dans laquelle on dispose de données venant d'un phénomène dont les variations sont bien modélisées par les réalisations indépendantes d'une variable aléatoire. À partir de ces données (on parle aussi d'observations), on peut être amené à faire une ou des hypothèses, que l'on aimerait pouvoir rejeter ou accepter. Il semble clair que la décision va être prise avec un certain niveau de confiance, car elle se base sur des données aléatoires.

Exemple 1

On jette une pièce vingt fois. On observe douze fois « pile » et huit fois « face ». Peut-on dire que la pièce est équilibrée ?

Exemple 2

Une fuite dans un circuit de refroidissement fait perdre de l'eau à chaque heure d'utilisation. En moyenne, 10 litres par heure sont perdus, avec un écart-type de 1 litre (statistique établie sur un temps long pendant lequel la fuite est stable). Une réparation est censée ramener la moyenne des pertes à 9 litres sans changer l'écart-type. Après réparation, on mesure pendant 5 heures des fuites de 7,1 litres, 9,5 litres, 8,8 litres, 10,1 litres, 9,2 litres par heure. La réparation a-t-elle fait passer les pertes moyennes par heure à 9 litres comme le prétend le réparateur, ou sont-elles restées à 10 litres ?

La théorie des **tests statistiques d'hypothèse** est le cadre qui fournit une méthodologie permettant de répondre à ces questions. Cette théorie a été développée au début du XXème

siècle par des statisticiens comme Karl Pearson (dont vous avez peut-être entendu parler du coefficient de corrélation), William Sealy Gosset (dont le pseudonyme est Student), Ronald Fisher, Jerzy Neyman, et Egon Pearson (qui ont donné leurs noms à différents concepts sur lesquels nous reviendrons). Comme souvent, rien n'est plus pratique qu'une bonne théorie. Comme il s'agit de prendre des décisions dont l'impact est potentiellement important, la pratique doit être menée de manière très rigoureuse.

1.1.1 Hypothèse nulle et hypothèse alternative

La démarche générale est toujours la même. Les données du problème étant n réalisations indépendantes d'une variable aléatoire X, on considère un n-échantillon $(X_1, X_2, \ldots, X_n)$, chaque variable aléatoire X_i étant distribuée comme X. Les données, ou observations, sont une réalisation du n-échantillon. Considérer un n-échantillon permet de tenir compte des variations possibles inhérentes au problème. Ensuite, on définit une statistique S sur cet échantillon (c'est-à-dire une fonction de $X_1, X_2, \ldots, X_n$). La statistique S est elle-même une variable aléatoire, et elle prend dans ce cours des valeurs réelles : à partir de la réalisation observée de l'échantillon, on calcule une réalisation s de S. Cette statistique est appelée **statistique de test**. Elle doit capturer l'information pertinente pour le problème à partir de l'échantillon, synthétisée dans une valeur réelle.

Une hypothèse $\mathscr{H}_0$ (dite **hypothèse nulle**) est formulée sur le problème ayant fourni les données, ainsi qu'une **hypothèse alternative** $\mathscr{H}_1$. Ces hypothèses portent sur une propriété caractérisant la loi de X, donc celle de S. On va voir qu'elles n'ont pas un rôle symétrique : l'hypothèse nulle est celle qui nous intéresse le plus, la plus critique pour le problème considéré. Réfuter l'hypothèse nulle alors qu'elle est vraie aurait des conséquences importantes.

Exemple 1 (suite)

Soit p la probabilité d'obtenir « pile » sur un lancer de pièce. La variable aléatoire X modélisant bien un lancer est celle qui vaut 1 avec probabilité p en cas de « pile », 0 avec probabilité $1-p$ en cas de « face ». Les hypothèses sont alors : $\mathscr{H}_0 : p = 1/2$ et $\mathscr{H}_1 : p \neq 1/2$. L'hypothèse nulle est bien la plus importante (elle suppose la pièce bien équilibrée, qui est la caractéristique qu'on cherche logiquement à établir). La variable aléatoire X suit une loi de Bernoulli de paramètre p. L'information pertinente issue du n-échantillon pourrait être le nombre de « pile » sur n lancers. On choisit donc la statistique $S = X_1 + \cdots + X_n$. Elle suit une loi binomiale de paramètres n (le nombre de tirages) et p. Sous l'hypothèse $\mathscr{H}_0$, S suit une loi binomiale de paramètres n et $1/2$.

Exemple 2 (suite)

La variable aléatoire X est ici celle représentant le nombre de litres d'eau perdus chaque heure. Sa loi pourra être modélisée comme une loi gaussienne de moyenne m et d'écart-type $\sigma = 1$ connu. Les hypothèses à départager sont : $\mathscr{H}_0 : m = 10$ et $\mathscr{H}_1 : m = 9$. L'hypothèse nulle correspond à l'hypothèse critique : on veut être certain de ne pas réfuter $\mathscr{H}_0$ (et donc juger la réparation convenable) à tort. Comme on s'intéresse à des pertes d'eau moyennes, il semble logique de choisir comme statistique de test la moyenne empirique : $S = \sum_{i=1}^{n} X_i / n$, où n est

le nombre de mesures disponibles (ici, $n = 5$). Sous $\mathscr{H}_0$, S suit une loi normale de moyenne 10 et de variance 1/5, et sous $\mathscr{H}_1$, de moyenne 9 et de même variance 1/5.

Dans ces exemples, les hypothèses portent sur un paramètre de la loi de X : p dans l'exemple 1, et m dans l'exemple 2. On parle de **test d'hypothèse paramétrique**, ils nous occuperont une grande partie de ce cours. Une hypothèse qui impose une valeur unique au paramètre est qualifiée d'**hypothèse simple**. Une hypothèse qui impose au paramètre un ensemble de valeurs possibles (par exemple un intervalle de valeurs) est qualifiée d'**hypothèse composite**.

Dans l'exemple 2, $\mathscr{H}_1$ est une hypothèse simple, et dans l'exemple 1, $\mathscr{H}_1$ est une hypothèse composite.

Les deux hypothèses ne sont pas compatibles : si on accepte $\mathscr{H}_0$, on refuse $\mathscr{H}_1$, et réciproquement.

1.1.2 Région d'acceptation, région de rejet

La décision de rejeter $\mathscr{H}_0$ (et donc d'accepter $\mathscr{H}_1$) va être prise si le n-uplet des données observées (la réalisation $(x_1,\dots,x_n)$ du n-échantillon) est dans une région W (comme *wrong*) de $\mathbb{R}^n$ appelée **région critique** (ou **région de rejet de** $\mathscr{H}_0$). Réciproquement, si le n-uplet des observations appartient au complémentaire $\overline{W}$ de W, appelé **région d'acceptation de** $\mathscr{H}_0$, on accepte $\mathscr{H}_0$ (et on rejette $\mathscr{H}_1$).

Construire un test d'hypothèse, c'est construire l'ensemble W, qui permet de mettre en œuvre la règle de décision : si $(x_1,\dots,x_n) \in W$, alors on rejette $\mathscr{H}_0$ (et on accepte $\mathscr{H}_1$), et inversement, si $(x_1,\dots,x_n) \in \overline{W}$, alors on rejette $\mathscr{H}_1$ (et on accepte $\mathscr{H}_0$).

Comme il n'est pas pratique de définir W sur les réalisations du n-échantillon (qui sont dans $\mathbb{R}^n$), on va plutôt prendre une décision selon la réalisation s de la statistique de test S, à valeurs réelles. On définira W à partir de la valeur s calculée à partir de $(x_1,\dots,x_n)$ comme l'ensemble :

$$W = \left\{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } s \in I\right\}$$

où I sera généralement un intervalle (borné ou non) ou une réunion d'intervalles de $\mathbb{R}$.

Remarque. Par léger abus de langage, c'est souvent l'intervalle I qui est appelé région de rejet (et est noté W).

1.1.3 Erreurs de première et deuxième espèce

Toute décision peut amener à faire des erreurs. Deux types d'erreurs sont possibles :
— on choisit $\mathscr{H}_1$ alors que $\mathscr{H}_0$ est vraie ;
— on choisit $\mathscr{H}_0$ alors que $\mathscr{H}_1$ est vraie.

Ces erreurs sont quantifiées par :
— la probabilité de choisir $\mathscr{H}_1$ alors que $\mathscr{H}_0$ est vraie : il s'agit de l'**erreur de première espèce** (notée α) ;

— la probabilité de choisir $\mathscr{H}_0$ alors que $\mathscr{H}_1$ est vraie : il s'agit de l'**erreur de seconde espèce** (notée β).

La quantité $1-\beta$ s'appelle la **puissance** du test : c'est la probabilité de choisir $\mathscr{H}_1$ (ou de rejeter $\mathscr{H}_0$) en ayant raison.

Bien entendu, on aimerait que les deux types d'erreurs soient les plus petites possible. Néanmoins, si on diminue le risque de première espèce (on accepte $\mathscr{H}_1$ moins facilement), c'est que la taille de la région critique W diminue, et donc que la région d'acceptation $\overline{W}$ de $\mathscr{H}_0$ augmente, ce qui va augmenter la valeur du risque de seconde espèce car on accepte $\mathscr{H}_0$ plus facilement. Les deux types d'erreurs sont donc antagonistes.

On a dit plus haut que les deux hypothèses ne jouaient pas le même rôle. L'hypothèse $\mathscr{H}_0$ est choisie de manière à ce que l'erreur de première espèce soit la plus importante des deux erreurs à contrôler.

Un test tel que sa puissance est supérieure à son risque de première espèce est appelé **test sans biais**. Autrement dit, pour un test sans biais, la probabilité de rejeter $\mathscr{H}_0$ alors que $\mathscr{H}_1$ est vraie est supérieure à celle de rejeter $\mathscr{H}_0$ alors que $\mathscr{H}_0$ elle-même est vraie. Dans le cas contraire, le test est dit biaisé. Attention, il ne s'agit pas de la même notion de biais que pour un estimateur statistique.

On va chercher W tel que l'erreur de première espèce α est égale (ou inférieure) à une certaine valeur α^* appelée **niveau de signification du test** ou **seuil de signification** (on parle aussi de **significativité**. Une valeur typique est $\alpha^* = 0,05$, d'autres valeurs standards étant $0,01$ ou $0,1$. La quantité $1-\alpha^*$ (typiquement 95%) est le **niveau de confiance** du test.

Fixer l'erreur de première espèce ne suffit pas à caractériser W. On peut s'en convaincre en considérant le cas où $\mathscr{H}_0$ est une hypothèse simple permettant de caractériser la loi $p_{\mathscr{H}_0}(\boldsymbol{x})$ du n-échantillon. En effet, l'erreur de première espèce est l'espérance du nombre d'erreurs sous l'hypothèse $\mathscr{H}_0$ et s'écrit comme :

$$\alpha = \int_{\mathbb{R}^n} \mathbf{1}_W(\boldsymbol{x})\, p_{\mathscr{H}_0}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_W p_{\mathscr{H}_0}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

où $\mathbf{1}_W$ désigne la fonction indicatrice de l'ensemble W, qui vaut 1 si $\boldsymbol{x} \in W$ ($\mathscr{H}_0$ a été rejetée à tort) et 0 si $\boldsymbol{x} \notin W$. Il n'y a pas de raison qu'il y ait un unique domaine W tel que cette intégrale soit égale à α^*.

Il faut donc ajouter une contrainte pour définir W, et donc le test statistique.

La première possibilité, qui est celle qu'on va adopter dans ce cours et qui sera développée dans la suite, est de définir l'allure de la région critique de manière intuitive en fonction de $\mathscr{H}_1$.

La seconde est d'ajouter une contrainte supplémentaire sur l'erreur de seconde espèce : on cherche la région critique W de manière à ce que l'erreur de seconde espèce β du test soit minimale, ou, de manière équivalente, de manière à ce que la puissance $1-\beta$ du test soit maximale. C'est le principe de Neyman. Au passage, comme W est défini à partir de la statistique de test S, on voit que la puissance du test dépend de S qu'il faut choisir judicieusement. Sous certaines hypothèses plus ou moins restrictives, le lemme de Neyman-Pearson permet de conclure à l'existence d'un tel ensemble. La théorie de Neyman-Pearson pose néanmoins un certain nombre de difficultés pratiques (on ne peut pas la mettre en œuvre aisément) et

théoriques, on ne l'abordera donc pas dans ce cours. Il se trouve que pour un certain nombre de cas, posant notamment des conditions sur la loi de la statistique de test, le test construit « intuitivement » correspond à celui de puissance maximale (voir l'exercice 6). On reviendra sur des questions similaires dans la discussion sur la décision bayésienne dans le cours de tronc commun *De l'apprentissage automatique aux IA génératives* au second semestre.

1.1.4 Résumé : mise en œuvre d'un test statistique d'hypothèse

La démarche est la suivante :

1. choix du niveau de signification α^* ;
2. choix des deux hypothèses $\mathscr{H}_0$ et $\mathscr{H}_1$;
3. choix de la statistique de test S ;
4. définition de l'allure de la région critique en fonction de $\mathscr{H}_1$: comme dit précédemment, cela revient à choisir un intervalle I sur les valeurs de réalisation de S ;
5. détermination de la région critique W ou des bornes de l'intervalle I en fonction du seuil α^* ;
6. calcul (si possible) de la puissance $1-\beta$ du test ;
7. calcul de la valeur s prise par la statistique de test ;
8. *décision* : rejet ou acceptation de $\mathscr{H}_0$ respectivement si $s \in I$, ou si $s \notin I$.

Plusieurs tests peuvent exister dans la littérature pour départager deux hypothèses : ils dépendront essentiellement de la statistique de test choisie, qui ne conduira pas à la même puissance.

Remarque. Seules les étapes 6 et 7 font intervenir les données observées.

Illustrons la démarche sur les exemples. Avant de mettre en œuvre un test, il faut commencer par fixer le niveau de signification. Nous choisissons $\alpha^* = 0,05$. Les points 1 et 2 ont été discutés précédemment.

Exemple 1 (fin)

Ici, on est amené à rejeter l'hypothèse $\mathscr{H}_0$, et donc à choisir l'hypothèse $\mathscr{H}_1$ (la pièce est mal équilibrée) si la statistique de test, qui compte le nombre de « pile », atteint des grandes ou des petites valeurs par rapport à $n/2$, n étant le nombre de tirages. Comme on cherche à maîtriser le risque de première espèce, on va chercher la plus petite valeur seuil s_0 telle que

$$P(S \geqslant n/2 + s_0 \text{ ou } S \leqslant n/2 - s_0 \mid \mathscr{H}_0) < 0,05$$

On s'oriente vers une règle de décision symétrique par rapport à $n/2$ car on n'a pas vraiment de raison de faire un choix différent.

La décision sera de choisir $\mathscr{H}_1$ si la valeur prise par la statistique est inférieure à $n/2 - s_0$ ou supérieure à $n/2 + s_0$. Cela assure bien que le risque de première espèce est inférieur à 0,05. Dans cet exemple, on n'impose pas que le risque de première espèce soit égal à 0,05 car, les distributions étant discrètes, il n'existe pas nécessairement de valeur s_0 réalisant l'égalité.

Comme S suit une loi binomiale et comme $n = 20$, pour $k \in \{0, 1, \ldots, 19, 20\}$:

$$P(S = k \mid \mathscr{H}_0) = \binom{20}{k} \frac{1}{2^{20}}$$

On trouve alors numériquement $s_0 = 5$, plus petit entier tel que :

$$\begin{aligned} &P(S = 0 \mid \mathscr{H}_0) + P(S = 1 \mid \mathscr{H}_0) + \cdots + P(S = 10 - s_0 \mid \mathscr{H}_0) \\ &+ P(s = 10 + s_0 \mid \mathscr{H}_0) + \cdots + P(S = 19 \mid \mathscr{H}_0) + P(S = 20 \mid \mathscr{H}_0) \; < \; 0,05 \end{aligned}$$

La règle de décision, définissant la région critique, est donc :
si $s \in W = \{0, 1, 2, 3, 4, 5, 15, 16, 17, 18, 19, 20\}$,
alors on rejette $\mathscr{H}_0$ et on accepte $\mathscr{H}_1$ (on décide que la pièce est mal équilibrée), au risque au plus 0,05 de se tromper.

La puissance de ce test est alors $P(S \geqslant n/2 + s_0$ ou $S \leqslant n/2 - s_0 | \mathscr{H}_1)$ pour $s_0 = 5$. Toute la difficulté ici est que $\mathscr{H}_1$ est une hypothèse composite (on sait juste $p \neq 1/2$). Il y a alors autant de valeurs de puissance $1 - \beta$ que de valeurs pour p sous l'hypothèse $\mathscr{H}_1$. Lorsque $\mathscr{H}_1$ est composite, on considère la **courbe de puissance** $\gamma(\pi)$, donnée par la formule :

$$\begin{aligned} \gamma(\pi) = &P(S = 0 \mid p = \pi) + \cdots + P(S = 10 - s_0 \mid p = \pi) \\ &+ P(s = 10 + s_0 \mid p = \pi) + \cdots + P(S = 20 \mid p = \pi) \end{aligned}$$

où chaque terme est donné par la loi binomiale de paramètres $n = 20$ et π.

Dans le cas d'une hypothèse $\mathscr{H}_1$ composite, la comparaison entre différents tests se fait en confrontant ces courbes de puissance. On est alors intéressé par le test dont la courbe de puissance domine celle de tous les autres tests de niveau au plus α^*. Un tel test est dit **uniformément le plus puissant** (UPP).

Notons que la mise en place du test est indépendante des données observées. À présent, mettons le test en œuvre. On observe $s = 12$ « pile » sur les 20 tirages : on ne peut donc pas rejeter l'hypothèse $\mathscr{H}_0$. Autrement dit, les observations ne permettent pas de rejeter l'hypothèse selon laquelle la pièce est équilibrée. On accepte donc l'hypothèse $\mathscr{H}_0$.

Exemple 2 (fin)

Ici, on est amené à rejeter l'hypothèse $\mathscr{H}_0$ ($m = 10$) en faveur de $\mathscr{H}_1$ ($m = 9$) si la valeur de la statistique S est trop petite. On cherche donc s_0 tel que $P(S < s_0 | \mathscr{H}_0) = 0,05$, de manière à ce que le risque de première espèce soit 0,05. Comme, sous l'hypothèse $\mathscr{H}_0$, S suit une loi normale de moyenne 10 et de variance $1/5$, s_0 est obtenu par :

$$s_0 = F^{-1}_{10;1/5}(0,05)$$

où $F_{\mu;\sigma^2}$ est la fonction de répartition de la loi normale de moyenne μ et de variance σ^2.

On calcule $s_0 = 9,67$. La puissance $1 - \beta$ du test est donnée par :

$$1 - \beta = P(S < s_0 | \mathscr{H}_1) = F_{9,1/5}(9,67) = 0,99$$

Cela signifie que lorsqu'on choisit $\mathscr{H}_1$, on a très peu de chance de se tromper.

Dans le cas présent, la moyenne empirique des observations est :

$$s = (7,8 + 9,5 + 8,8 + 10,1 + 9,2)/5 = 9,08$$

On peut conclure, comme $9,08 < 9,67$: on rejette l'hypothèse $\mathscr{H}_0$ et on accepte l'hypothèse $\mathscr{H}_1$. On conclut donc à l'efficacité de la réparation, au risque 5% de se tromper (ou de manière équivalente avec un niveau de confiance de 95%). On dit aussi que l'effet de la réparation est **statistiquement significatif** au risque 5%.

1.1.5 Tests unilatéraux et bilatéraux, *p*-valeur

La règle de décision de l'exemple 2 revient à trouver un seuil s_0 tel que $P(S < s_0 \mid \mathscr{H}_0) = \alpha^*$, où le niveau de signification α^* vaut généralement 0,05.

On parle de **test unilatéral à gauche**. De la même manière, un test défini par s_0 tel que $P(S > s_0 \mid \mathscr{H}_0) = \alpha^*$ est appelé **test unilatéral à droite**, et un test basé sur $P(S > m_0 + s_0$ ou $S < m_0 - s_0 \mid \mathscr{H}_0) = \alpha^*$ (c'est-à-dire, lorsque $s_0 > 0$, $P(|S - m_0| > s_0 \mid \mathscr{H}_0) = \alpha^*$) est appelé **test bilatéral**.

Dans le cas du test bilatéral, une ambiguïté subsiste pour fixer les intervalles de rejet, définis par les deux inconnues m_0 et s_0 et une seule équation. On choisit de répartir également le risque sur les deux queues de distributions, de sorte qu'on impose : $P(S > m_0 + s_0 \mid \mathscr{H}_0) = P(S < m_0 - s_0 \mid \mathscr{H}_0) = \alpha^*/2$[1]. Lorsque la distribution de S est symétrique par rapport à 0 sous H_0 (loi de Gauss, loi de Student), cela revient à choisir $m_0 = 0$ et s_0 tel que $P(S > s_0 \mid \mathscr{H}_0) = \alpha^*/2$, s_0 étant déterminée à l'aide de la fonction de répartition F par $s_0 = F^{-1}(1 - \alpha^*/2)$. Lorsque la loi sous H_0 n'est pas symétrique (loi du χ^2, loi de Fisher-Snedecor que l'on verra par la suite), on détermine les bornes des intervalles de rejet à l'aide de la fonction de répartition par $m_0 - s_0 = F^{-1}(\alpha^*/2)$ et $m_0 + s_0 = F^{-1}(1 - \alpha^*/2)$.

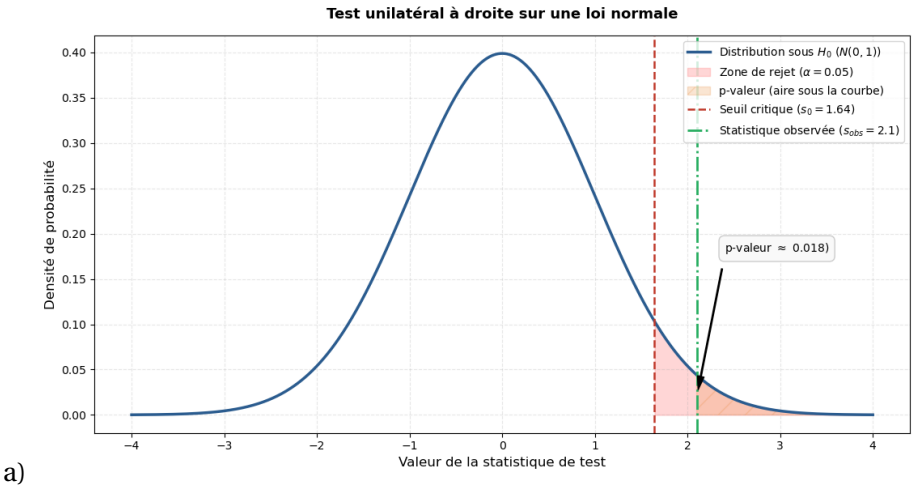
Dans les cas unilatéraux à gauche ou à droite, la valeur de la probabilité $P(S < s \mid \mathscr{H}_0)$, $P(S > s \mid \mathscr{H}_0)$, qui dépend de la valeur s calculée sur les données, est appelée ***p*-valeur**. Dans le cas bilatéral avec une loi de S symétrique, la p-valeur est donnée par $2 \times P(|S| > |s| \mid \mathscr{H}_0)$. Lorsque la loi n'est pas symétrique et que les risques sont répartis également dans les queues de distribution, la p-valeur est définie par la méthode du double de la plus petite queue : lorsque s est plus grand que la médiane, la p-valeur est $2 \times P(|S| > s \mid \mathscr{H}_0)$, et lorsque s est plus petit que la médiane, la p-valeur est $2 \times P(|S| < s \mid \mathscr{H}_0)$.

On voit que rejeter l'hypothèse $\mathscr{H}_0$ en comparant s à s_0 équivaut à rejeter $\mathscr{H}_0$ si la p-valeur du test est inférieure au risque-seuil α^*. Il est équivalent de raisonner sur les valeurs critiques (comparer s à s_0) ou sur les p-valeurs (à comparer à α^*).

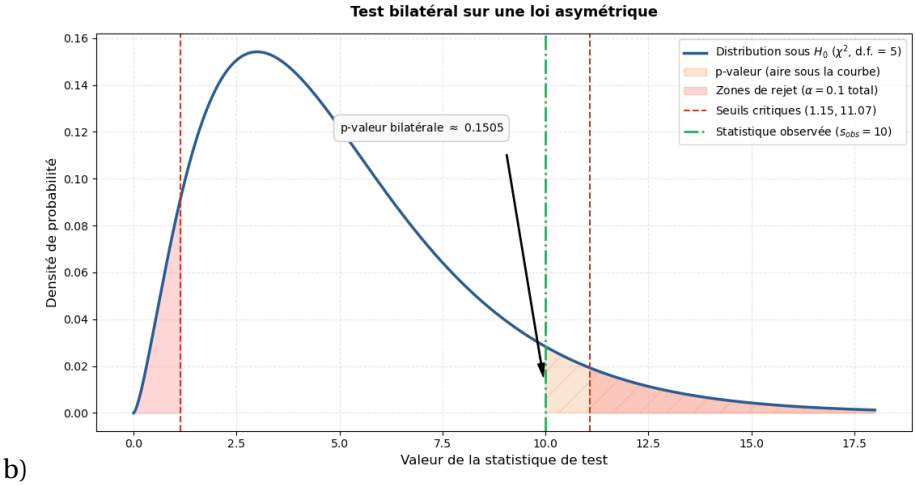
La figure 1.1 illustre la notion de p-valeur.

D'un point de vue fréquentiste, dans une proportion des cas égale à cette probabilité α^*, la réalisation d'un échantillon sous la loi définie par $\mathscr{H}_0$ va fournir « par hasard » une valeur de la statistique de test inférieure à s (pour un test unilatéral à gauche). Voir section 1.1.7.

1. La méthode des lignes de densité détermine les seuils par : $f(m_0 + s_0) = f(m_0 - s_0)$ où f est la densité de S sous $\mathscr{H}_0$, et $P(S > m_0 + s_0 \mid \mathscr{H}_0) + P(S < m_0 - s_0 \mid \mathscr{H}_0) = \alpha^*$.



a)



b)

FIGURE 1.1 – **En haut** *: test unilatéral à droite sur une loi symétrique (loi de la statistique S :* $N(0,1)$ *sous* H_0*), avec un seuil de signification* $\alpha^* = 0,05$*. La zone de rejet* $[s_0, +\infty[$ *est définie par* $P(S > s_0) = \alpha^*$*, soit :* $s_0 = 1,64$*. Ici, la valeur observée de la statistique de test est* $s_{obs} = 2,1$*. On calcule* $P(S > s_0) = 0,018$*, qui est la p-valeur du test. Comme la p-valeur est inférieure à* α^**, on rejette* H_0*. On peut aussi aboutir à la même conclusion en constatant* $s_{obs} > s_0$*.* **En bas** *: test bilatéral sur une loi asymétrique (loi de la statistique S :* χ^2_5 *sous* H_0*), avec un seuil de signification* $\alpha^* = 0,1$*. La zone de rejet, telle que définie dans le texte, est* $]0;1,15] \cup [11,07;+\infty[$*. On observe* $s_{obs} = 10$*, qui est en dehors de la zone de rejet : on ne peut pas rejeter* $\mathcal{H}_0$*. En raisonnant sur la p-valeur : elle vaut* $2 \times P(S > s_{obs}) = 0.15$ *qui est plus grand que* $\alpha^* = 0,1$ *: on aboutit à la même conclusion.*

1.1.6 Points de vue de Neyman-Pearson et de Fisher

La théorie statistique des tests a été développée au début du vingtième siècle. Deux grandes familles d'interprétations se sont opposées, parfois de manière virulente, sur des arguments tenant de la discussion philosophique.

Le point de vue développé dans l'exemple 2 correspond à celui de Neyman et Pearson : deux hypothèses sont confrontées, le seuil α^* permet de définir la région de rejet de $\mathscr{H}_0$ (et la région d'acceptation de $\mathscr{H}_0$) qui sont complémentaires, puis on calcule la puissance du test lorsque c'est possible. Lorsqu'on rejette une hypothèse, on accepte l'autre. Accepter $\mathscr{H}_1$ en rejetant $\mathscr{H}_0$ ne veut pas dire que $\mathscr{H}_1$ est vraie : dans notre exemple, on ne sait pas si la « vraie » valeur de la moyenne du volume horaire de la fuite est de 9 : c'est seulement ce que promet le réparateur. Par ailleurs, il peut sembler paradoxal d'accepter $\mathscr{H}_0$ si on ne maîtrise pas la valeur de l'erreur de seconde espèce β (qui est la probabilité de choisir $\mathscr{H}_0$ à tort) : cette probabilité peut elle-même être grande, en particulier si le test est biaisé.

Le point de vue de Pearson est de travailler essentiellement sur une hypothèse $\mathscr{H}_0$ et de calculer la p-valeur correspondant à la probabilité que la statistique de test S dépasse (pour un test unilatéral à droite) la valeur observée s sous l'hypothèse $\mathscr{H}_0$. La p-valeur quantifie la probabilité qu'une aussi grande valeur soit explicable par le hasard gouvernant les échantillons sous l'hypothèse $\mathscr{H}_0$. Chez Pearson, il n'y a pas vraiment d'hypothèse alternative $\mathscr{H}_1$.

Les cours modernes de statistique tentent de faire une synthèse de ces deux approches. Les logiciels de statistique calculent des p-valeurs : l'interprétation du test se fait uniquement en comparant une p-valeur au seuil de signification décidé au préalable. Si la p-valeur est inférieure au seuil α^*, alors on rejette l'hypothèse $\mathscr{H}_0$ avec un risque α^*. L'effet mis en évidence est statistiquement significatif. Par contre, si la p-valeur est supérieure à α, on ne conclut pas : le test ne permet pas de rejeter $\mathscr{H}_0$, car on ne maîtrise pas l'erreur de seconde espèce β. La valeur s peut être expliquée par le hasard (il n'est pas étonnant d'observer une telle valeur produite par un échantillon aléatoire sous l'hypothèse $\mathscr{H}_0$) ou autre chose. Un autre test ou des données complémentaires permettront peut-être de conclure. En d'autres termes, l'absence de preuve d'un effet n'est pas preuve d'absence de cet effet.

Il semble que la majorité des auteurs suive l'approche de Pearson : les données permettent (ou pas) de rejeter l'hypothèse $\mathscr{H}_0$, avec un certain risque d'erreur, mais ne permettent pas pour autant d'accepter une hypothèse alternative.

1.1.7 Complément facultatif : *p*-hacking

Si on fixe la valeur de α^* à 5%, on est amené à rejeter l'hypothèse $\mathscr{H}_0$ par erreur avec une probabilité de 0,05. Généralement, rejeter $\mathscr{H}_0$ permet de juger qu'un phénomène est statistiquement significatif.

Si on fait un grand nombre de tests aléatoires indépendants entre eux sur un jeu de données, un certain nombre, certes petit, d'hypothèses $\mathscr{H}_0$ vont donc être rejetées par hasard : si on conduit N tests, on obtiendra de l'ordre de $N\alpha$ rejets « par hasard ». Cela conduit à juger qu'un phénomène est statistiquement significatif alors que le résultat est le fruit du hasard. Un chercheur inattentif (ou peu scrupuleux…) pourrait être amené à sélectionner le test conduisant à un résultat favorable pour son étude sans parler des nombreux autres tests

défavorables. Ces démarches peu rigoureuses sont regroupées sous le terme générique de *p-hacking*. Voir l'illustration humoristique à cette URL : https://xkcd.com/882/.

Dans le cas de tests d'hypothèses multiples, une manière simple de contrer ce phénomène est d'appliquer la *correction de Bonferroni*, qui consiste à utiliser comme niveau de confiance α^*/N, où N est le nombre de tests. Ainsi, le nombre de rejets « par hasard » restera globalement de l'ordre de α^* (typiquement 5%). Un autre méthode est de contrôler le taux de fausses découvertes, voir : https://en.wikipedia.org/wiki/False_discovery_rate. Cette approche est basée sur le fait que les p-valeurs sont distribuées uniformément sur $[0,1]$ sous $\mathscr{H}_0$[2].

1.2 Quelques tests paramétriques basiques

La suite de ce document présente quelques tests paramétriques classiques s'inscrivant dans la démarche de la section 1.1.4.

On suppose disposer d'un n-échantillon $(X_1,\dots,X_n)$ issu d'une variable aléatoire gaussienne.

1.2.1 Test sur la moyenne d'une loi normale

On teste l'hypothèse $\mathscr{H}_0 : m = \mu_0$ contre l'hypothèse $\mathscr{H}_1 : m \neq \mu_0$.

1.2.1.1 Variance σ^2 connue

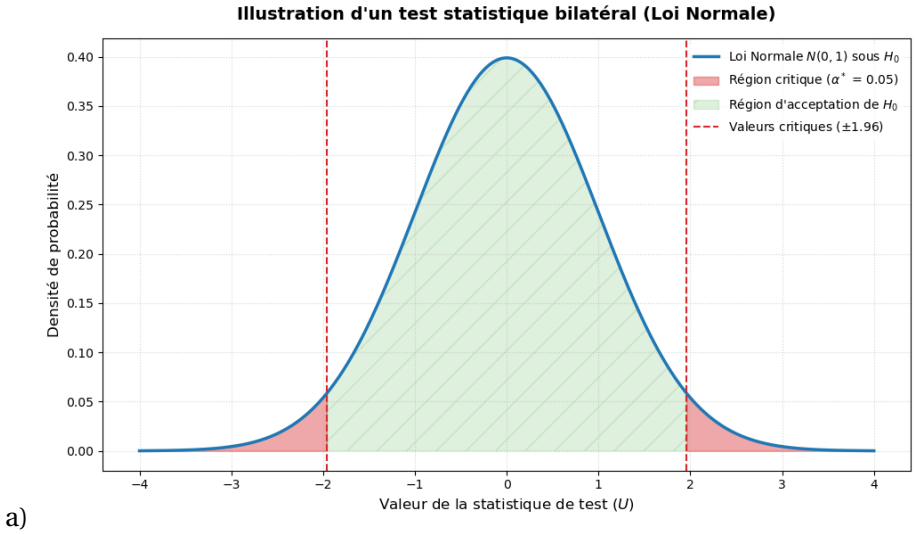
La statistique de test est la moyenne empirique $\overline{X}$: sous $\mathscr{H}_0$, $\overline{X}$ suit la loi normale de moyenne μ_0 et de variance σ^2/n, donc $U = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ suit une loi normale centrée réduite. On veut savoir à quel point la variable aléatoire U peut s'éloigner de 0 à l'aide d'un test bilatéral, symétrique par rapport à 0. On cherche donc une valeur seuil u_0 telle que : $P(|U| > u_0) = \alpha^*$. Notons $u_{\alpha^*/2}$ le quantile tel que $P(U > u_{\alpha^*/2}) = \alpha^*/2$.

La conclusion est la suivante : si la moyenne $\overline{x}$ des observations est telle que $(\overline{x}-\mu_0)/(\sigma/\sqrt{n})$ n'est pas dans l'intervalle $[-u_{\alpha^*/2}, u_{\alpha^*/2}]$, alors on peut rejeter l'hypothèse $\mathscr{H}_0$ au risque α^*.

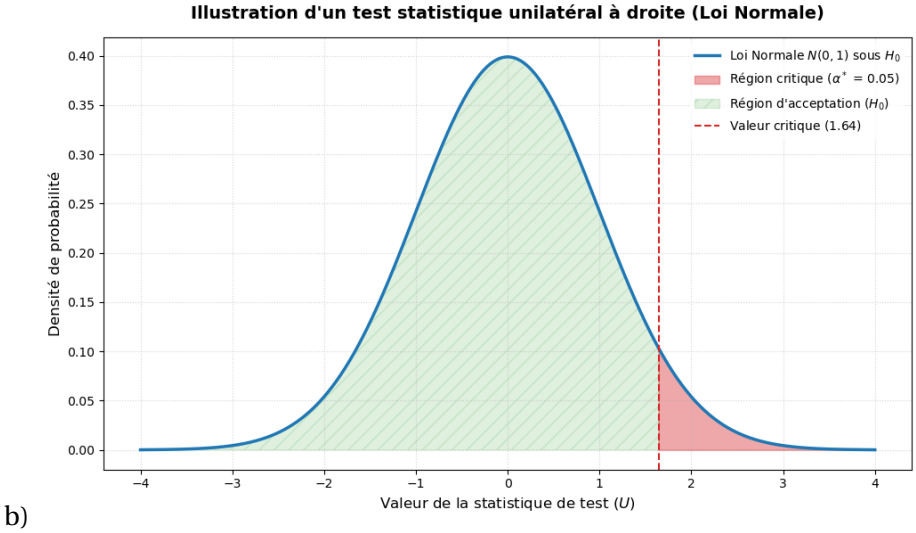
La figure 1.2 a) illustre ce test bilatéral. On note que dans le cas bilatéral, le risque est réparti symétriquement dans les deux queues de la distribution.

La figure 1.2 b) illustre un test unilatéral à droite basé sur $P(U > k) = \alpha^*$, que l'on utiliserait dans le cas de l'hypothèse alternative $\mathscr{H}_1 : m > \mu_0$.

2. Une p-valeur pour un test à gauche s'écrit $F(s) = P(S < s \mid \mathscr{H}_0)$ où F est la fonction de répartition (supposée ici inversible) de la statistique de test S et s une réalisation de S. On considère la variable aléatoire $F(S)$ des p-valeurs vues comme des valeurs aléatoires. $F(S)$ prend donc des valeurs sur $[0,1]$. Notons $F_{F(S)}$ sa fonction de répartition. Alors $F_{F(s)}(p) = P(F(S) < p) = P(S < F^{-1}(p)) = F(F^{-1}(p)) = p$. La fonction de répartition suffisant à caractériser une loi, les p-valeurs sont bien distribuées selon la loi uniforme sur $[0,1]$. Le raisonnement s'adapte pour un test à droite ou un test bilatéral.



a)



b)

FIGURE 1.2 – **En haut** *: test bilatéral pour une statistique suivant la loi normale. Avec un risque* $\alpha^* = 5\%$*, les seuils critiques sont* $\pm u_{2,5\%} = \pm 1,96$*.* **En bas** *: test unilatéral à droite pour une statistique suivant la loi normale. Avec un risque* $\alpha^* = 5\%$*, la valeur critique est* $u_{5\%} = 1,645$*.*

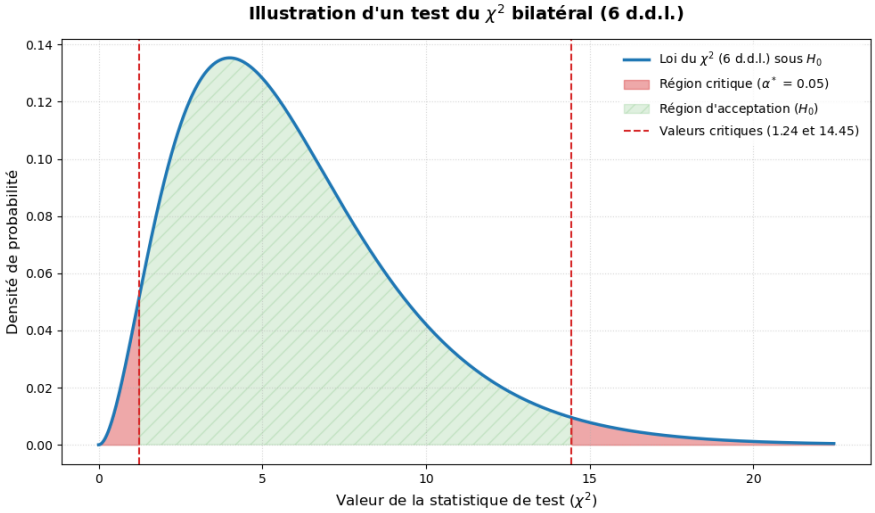


FIGURE 1.3 – *Test bilatéral pour une statistique suivant la loi du χ^2 à six degrés de liberté.*

1.2.1.2 Variance inconnue

On ne peut plus normaliser pour calculer U car on ne connaît pas la variance σ^2 : il faut utiliser la variance empirique.

La statistique de test est : $T = \sqrt{n-1}(\overline{X} - m)/S_n$, où $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. On sait que T suit une loi de Student à $n-1$ degrés de liberté[3]. On cherche le seuil t_0 tel que $P(|T| > t_0) = \alpha$, ce qui ramène à regarder les quantiles de la loi de Student. La représentation graphique est proche de celle de la figure 1.2 ; bien entendu les quantiles ne sont pas les mêmes lorsque n « n'est pas très grand » (rappelons que la loi de Student tend vers la loi Gaussienne quand $n \to +\infty$).

Remarque. Dans ces deux cas, ce qui est important est que la moyenne empirique suive une loi normale. Si les X_i ne suivent pas une loi normale mais que la taille n de l'échantillon est « assez grande », on peut appliquer les résultats précédents car le théorème central-limite nous assure que la loi de la moyenne empirique est proche d'une loi normale. Un test dont les conclusions restent valables quand on s'écarte un peu des suppositions de départ sur la loi de X (ici, le caractère gaussien) est appelé **test robuste**.

1.2.2 Test sur la variance d'une loi normale

On veut tester $\mathscr{H}_0 : \sigma = \sigma_0$ contre $\mathscr{H}_0 : \sigma \neq \sigma_0$

3. C'est le corollaire 16 du chapitre 2 de la première partie du cours, qui est une conséquence du théorème de Cochran.

1.2.2.1 Moyenne *m* connue

La statistique de test est nS_n/σ^2 où $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$. On sait que nS_n/σ^2 suit une loi du χ^2 à n degrés de liberté (par définition de cette loi) car cette quantité est le carré de la norme du vecteur gaussien centré réduit de dimension n :

$$((X_1 - m)/\sigma, (X_2 - m)/\sigma, \ldots, (X_n - m)/\sigma)$$

On utilise un test bilatéral.

La règle de décision est illustrée par la figure 1.3. Remarquons que la loi du χ^2 n'est pas symétrique. Le risque est réparti à part égale entre les queues de droite et de gauche de la distribution, comme expliqué en section 1.1.5.

1.2.2.2 Moyenne inconnue

On utilise la statistique nS/σ^2 où cette fois $S_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est la variance empirique. On sait que nS/σ^2 suit une loi du χ^2 à $n-1$ degrés de liberté[4]. Le principe de décision reste le même que lorsque la moyenne est connue.

1.3 Exercices

1.3.1 Risques

Des pièces industrielles sont produites selon un procédé qui nécessite des matières premières. Il y a une certaine variabilité de la quantité de matières premières nécessaires : chaque pièce nécessite une masse (en kilogrammes) de matières premières répartie selon une loi normale de moyenne $m = 100$ et d'écart-type $\sigma = 10$. Le procédé de production est modifié dans l'objectif d'atteindre une quantité de matière première nécessaire de moyenne $m = 80$ et de même écart-type.

Le responsable de production (qui n'a pas suivi le cours de statistique) juge que l'objectif est atteint si la moyenne de la quantité de matière première atteint 88 kilogrammes par pièce.

1) Quels sont les risques associés à cette décision ?

2) Proposez un critère objectif permettant de juger du succès de la modification du procédé de production.

1.3.2 Courbe de puissance

D'après un exercice du cours de statistique 2018 de l'ENSTA.

Soit $(X_1, \ldots, X_n)$ un échantillon de loi gaussienne d'espérance μ inconnue et de variance $\sigma^2 = 100$ connue.

4. Voir la proposition 15 du chapitre 2 de la première partie du cours de statistique, conséquence du théorème de Cochran.

1) Construire un test de niveau $\alpha = 0.10$ de l'hypothèse nulle "$\mu = 0$" contre l'hypothèse alternative "$\mu = 1.5$", fondé sur la moyenne empirique, estimateur du paramètre μ.

Supposons $n = 25$. On observe $\overline{x} = 1$. Quelle est la décision du test?

Quelle est l'erreur de seconde espèce du test?

2) On souhaite tester $\mathscr{H}_0 : \mu = 2$ contre $\mathscr{H}_1 : \mu < 2$. Définir la région de rejet. Calculer la puissance du test et étudier ses variations en fonction de μ, n et σ.

1.3.3 Test de données appariées

On souhaite mesurer l'effet sur le poids d'un programme d'arrêt du tabac. On dispose du poids (en kilogrammes) de dix personnes avant et après l'arrêt du tabac.

Individu	1	2	3	4	5	6	7	8	9	10
Avant	61,2	66,3	80,4	70,5	73,0	78,1	57,9	71,4	74,2	75,3
Après	62,0	64,3	81,5	68,3	72,9	77,8	63,2	70,3	75,1	75,0

L'effet est-il statistiquement significatif?

Indication : si $d_i = x'_i - x_i$ avec x_i et x'_i les poids respectivement avant et après arrêt de l'individu i, on calcule : $\bar{d} = \frac{1}{10}\sum_i d_i = 0,21$ et : $s_d^2 = \frac{1}{9}\sum_i (d_i - \bar{d})^2 \simeq 4,505$.

1.3.4 Contrôle par échantillonnage

Un ingénieur dans un service achat est chargé du contrôle de la qualité de pièces livrées par un fournisseur. Contractuellement, celui-ci s'engage à fournir une proportion $p = 0,9$ de pièces conformes. Sur une livraison de plusieurs milliers de pièces, notre ingénieur prélève 10 pièces, et en juge 3 non conformes.

1) Montrez qu'il est difficile de rejeter la livraison.

Indication : Si X est une variable aléatoire binomiale de paramètres $10; p$, vous calculerez la p-valeur de $X \geqslant 3$ sous l'hypothèse $\mathscr{H}_0 : p = 0,1$.
On donne, si $p = 0,1$: $P(X = 0) = 0,3487$; $P(X = 1) = 0,3874$; $P(X = 2) = 0,1937$.

2) Pour en avoir le cœur net, l'ingénieur fait prélever 50 pièces, et en trouve 15 non conformes. Que conclure?

Indication : l'inégalité de Hoeffding dit que si S_n suit une loi binomiale de paramètres n, p, alors :

$$P(S_n - np \geqslant t) \leqslant e^{-2t^2/n}$$

1.3.5 Vérification de précision

On utilise une balance de laboratoire censée fournir des mesures de poids à une précision de $\pm 0,10$ gramme. Une manière raisonnable d'interpréter cette notion de précision (en suivant la norme ISO 5725) est de supposer que les mesures sont affectées d'une erreur aléatoire de moyenne nulle et d'écart-type 0,1 gramme.

On effectue 10 mesures du même objet, on obtient des valeurs x_i telles que :
$\sum_{i=1}^{10} x_i = 10,043$ et $\sum_{i=1}^{10} x_i^2 = 10,456$.
Ces données sont-elles conformes à la précision annoncée par le fabricant ?

1.3.6 Lemme de Neyman-Pearson dans un cas particulier

Soit $(X_1,\dots,X_n)$ un n-échantillon dépendant d'un paramètre θ, de densité jointe $f(x_1,\dots,x_n;\theta)$ (dans le cas où l'échantillon est i.i.d., la loi jointe est le produit $g(x_1;\theta)\dots g(x_n;\theta)$ où g est la densité de la loi commune aux X_i) . On suppose que la densité f ne s'annule pas. On considère le problème du test d'une hypothèse simple $H_0 : \theta = \theta_0$ contre une hypothèse alternative simple $H_1 : \theta = \theta_1$.

On définit la région de rejet W par le rapport de vraisemblance suivant :

$$W = \left\{ x \in \mathbb{R}^n : \frac{f(x;\theta_1)}{f(x;\theta_0)} \geq k \right\}$$

où $k > 0$ est une constante telle que le risque de première espèce soit $P(X \in W \mid H_0) = \alpha$. Une telle constante k existe sous des hypothèses réalistes : si k est grand, alors $W = \emptyset$ et $P(X \in W|H_0) = 0$, et si k est trop petit, alors $W = \mathbb{R}$ et $P(X \in W|H_0) = 0$. Par le théorème des valeurs intermédiaires (sous hypothèse de continuité), k existe.

Partie A

L'objectif est de démontrer que pour toute autre région de rejet W' telle que $P(X \in W' \mid H_0) \leq \alpha$, la puissance de W est supérieure ou égale à celle de W'.

1) Soient $\mathbb{1}_W$ et $\mathbb{1}_{W'}$ les fonctions indicatrices des régions W et W'. Justifier que pour tout $x \in \mathbb{R}$:

$$(\mathbb{1}_W(x) - \mathbb{1}_{W'}(x)) \cdot (f(x;\theta_1) - k f(x;\theta_0)) \geq 0$$

2) En intégrant l'inégalité précédente sur tout l'espace des observations, montrer que :

$$\int (\mathbb{1}_W(x) - \mathbb{1}_{W'}(x)) f(x;\theta_1)\,dx \geq k \int (\mathbb{1}_W(x) - \mathbb{1}_{W'}(x)) f(x;\theta_0)\,dx$$

3) Exprimer les termes de gauche en fonction de la puissance du test pour la région W (notée $1-\beta$) et pour la région W' (notée $1-\beta'$).

4) En utilisant la condition sur le risque α, conclure la démonstration.

Partie B

Soit $X_1,\dots,X_n$ un échantillon i.i.d. suivant $\mathcal{N}(\mu,\sigma^2)$ avec σ^2 connu. On teste $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1$ (où $\mu_1 > \mu_0$).

1) Établir le rapport de vraisemblance $\Lambda(x_1,\dots,x_n)$.

2) Montrer que le test le plus puissant de niveau α rejette H_0 si $\bar{X}_n > c$.

3) Déterminer c pour un risque α donné.

Chapitre 2

Tests statistiques d'hypothèse (2)

Le chapitre précédent porte sur les tests statistiques d'hypothèse dont la mise en œuvre nécessite que certaines propriétés sur la loi des échantillons soient satisfaites, notamment une distribution gaussienne. Dans le présent chapitre, nous discuterons plusieurs tests permettant d'étudier les échantillons : comparaison d'échantillons ou étude de leur distribution. De très nombreux tests sont disponibles dans la littérature, nous n'en évoquerons que quelques-uns pour donner un panel des possibilités qui s'offrent au statisticien. L'essentiel est d'avoir compris le principe de fonctionnement des tests, afin d'être capable de mettre en œuvre un test qui ne serait pas listé ici, mais qu'on pourrait trouver dans la littérature...ou qui pourrait être proposé par une IA générative.

2.1 Tests de comparaison de deux échantillons

Dans cette section, la question est de savoir si deux échantillons peuvent être discernés d'un point de vue statistique. Autrement dit, on veut savoir si les deux échantillons ont été prélevés de manière indépendante dans la même population homogène. D'un point de vue pratique, on va se demander si deux jeux d'observations sont des réalisations d'échantillons de mêmes lois.

Exemple 1

Les notes à l'examen du groupe A1 et du groupe A2 sont-elles comparables ? Ou alors la différence est-elle significative ?

2.1.1 Comparaison de deux échantillons gaussiens indépendants

On suppose disposer de deux échantillons gaussiens indépendants de moyennes m_1 et m_2, de variances σ_1^2 et σ_2^2, et de tailles n_1 et n_2, respectivement.

Les hypothèses sont :

— $\mathscr{H}_0 : m_1 = m_2$ et $\sigma_1 = \sigma_2$

— $\mathscr{H}_1 : m_1 \neq m_2$ ou $\sigma_1 \neq \sigma_2$

Pour une raison qu'on va voir, on commence par tester l'égalité des variances.

On sait que si S_1^2 et S_2^2 désignent les estimateurs non corrigés de la variance des deux échantillons, alors les statistiques $n_1 S_1^2/\sigma_1^2$ et $n_2 S_2^2/\sigma_2^2$ sont distribuées selon des lois du χ^2 à respectivement $n_1 - 1$ et $n_2 - 1$ degrés de liberté (d.d.l.) [1].

Introduisons la statistique de test :

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

Comme S_1 et S_2 sont indépendants (comme les deux échantillons), cette statistique suit une **loi de Fisher-Snedecor** à $(n_1 - 1, n_2 - 1)$ degrés de liberté. Voir l'annexe A.4 pour la définition et la table de cette loi.

Sous l'hypothèse $\mathscr{H}_0$ ($\sigma_1 = \sigma_2$), F devient :

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

où S_1^* et S_2^* désignent les estimateurs corrigés de la variance. Si on met la plus grande des deux quantités au numérateur, le rapport est supérieur à 1, et proche de 1 lorsque $\sigma_1 = \sigma_2$.

On va donc mettre en place un test statistique unilatéral à droite. On cherche une région critique de la forme $[k, +\infty[$ où $k > 1$ vérifie $P(F > k \mid \mathscr{H}_0) = \alpha^*$.

Si les variances sont jugées égales ($\sigma_1 = \sigma_2 = \sigma$, car l'hypothèse $\mathscr{H}_0$ n'est pas rejetée par le test précédent), on passe au test des espérances. L'objectif est de tester si la différence des moyennes empiriques sur les deux échantillons est significative. On va introduire une statistique de test constuire en centrant et réduisant la différence des moyennes empiriques, dans l'esprit de ce qui a été fait en section 1.2.1.2.

Comme la somme de deux variables indépendantes suivant une loi du χ^2 suit encore une loi du χ^2 dont le d.d.l. est la somme des d.d.l. des deux variables, la variable aléatoire $\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ suit une loi du χ^2 à $n_1 + n_2 - 2$ d.d.l.

Par ailleurs, les moyennes empiriques $\overline{X_1}$ et $\overline{X_2}$ sont distribuées selon des lois gaussiennes de moyennes m_1 et m_2 et de variances σ^2/n_1 et σ^2/n_2, respectivement. Ainsi, $\overline{X_1} - \overline{X_2}$ suit une loi gaussienne de moyenne $m_1 - m_2$ et de variance $\sigma^2(1/n_1 + 1/n_2)$ (car $\overline{X_1}$ et $\overline{X_2}$ sont indépendantes).

Ici, σ est inconnu. On introduit donc :

$$T = \frac{(\overline{X_1} - \overline{X_2} - (m_1 - m_2))/(\sigma\sqrt{1/n_1 + 1/n_2})}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)/(\sigma^2(n_1 + n_2 - 2))}} = \frac{\overline{X_1} - \overline{X_2} - (m_1 - m_2)}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)}}\sqrt{n_1 + n_2 - 2}$$

qui suit une loi de Student à $n_1 + n_2 - 2$ d.d.l. (c'est la conséquence de la définition de cette loi : T s'écrit sous la forme $Z/\sqrt{U/(n_1 + n_2 - 2)}$ où $Z \sim N(0,1)$ et $U \sim \chi^2_{n_1+n_2-2}$).

1. Si $(X_1, \ldots, X_n)$ est un n-échantillon de loi $\mathscr{N}(\mu, \sigma^2)$ et S^2 est la variance empirique (non corrigée), alors nS^2/σ^2 suit la loi du χ^2 à $n-1$ d.d.l. Voir la proposition 15 du chapitre 2 de la première partie du cours de statistique, conséquence du théorème de Cochran.

Sous $\mathscr{H}_0 : m_1 = m_2$, on fait un test bilatéral sur la statistique de test

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)}} \sqrt{n_1 + n_2 - 2}$$

Remarque. Le test de Student est robuste. Si n_1 et n_2 sont « grands » (supérieurs à 20, ou 30 selon les auteurs...), alors le test d'égalité des moyennes peut tout de même être utilisé si les échantillons ne sont pas gaussiens et même si $\sigma_1 \neq \sigma_2$.

Remarque. Dans l'exercice 3 de la séance précédente (données appariées), on ne peut pas utiliser ces tests car les deux populations ne sont pas indépendantes.

2.1.2 Comparaison de plusieurs échantillons décrits par une variable qualitative

Une variable aléatoire **qualitative** (par opposition aux variables aléatoires **quantitatives** vues jusqu'à présent) est une variable prenant un nombre fini de valeurs. Les valeurs possibles sont appelées les **modalités**.

On suppose disposer de k échantillons d'une variable aléatoire à r modalités dont on observe une réalisation. Si on note n_{ij} le nombre d'individus de l'échantillon i qui possèdent la modalité j, on peut représenter les données sous la forme d'un tableau :

	modalité 1	modalité 2	...	modalité r	total
échantillon 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
échantillon 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
$\vdots$					
échantillon k	n_{k1}	n_{k2}		n_{kr}	$n_{k.}$
total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

L'hypothèse que l'on teste est $\mathscr{H}_0$: « les échantillons proviennent de la même distribution »

Notons qu'il s'agit d'un **test non-paramétrique**.

Exemple 2

On se demande si trois traitements présentent la même efficacité. On forme trois groupes ($k = 3$ échantillons) de malades traités chacun avec un des traitements. Chaque malade a, après traitement, des symptômes modérés ou faibles ($r = 2$ modalités). Les trois groupes sont-ils discernables d'un point de vue statistique ? (l'alternative étant qu'on ne peut pas exclure qu'ils sont distribués de la même manière, ce qui suggère que les traitements ont la même efficacité, ou la même inefficacité).

Le tableau regroupant les effectifs observés est :

n_{ij} observé	symptômes modérés	symptômes faibles	total
groupe 1	50	72	122
groupe 2	45	80	125
groupe 3	35	48	83
total	130	200	330

Sous $\mathscr{H}_0$, notons $p_1, p_2, \dots p_r$ les probabilités des r modalités (les mêmes probabilités pour tous les échantillons donc). Les p_j ne sont pas connus, on les estime par les fréquences empiriques $n_{.j}/n$. Sous cette hypothèse, l'effectif moyen attendu pour l'échantillon i et modalité j est $n_{i.}p_j \simeq n_{i.}n_{.j}/n$. On note m_{ij} cette dernière quantité.

On admet alors (c'est le **théorème de Pearson**) que, sous $\mathscr{H}_0$,

$$d^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} \tag{2.1}$$

est réalisation d'une variable aléatoire D^2 qui converge en loi lorsque $n \to +\infty$ vers une variable aléatoire de loi du χ^2 à $(k-1)(r-1)$ d.d.l. C'est une conséquence du théorème central limite dans sa version multivariée (ce qui explique pourquoi d^2 apparaît comme le carré de la norme d'un vecteur gaussien, d'où la loi du χ^2). Sous $\mathscr{H}_1$, D^2 tend vers l'infini presque sûrement.

Comme d^2 est petit sous $\mathscr{H}_0$ (et grand sous $\mathscr{H}_1$), cela veut dire que si les effectifs sont « grands » (pour utiliser la loi limite lorsque $n \to +\infty$), on peut rejeter $\mathscr{H}_0$ sous un test unilatéral à droite selon la loi du χ^2 à $(k-1)(r-1)$ d.d.l.

Il s'agit du **test d'homogénéité du χ^2**. Nous avons déjà vu et nous verrons d'autres tests du χ^2.

Exemple 2 (suite)

On calcule les effectifs attendus sous hypothèse $\mathscr{H}_0$:

m_{ij} attendu	symptômes modérés	symptômes faibles	total
groupe 1	48,06	73,94	122
groupe 2	49,24	75,76	125
groupe 3	32,70	50,30	83
total	130	200	330

La valeur de d^2 est donc : $d^2 \approx 0,0783 + 0,0509 + 0,3651 + 0,2373 + 0,1618 + 0,1052 \approx 0,9986$.

Le valeur critique du test du χ^2 à $(3-1)(2-1) = 2$ degrés de liberté à droite vaut, d'après l'annexe A.3 : 5,99. Comme $1 < 5,99$, on ne peut pas rejeter l'hypothèse que les trois groupes de patients sont distribués identiquement, ce qui indique une équivalence de l'efficacité des traitements.

Remarque. Ce test peut être appliqué à des échantillons décrits par une variable quantitative, en regroupant les valeurs. Par exemple, si X est une variable aléatoire mesurant la tem-

pérature, on peut former les modalités « froid », « tempéré », « chaud » en fixant des seuils de température.

2.2 Tests d'ajustement à une loi

Il s'agit à présent de savoir si un échantillon est distribué selon une loi donnée. On a vu (dans le chapitre précédent) que de nombreux tests supposent que l'échantillon est gaussien. La première étape, avant même la mise en œuvre du test, serait donc de vérifier cette hypothèse de gaussianité.

2.2.1 Méthodes qualitatives empiriques

Avant de mettre en œuvre des tests statistiques, un certain nombre de méthodes empiriques permettent de vérifier que l'hypothèse sur la distribution des données est réaliste. C'est la première chose à faire dans toute étude statistique, d'autant plus sur un logiciel où ces informations sont facilement accessibles.

On peut commencer par tracer l'histogramme des données, et voir par exemple si elles ont la forme de la distribution attendue, comme une « courbe en cloche » pour une distribution gaussienne. Voir la figure 2.1 et le chapitre 1 de la première partie du cours. Ensuite, on calcule des moments empiriques. Par exemple, pour une loi de Poisson on sait qu'espérance (moment d'ordre 1) et variance (moment d'ordre 2 centré) doivent être égales, et pour une loi gaussienne le coefficient d'asymétrie (*skewness*, moment d'ordre 3 centré réduit) est nul et le coefficient d'aplatissement (*kurtosis*, moment d'ordre 4 centré réduit) est 3. Il existe également des ajustements graphiques que l'on ne détaillera pas ici : une transformation des observations permet de les représenter dans un domaine où elles sont réparties sur une droite. Par exemple, les logiciels fournissent généralement un *QQ-plot* qui représentent les observations sur la *droite de Henry* si elles sont bien distribuées de manière gaussienne.

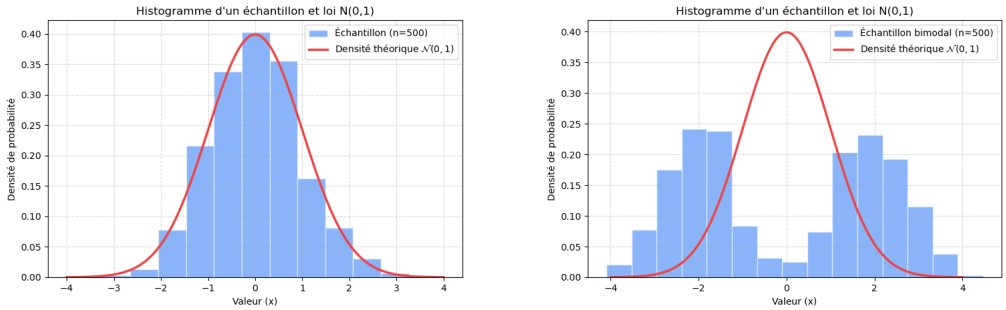


FIGURE 2.1 – *Histogrammes d'un échantillon de taille 500 et densité de la loi $N(0,1)$. À gauche : l'allure de l'histogramme permet de dire que l'échantillon est sans doute tiré selon la loi normale $N(0,1)$. À droite la présence de deux modes dans l'histogramme rend cette hypothèse très peu plausible.*

On aura compris que ces méthodes empiriques donnent des indications mais ne permettent pas de trancher dans un sens ou dans l'autre, de par leur nature qualitative.

2.2.2 Tests statistiques

Parmi les nombreux **tests d'ajustement** de la littérature (on parle aussi de **tests d'adéquation** ou de **tests de conformité**), nous en discutons deux.

2.2.2.1 Test d'ajustement du χ^2

On considère une variable aléatoire X qualitative ou quantitative mais discrétisée comme dans la section 2.1.2 : X prend des valeurs dans un ensemble fini de taille k, chaque valeur ayant une probabilité p_j ($1 \leq j \leq k$).

On dispose d'observations : n_j dans la classe j ($1 \leqslant j \leqslant k$). La taille de l'échantillon est $n = \sum_{j=1}^{k} n_j$. Si les observations sont la réalisation d'un n-échantillon de X (hypothèse $\mathscr{H}_0$), l'effectif moyen attendu dans chaque classe j est $m_j = np_j$

On calcule alors :

$$d^2 = \sum_{j=1}^{k} \frac{(n_j - m_j)^2}{m_j}$$

On admet que, sous $\mathscr{H}_0$, d^2 est réalisation d'une variable aléatoire D^2 qui converge en loi vers une variable aléatoire de loi du χ^2 à $k-1$ degrés de liberté, et sous $\mathscr{H}_1$ (le n-échantillon n'est pas distribué comme X), D^2 converge vers l'infini p.s.

Il s'agit du **test d'ajustement du χ^2**. Attention, il ne doit pas être confondu avec le test d'homogénéité du χ^2 (décrit en section 2.1.2) malgré les similitudes.

Comme d^2 est petit sous $\mathscr{H}_0$, on peut rejeter $\mathscr{H}_0$ à l'aide d'un test unilatéral à droite selon cette loi, si n est « suffisamment grand ». En pratique, on demande à ce que tous les np_i soit supérieurs à 5 (cela dépend des auteurs). Si ce n'est pas le cas, on peut regrouper certaines des modalités de X.

Remarque. Lorsque les probabilités p_j sont issues d'une loi paramétrique dont le (ou les) paramètre(s) est estimé sur les observations (par exemple : on veut tester l'ajustement à une variable gaussienne discrétisée, il faut alors estimer moyenne et variance), il faut diminuer le nombre de degrés de libertés et considérer, si p est le nombre de paramètres estimés, $k-1-p$ d.d.l.

2.2.2.2 Test d'ajustement de Kolmogorov-Smirnov

Soit X une variable aléatoire de fonction de répartition F et $(X_1, \dots X_n)$ un n-échantillon distribué comme X.

La fonction de répartition empirique de cet échantillon est définie pour tout $x \in \mathbb{R}$ par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant x}$$

où $\mathbf{1}_A$ désigne l'indicatrice de l'événement A.

$F_n(x)$ désigne donc la proportion d'éléments du n-échantillon qui sont inférieurs à x. Il s'agit d'une variable aléatoire car $F_n(x)$ dépend du n-échantillon.

On admet la convergence suivante :

$$P\left(\sup_x |F_n(x) - F(x)| > \frac{c}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} \alpha(c)$$

pour tout $c > 0$, où $\alpha(c)$ est une fonction que l'on sait calculer explicitement (la formule est un peu compliquée, on ne la précise pas ici), qui est décroissante par rapport à c. Le point remarquable est que $\alpha(c)$ ne dépend pas de F (et donc de la distribution de X). Voir aussi la **distance de Kolmogorov-Smirnov**, définition 13 du chapitre 2 de la première partie du cours « Inférence statistique ».

On en déduit donc un test statistique unilatéral à droite si n est « assez grand », basé sur la statistique $\sqrt{n}D_n$ où $D_n = \sup_x |F_n(x) - F(x)|$ mesure l'écart entre la fonction de répartition empirique F_n détermnée sur l'échantillon et la fonction de répartition supposée F . Par exemple, $\alpha(c) = 0,05$ si $c = 1,358$: on rejette donc « le n-échantillon est distribué comme X » (hypothèse $\mathscr{H}_0$) au risque 5% si $D_n > 1,358/\sqrt{n}$. Par ailleurs, $\alpha(c) = 0,01$ si $c = 1,629$: cette valeur est utile pour un test au risque de 1%.

Il s'agit du **test de Kolmogorov-Smirnov**. Il s'étend à la comparaison de deux fonctions empiriques. Il permet alors de faire un test d'homogénéité.

Remarque : attention, si la fonction de répartition F fait intervenir des paramètres estimés empiriquement sur l'échantillon, en toute rigueur le test n'est plus valable. Par exemple, si on suppose X de loi gaussienne dont les paramètres sont estimés sur l'échantillon, alors on utilise plutôt le test de Lilliefors ou le test de Shapiro-Wilk.

2.3 Tests d'indépendance

Dans de nombreuses situations, on est intéressé par savoir si deux variables aléatoires sont statistiquement indépendantes.

Exemple 3

Soit X une variable aléatoire modélisant la catégorie socio-professionnelle (selon 7 modalités) et Y l'orientation politique (selon 5 modalités). On fait un sondage qui fournit des effectifs pour chaque couple de catégorie socio-professionnelle et orientation politique. Ces données peuvent-elles permettre de rejeter ou accepter une hypothèse d'indépendance entre X et Y ?

Exemple 4

Soient X et Y deux variables aléatoires modélisant le taux de deux hormones dans le sang. Après avoir recueilli des échantillons sur différentes personnes, on aimerait décider si X et Y sont indépendantes.

Nous donnons ici quelques tests utiles pour des variables qualitatives ou quantitatives comme dans les exemples précédents (variable qualitative dans l'exemple 3, et quantitative dans l'exemple 4).

On considère dans la suite deux variables X et Y, et l'hypothèse $\mathscr{H}_0$: « les deux variables sont indépendantes ».

2.3.1 Variables quantitatives

Les deux tests les plus courants sont des tests sur des coefficients de corrélation.

2.3.1.1 Cas gaussien : coefficient de corrélation de Pearson

Lorsque (X,Y) est un vecteur gaussien, l'indépendance de X et Y est équivalente à la nullité du **coefficient de corrélation de Pearson** (c'est celui que vous connaissez) entre X et Y.

Rappelons que ce coefficient de corrélation est :

$$\rho_P(X,Y) = \frac{\text{cov}(X,Y)}{\text{var}(X)\text{var(Y)}}$$

À partir de deux n-échantillons $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_n)$, on calcule le coefficient de corrélation empirique :

$$R = \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2 \sum_{i=1}^n (Y_i-\overline{Y})^2}}$$

Il est possible d'exprimer la loi de R pour construire un test.

Néanmoins, il est plus simple d'effectuer la **transformée de Fisher** de R qui consiste à calculer :

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

Lorsque n est « grand » ($n > 25$), la variable aléatoire Z est distribuée selon une loi normale de moyenne $\frac{1}{2}\ln((1+\rho_P)/(1-\rho_P))$ et de variance $1/(n-3)$ qui ne dépend pas de ρ_P.

On construit ainsi un test sous l'hypothèse $\mathscr{H}_0 : \rho_P = 0$.

2.3.1.2 Cas non gaussien : coefficient de corrélation de Spearman (facultatif)

Lorsque les variables ne sont pas gaussiennes, le test précédent n'est pas utilisable. On introduit le **coefficient de corrélation de Spearman**.

On considère le n-échantillon $(U_1,\dots,U_n)$ établi en classant dans l'ordre croissant le n-échantillon $(X_1,\dots,X_n)$ de manière à ce que U_i est le rang de la variable X_i après classement (on suppose qu'il n'y a pas d'ex-aequo). Même chose pour $(V_1,\dots,V_n)$ avec $(Y_1,\dots,Y_n)$.

Le coefficient de corrélation de Spearman de X et Y est le coefficient de corrélation de Pearson des rangs U et V :

$$r_S(X,Y) = \frac{\text{cov}(U,V)}{\text{var}(U)\text{var(V)}}$$

L'intérêt de r_S est qu'il est invariant par transformation monotone croissante des variables (une telle transformation ne change pas les rangs).

Exemple 5

Sur une réalisation : si $(x_1, x_2, x_3) = (3,4;2,1;5,3)$, alors $(u_1, u_2, u_3) = (2,1,3)$.

D'une part : $\overline{U} = \overline{V} = (n+1)/2$, car les n-échantillons U_i et V_i sont des listes d'indices obtenues par permutation de $\{1,2,\ldots,n\}$., et $\sum_{k=1}^{n} k = n(n+1)/2$.

D'autre part : $S_U^2 = S_V^2 = (n^2-1)/12$ après développement de $\sum_{k=1}^{n}(k-(n+1)/2))^2$

On en déduit l'expression du coefficient de corrélation de Spearman empirique :

$$\begin{aligned} R_S(X,Y) &= \frac{\sum_{i=1}^{n}(U_i-\overline{V})(V_i-\overline{V})}{\sqrt{\sum_{i=1}^{n}(U_i-\overline{U}^2\sum_{i=1}^{n}(V_i-\overline{V}^2}} \\ &= \frac{\sum_{i=1}^{n} U_i V_i/n-(n+1)^2/4}{(n^2-1)/12} \\ &= 1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n} D_i^2 \end{aligned}$$

avec $D_i = U_i - V_i$, une fois tous les calculs faits.

On voit (en réfléchissant un peu...) que :

— si $R_S = 1$, alors pour tout i, $D_i = 0$: les classements sont identiques (il y a une relation croissante entre X et Y) ;

— si $R_S = -1$, alors pour tout i, les classements sont l'inverse l'un de l'autre (il y a une relation décroissante entre X et Y) ;

— si $R_S = 0$, alors pour tout i, $\sum D_i^2 = n(n^2-1)/6$. C'est le cas qui nous intéresse.
Tout d'abord, $\sum D_i^2 = \sum(U_i - V_i)^2 = 2\sum i^2 - 2\sum U_i V_i = n(n+1)(2n+1)/3 - 2\sum U_i V_i$
Maintenant, si U et V sont indépendants, $E(\sum U_i V_i) = \sum E(U_i)E(V_i) = n(n+1)^2/4$, et donc : $E(\sum D_i^2) = n(n^2-1)/6$ après calcul. Ainsi $R_S = 0$ correspond au cas où les différences entre les rangs sont « statistiquement banales » sous hypothèse $\mathscr{H}_0$.

Pour les « grandes valeurs » de n ($n > 30$), sous hypothèse d'indépendance de X et Y, R_S est distribué selon une loi normale de moyenne nulle et de variance $1/(n-1)$, ce qui permet de construire un test d'hypothèse.

2.3.2 Variables qualitatives

Soient deux variables qualitatives X et Y prenant respectivement q et r modalités notées respectivement x_i ($1 \le i \le q$) et y_j ($1 \le j \le r$). Les observations permettent de remplir le tableau à double-entrée (dit *tableau de contingence*) suivant :

	y_1	y_2	...	y_r	total
x_1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
$\vdots$					
x_q	n_{q1}	n_{q2}		n_{qr}	$n_{q.}$
total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

Sous hypothèse d'indépendance de X et Y, on a : $p(X = x_i, Y = y_j) = p(X = x_i)p(Y = y_j)$. Sous cette hypothèse, l'effectif moyen attendu pour $(X = x_i, Y = y_j)$ est $m_{ij} = Np(X = x_i)p(Y = y_j) \simeq N(n_{i.}/N)(n.j/N) = n_{i.}n.j/N$.

On calcule alors :

$$d^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

qui a des petites valeurs si X et Y sont indépendants. On admet que, sous $\mathscr{H}_0$ (indépendance de X et Y), d^2 est réalisation d'une variable aléatoire D^2 qui converge en loi vers une variable aléatoire de loi du χ^2 à $(q-1)(r-1)$ degrés de liberté ; sous $\mathscr{H}_1$ (non-indépendance), D^2 tend vers l'infini p.s.

Cela nous permet de construire un test unilatéral à droite.

Il s'agit du **test d'indépendance du χ^2** (différent donc, des tests du χ^2 introduits en sections 2.1.2 et 2.2.2.1).

2.4 Limite des tests d'hypothèse à l'heure des données massives

Dans de nombreuses applications pratiques il est « facile » de disposer de grandes quantités de données, et par ailleurs les ressources de calcul disponibles de nos jours permettent de les traiter. On est dans le cas où la taille n des échantillons est très grande. Le problème est que dans ce cas, tout devient significatif. Aucune paire de variables aléatoires ne peut être déclarée comme indépendante. Par exemple, en sections 2.3.1.1 ou 2.3.1.2, la variance de Z ou R_S varie en $1/n$, il faut donc que la corrélation soit à un niveau irréalistement faible pour que l'indépendance ne soit pas rejetée. De la même manière, les ajustements sont toujours rejetés. En effet, le même argument sur la variance tient pour le test de Kolmogorov-Smirnov en section 2.2.2.2. Cette situation est logique : déclarer qu'un phénomène suit une certaine loi est toujours un choix de modélisation forcément imparfait : souvenez-vous que tous les modèles sont faux, mais certains sont utiles (George E.P. Box). Il est donc logique que disposer d'un grand nombre d'observations permette d'invalider ce modèle simplificateur, qui garde sans doute un intérêt par ailleurs. Le problème du p-hacking évoqué à la séance précédente est un autre écueil que l'on peut rencontrer lorsqu'un grand nombre de tests est possible sur ces données massives.

Pour cette raison, nous ne nous poserons pas de problèmes de significativité statistique dans le cours *De l'apprentissage automatique aux IA génératives* au second semestre. La théorie des tests statistiques d'hypothèse fait néanmoins partie de la « caisse à outils » de l'ingénieur, et doit avoir été comprise avant d'envisager d'aller plus loin. Il existe de nombreux cas

pratiques dans lesquels les échantillons ont une taille limitée (parce que les données sont difficiles ou onéreuses à acquérir) : la statistique classique reste alors très utile.

2.5 Exercices

2.5.1 Deux procédés de fabrication

Dans une usine, des tiges d'acier sont produites sur deux chaînes de production distinctes. La longueur des tiges est affectée d'une certaine variabilité.

On prélève six tiges produites sur la première chaîne, leur longueur est mesurée à :

$$12,8; 12,9; 10,9; 11,5; 11,7; 10,5$$

On prélève huit tiges produites sur la seconde chaîne, leur longueur est mesurée à :

$$12,7; 10,5; 11,3; 10,8; 10,7; 11,6; 11,8; 10,9$$

Y a-t-il une différence significative dans la production des deux chaînes? On supposera que la longueur des tiges est distribuée selon une loi normale.

Indication : sur la première chaîne, la moyenne des longueurs des tiges est $11,717$ et leur variance corrigée est de $0,9657$; sur la deuxième chaîne, la moyenne des longueurs des tiges est $11,288$ et leur variance corrigée est de $0,5241$.

2.5.2 Générateur aléatoire

Les générateurs aléatoires disponibles dans les langages de programmation sont en fait pseudo-aléatoires. Ils sont définis à l'aide d'une suite déterministe. Il est donc important de tester les propriétés qu'ils sont censés suivre.

On a généré 20 valeurs entre 0 et 1, présentées dans l'ordre croissant ici pour une meilleure lisibilité :
0,008 ; 0,012 ; 0,027 ; 0,087 ; 0,207 ; 0,327 ; 0,392 ; 0,396 ; 0,433 ; 0,437 ; 0,453 ; 0,475 ; 0,590 ; 0,643 ; 0,645 ; 0,681 ; 0,736 ; 0,737 ; 0,823 ; 0,928.

1) Ces valeurs sont-elles cohérentes avec le fait que le générateur est censé produire des nombres aléatoires répartis de manière uniforme sur $[0,1]$? Vous utiliserez un test de Kolmogorov-Smirnov et utiliserez le graphique ci-après (à compléter).

2) Même question avec un test du χ^2 d'adéquation en séparant les données en répartissant les données dans les intervalles $[0,1/2]$, $[1/2,3/4]$, $[3/4,1]$.

2.5.3 Snack ou non ?

Un exploitant de cinéma a compté si ses clients achetaient des snacks ou pas, selon le type de film qu'ils venaient voir. Le tableau suivant rassemble les résultats :

Type de Film	Snack	Pas de Snack	**Total Ligne**
Action	50	75	**125**
Comédie	125	175	**300**
Famile	90	30	**120**
Horreur	45	10	**55**
Total Colonne	**310**	**290**	**600**

Y a-t-il un lien entre l'achat de snacks et le type de film ?

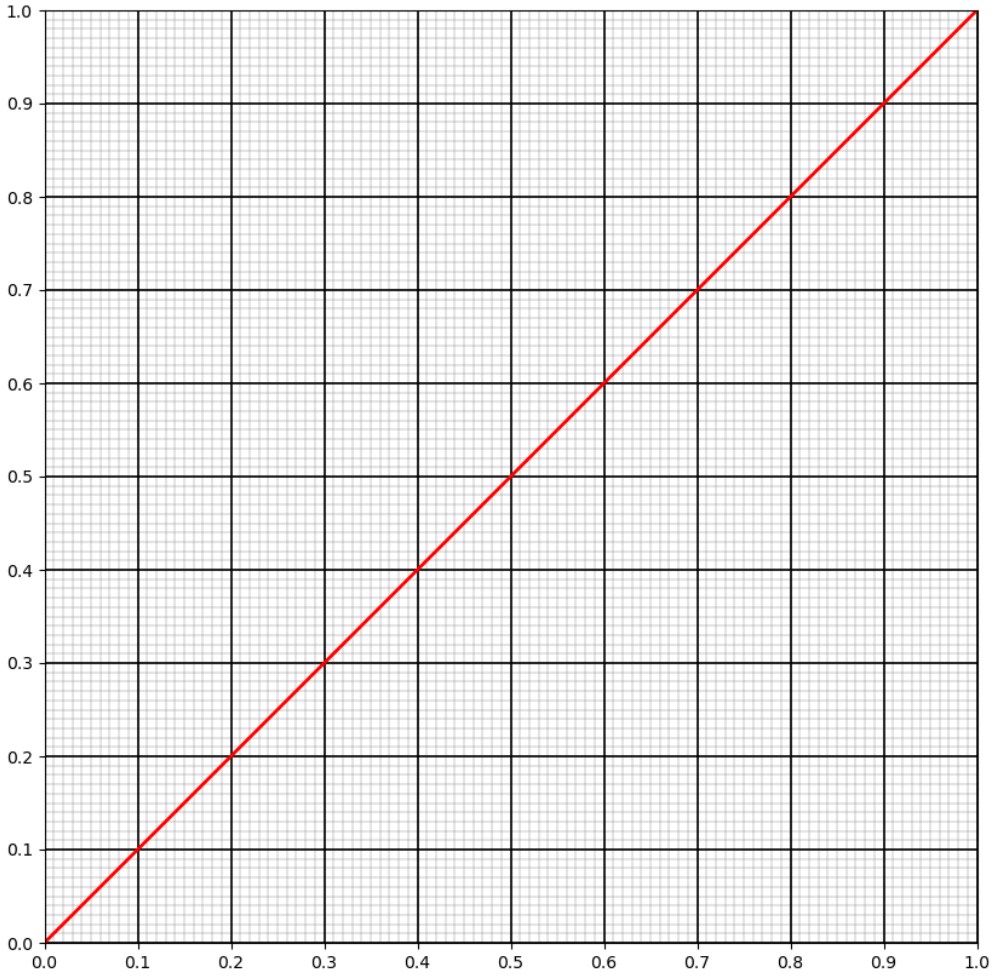


FIGURE 2.2 – *Graphique pour l'exercice 2.*

2.5.4 Test des signes (1)

Selon J. Hemerlijk[2] : *"The sign test is probably the oldest test in existence. It has been applied already in 1710 by John Arbuthnot"*.

On dispose d'un échantillon de paires de valeurs $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_n)$. On forme les différences $D_i = Y_i - X_i$ pour tout $1 \leqslant i \leqslant n$.

Le test des signes consiste à examiner le signe des différences D_i. L'idée est que s'il y a autant de signes + que de signes −, alors il n'y a pas de différence significative entre les X_i et les Y_i. Il s'agit d'un test d'appariement plus général que celui de l'exercice 3 de la séance précédente car il ne suppose pas que les D_i soient distribués normalement. Par ailleurs, il tient uniquement compte de la comparaison entre X_i et Y_i et pas de la *valeur* de la différence, ce qui est intéressant dans les situations où la valeur de la différence des variables aléatoires n'a pas de réelle signification physique.

Notons π la probabilité d'observer $D > 0$.

1) Soit S le nombre de valeurs D_i telles que $D_i > 0$ (signe +). Quelle est la loi de S ?

2) Proposez une hypothèse nulle $\mathscr{H}_0$ traduisant l'absence de différence significative entre X et Y, et déterminez la p-valeur du test bilatéral sur $\mathscr{H}_0$.

Remarque : pour simplifier, on suppose qu'il n'y pas de différences nulles. Les différences nulles sont discutées dans l'exercice suivant, et sont l'objet de l'article de Hemerlijk.

3) Appliquez le test aux données de l'exercice 3 de la séance précédente, rappelées ci-dessous :

Individu	1	2	3	4	5	6	7	8	9	10
Avant	61,2	66,3	80,4	70,5	73,0	78,1	57,9	71,4	74,2	75,3
Après	62,0	64,3	81,5	68,3	72,9	77,8	63,2	70,3	75,1	75,0

Indication : $\sum_{k=0}^{4} \binom{10}{k} = 386$.

2.5.5 Test des signes (2)

Une étude clinique évalue l'efficacité d'une application de méditation sur le niveau de stress perçu. On mesure le score de stress (sur une échelle de 0 à 100) chez $n = 50$ utilisateurs avant et après un mois de pratique.

Les résultats observés pour les différences $D_i = \text{Score}_{\text{après}} - \text{Score}_{\text{avant}}$ sont les suivants :

— 30 utilisateurs ont vu leur score baisser ($D_i < 0$).

— 10 utilisateurs ont vu leur score augmenter ($D_i > 0$).

— 10 utilisateurs n'ont observé aucun changement ($D_i = 0$).

On souhaite tester si l'application a un effet ou pas, à l'aide du test des signes.

1) Les valeurs nulles ($D_i = 0$) posent un problème. Il y a deux manière simple de traiter les zéros : méthode d'exclusion (suppression des observations pour lesquelles la différence est

2. J. Hemerlik. A theorem on the sign test when ties are present. Proceedings Nederl. Akademie van Wetenschappen Series A 55 (1952) p. 322.

nulle) ou méthode de redistribution (répartition équitable entre les signes + et −) des zéros. Laquelle de ces deux approches a sans doute plus tendance à privilégier l'absence d'effet ? (on attend une justification qualitative basée sur le comportement attendu de la puissance du test)

2) En utilisant la méthode d'exclusion ($n' = 40$ utilisateurs), on souhaite tester $\mathscr{H}_0 : \pi = 0,5$ contre l'alternative bilatérale $\mathscr{H}_1 : \pi \neq 0,5$.

On sait que que si X suit une loi binomiale, alors $X' = (X - E(X))/\sigma_X$ converge en loi vers la loi normale centrée réduite.

1. Rappelez l'espérance $E(S)$ et la variance $\mathrm{Var}(S)$ du nombre de signes positifs S sous $\mathscr{H}_0$.
2. Calculez la valeur de la statistique de test centrée réduite $Z = \frac{S - E(S)}{\sqrt{\mathrm{Var}(S)}}$.

3) Au seuil de signification $\alpha = 5\%$, concluez quant à l'efficacité de l'application de méditation. Calculez également la p-valeur approximative associée à ce test. Comparez à la p-valeur que vous auriez obtenu en utilisant la méthode de redistribution.

Chapitre 3

Régression linéaire

La modélisation statistique a pour objectif de formaliser et de quantifier les relations existant entre différentes variables mesurées sur une population d'individus. L'intérêt est, d'une part, la modélisation en elle-même, qui permet de mieux comprendre un phénomène de par les relations établies, et, d'autre part, la capacité de faire des prédictions (des **inférences**) grâce au modèle. Parmi les méthodes statistiques, la **régression linéaire** occupe une place prépondérante pour des raisons historiques et fondamentales. Elle est sans doute le modèle le plus simple d'inférence d'une variable quantitative, et, est, à ce titre, le premier modèle vu dans tout cours d'« apprentissage automatique » (la science de l'« intelligence artificielle »).

3.1 Introduction

L'idée est d'expliquer les variations d'une variable quantitative d'intérêt, appelée **variable expliquée**, **dépendante** ou **cible** (notée Y car il s'agira d'une variable aléatoire), à l'aide d'une ou plusieurs autres variables quantitatives ou qualitatives, appelées **variables explicatives**, **indépendantes** ou **régresseurs**. Le qualificatif « linéaire » signifie que le modèle est construit comme une combinaison linéaire des *paramètres* à estimer. Les relations entre la variable expliquée et les variables explicatives ne sont elles-mêmes pas nécessairement linéaires. Cette hypothèse permet des développements mathématiques élégants et d'une puissance prédictive très intéressante. Dans le cadre de ce cours, on se limitera au cas d'une seule variable explicative, appelé **régression linéaire simple**. Lorsque plusieurs variables explicatives sont considérées, on parle de **régression linéaire multiple**.

Nous nous intéresserons donc à des relations du type : $y = \beta_0 + \beta_1 x$, ou $y = \beta_0 + \beta_1 f(x)$ où f est une fonction de x (se ramène au premier cas en posant $x' = f(x)$), mais pas à des relations comme par exemple $y = \beta_0 e^{\beta_1 x}$ qui n'est pas linéaire en les coefficients β_0 et β_1. Bien entendu, dans ce dernier exemple il est possible de se ramener au cas linéaire par une transformation logarithmique de y.

La figure 3.1 donne un exemple d'observations (x_i, y_i), et un modèle linéaire qui permettrait d'expliquer y en fonction de x. Évidemment, les observations ne suivent pas rigoureuse-

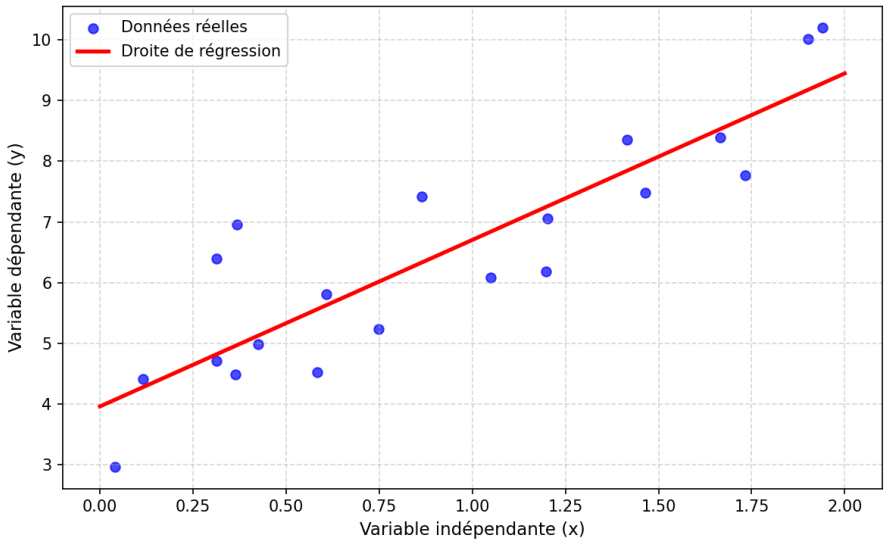


FIGURE 3.1 – *Un exemple de régression linéaire.*

ment le modèle linéaire. Le bon cadre d'étude est celui de la statistique : on va modéliser les écarts au modèle linéaire idéal par des fluctuations aléatoires. L'intérêt d'un tel modèle est double : tout d'abord en terme de modélisation du phénomène si on arrive à le valider sous certaines hypothèses statistiques, et aussi en terme d'inférence car le modèle permettrait de prédire de nouvelles valeurs y correspondant à des x pour lesquelles on n'a pas d'observations.

3.2 La régression linéaire simple

Le cas de la **régression linéaire simple** met en jeu une unique variable explicative x pour expliquer la variable Y. Cette deuxième variable est écrite en majuscule car nous allons la considérer comme étant aléatoire.

3.2.1 Définition mathématique du modèle

On suppose disposer d'un n-échantillon, noté $(x_i, Y_i)_{1 \leqslant i \leqslant n}$: autrement dit, on supposera (ce sera une conséquence des hypothèses **H1** à **H4** de la suite) les variables aléatoires indépendantes entre elles et identiquement distribuées.

Le modèle de régression linéaire simple postule que pour chaque individu (ou observation) i, la relation entre Y_i et x_i s'écrit :

$$\forall\, i \in \{1, \ldots, n\}, \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.1}$$

où :

— x_i est la valeur prise par la variable explicative pour l'observation i. Elle est ici supposée déterministe (fixée par l'expérimentateur ou observée sans erreur) ;
— β_0 est la constante du modèle (ordonnée à l'origine, *intercept*) ;
— β_1 est le paramètre de pente, traduisant l'effet marginal de x sur Y ;
— ε_i désigne la **perturbation aléatoire** ou le terme d'erreur. Cette variable aléatoire capte tout ce qui s'écarte de la relation linéaire exacte : erreurs de mesure, omission d'autres variables, résidu intrinsèquement aléatoire du comportement humain ou physique...
— Y_i est la valeur prise par la variable expliquée pour l'observation i : Y_i est donc considéré comme une variable aléatoire, fonction de ε_i.

Ici, les paramètres du modèle β_0 et β_1 sont fixés mais inconnus. Tout l'enjeu est d'en construire des estimateurs statistiques.

Comme d'habitude, en pratique on observe une réalisation $(x_i, y_i)_{1\leqslant i\leqslant n}$ du n-échantillon, comme dans la figure 3.1, à partir de laquelle on va déterminer des estimations des paramètres du modèle. Si β_0' et β_1' sont ces estimations obtenues à partir des observations, la **droite de régression** a pour équation : $y = \beta_0' + \beta_1' x$, c'est celle qui est tracée sur la figure 3.1.

3.2.2 Modélisation probabiliste

Pour mener à bien l'estimation des paramètres du modèle et l'inférence (prédiction), on formule un ensemble d'hypothèses sur le comportement des perturbations aléatoires ε_i :

H1 (moyenne nulle des perturbations). L'espérance des erreurs est nulle : $E(\varepsilon_i) = 0$ pour tout i. Cela signifie que le modèle linéaire est exact en moyenne : $\forall\, 1 \leqslant i \leqslant n$, $E(Y_i) = \beta_0 + \beta_1 x_i$.

H2 (homoscedasticité). La variance du terme d'erreur est constante pour toutes les observations :

$$\forall\, i \in \{1, \dots, n\}, \quad \mathrm{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 \tag{3.2}$$

H3 (absence d'autocorrélation). Les erreurs associées à deux observations distinctes sont non corrélées. La corrélation entre les termes d'erreurs est $\mathrm{Cov}(\varepsilon_i, \varepsilon_k)/\sigma^2$, Cov désignant la covariance, et l'espérance des ε_i étant nulle. On a :

$$\forall\, 1 \leqslant i \neq j \leqslant n, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \tag{3.3}$$

Les hypothèses **H1**, **H2**, **H3** sont les **hypothèses de Gauss-Markov**.

3.2.3 Estimation par la méthode des moindres carrés ordinaires (MCO)

Le principe des moindres carrés ordinaires (MCO) consiste à rechercher les valeurs des paramètres β_0 et β_1 qui minimisent la somme des carrés des écarts entre les valeurs observées y_i et les valeurs prédites par la **droite de régression** d'équation $\hat{y}_i = \beta_0 + \beta_1 x_i$. Considérer les carrés des écarts (plutôt que leur valeur absolue, ou toute autre fonction des écarts) simplifie les développements mathématiques suivants.

On cherche donc à minimiser en fonction de β_0 et β_1 la fonction suivante :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{3.4}$$

Proposition 3.1 *Les estimateurs des MCO, notés $\hat{\beta}_0$ et $\hat{\beta}_1$, sont donnés par les formules suivantes :*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{3.5}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{3.6}$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{Y} = \frac{1}{n}\sum_{Y=1}^n Y_i$ désignent les moyennes (moyenne empirique pour $\bar{Y}$).

Notons que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont, comme les Y_i, des variables aléatoires.

La prédiction de la i-ème observation selon ce modèle est donnée par $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: il s'agit également d'une variable aléatoire.

Démonstration

La fonction $S(\beta_0, \beta_1)$ est strictement convexe. Ses dérivées partielles par rapport aux deux paramètres s'écrivent :

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) \tag{3.7}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (Y_i - \beta_0 - \beta_1 x_i) \tag{3.8}$$

L'annulation des dérivées partielles donne le système de ce qu'on appelle les **équations normales** qui définissent les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$:

$$\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i = 0 \tag{3.9}$$

$$\sum_{i=1}^n x_i Y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \tag{3.10}$$

Au passage, on voit que considérer les carrés des écarts simplifie les calculs : en dérivant, on aboutit à des relations linéaires en $\hat{\beta}_0$ et $\hat{\beta}_1$ pour les équations normales.

De l'équation (3.9), en divisant par n, on obtient : $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, ce qui démontre la relation pour l'ordonnée à l'origine : $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

En injectant cette expression dans l'équation (3.10), il vient :

$$\begin{aligned}
\sum_{i=1}^n x_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0 \\
\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0 \\
\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} &= \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)
\end{aligned}$$

En utilisant les identités remarquables de la covariance et de la variance empirique, à savoir $\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}$ et $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$, on obtient l'expression finale de $\hat{\beta}_1$. □

Comme $Y_i - \bar{Y} = \beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$, on a également :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{3.11}$$

en développant le numérateur, car $\sum_i (x_i - \bar{x}) = 0$.

Ainsi,

$$\hat{\beta}_0 \;=\; \bar{Y} - \beta_1 \bar{x} - \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x} \;=\; \beta_0 - \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x} + \bar{\varepsilon} \tag{3.12}$$

où $\bar{\varepsilon}$ est la moyenne empirique des ε_i.

Ces deux relations seront utiles dans la section suivante.

Remarque. En pratique, on observe une réalisation du n-échantillon $(x_i, y_i)_{1 \leqslant i \leqslant n}$, qui permet de calculer une réalisation de $\bar{Y}$, $\hat{\beta}_0$, $\hat{\beta}_1$, et la valeur moyenne $\bar{x}$ (qui est déterministe). La relation $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ signifie que la droite de régression passe toujours par le point moyen $(\bar{x}, \bar{y})$ du nuage des observations (x_i, y_i).

3.2.4 Propriétés statistiques des estimateurs simples

Sous les hypothèses **H1** à **H3**, on démontre les propriétés essentielles suivantes :

1. $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des **estimateurs sans biais** de β_0 et β_1 : $E(\hat{\beta}_1) = \beta_1$ et $E(\hat{\beta}_0) = \beta_0$.
2. La variance des estimateurs s'écrit :

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{3.13}$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{3.14}$$

Démonstration

Ce sont des conséquences directes des équations 3.11 et 3.12, de la linéarité de l'espérance, de l'absence d'autocorrélation entre les ε_i, et de $E(\varepsilon_i) = E(\bar{\varepsilon}) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathrm{Var}(\bar{\varepsilon}) = \sigma^2/n$. □

Ces formules indiquent que, à amplitude σ^2 des perturbations aléatoires fixée, la précision des estimateurs augmente avec la taille n de l'échantillon et avec la dispersion des valeurs de la variable explicative x.

Sauf dans des cas pathologiques où $\sum_i (x_i - \bar{x})^2$ s'annule, **ces estimateurs sont convergents** car leur variance tend vers 0 quand $n \to +\infty$ ($\sum_i (x_i - \bar{x})^2 = O(n)$).

Par ailleurs, la covariance de $\hat{\beta}_0$ et $\hat{\beta}_1$ est (après des calculs basés sur la bilinéarité de la covariance et la non-corrélation des aléas) :

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{Cov}\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})\bar{x}\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \bar{\varepsilon}, \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{3.15}$$

Ces deux estimateurs ne sont donc pas indépendants.

Remarque. Les estimateurs donnés par les équations 3.5 et 3.6 sont des **estimateurs linéaires**, dans le sens où ils sont formés comme une combinaison linéaire des données observées Y_i. Le **théorème de Gauss-Markov** (admis) énonce que les estimateurs des MCO sont ceux de variance minimale parmi les estimateurs linéaires sans biais. On dit que cet estimateur est le BLUE (*Best Linear Unbiased Estimator*).

3.2.5 Estimation sans biais de la variance des résidus

Pour exploiter de manière pratique les formules des variances des estimateurs des MCO, $\text{Var}(\hat{\beta}_1)$ et $\text{Var}(\hat{\beta}_0)$, données par les équations 3.13 et 3.14, il est indispensable d'estimer le paramètre de variance inconnu σ^2. On définit pour chaque observation i le résidu empirique $\hat{e}_i$ par :

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{3.16}$$

Attention, le résidu empirique $\hat{e}_i$ est une variable aléatoire, comme Y_i, $\hat{\beta}_0$ et $\hat{\beta}_1$, mais ce n'est pas l'erreur ε_i dans le modèle (théorique) de l'équation 3.1, qui reste inconnue.

La **somme des carrés des résidus** (SCR), égale à $\sum_i \hat{e}_i{}^2$, correspond exactement au minimum de la fonction de coût $S(\hat{\beta}_0, \hat{\beta}_1)$. Bien que la variance théorique de chaque perturbation soit σ^2, l'estimateur naturel de cette variance $\frac{1}{n}\sum_i \hat{e}_i^2$ est biaisé, ce que nous allons démontrer à présent en déterminant l'espérance de $\sum_i \hat{e}_i^2$.

D'une part : $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$; et d'autre part : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, soit $\bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$. Donc :

$$\begin{cases} Y_i - \bar{Y} = \beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \\ \hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x}) \end{cases} \tag{3.17}$$

Puisque $\hat{e}_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})$, on obtient par substitution :

$$\hat{e}_i = (\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \tag{3.18}$$

Élevons cette expression au carré et sommons sur i :

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left[(\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \right]^2 \tag{3.19}$$

En développant l'identité remarquable, on obtient trois termes :

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x}) + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.20}$$

On sait d'après l'équation 3.11 que :

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.21}$$

Ce qui implique que $\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) = (\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n} (x_i - \bar{x})^2$.

En remplaçant cette somme dans le terme central de l'équation 3.20, l'expression se simplifie en :

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.22}$$

Par linéarité de l'espérance :

$$E\left(\sum_{i=1}^{n} \hat{e}_i^2\right) = E\left(\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2\right) - \sum_{i=1}^{n} (x_i - \bar{x})^2 E\left((\hat{\beta}_1 - \beta_1)^2\right) \tag{3.23}$$

Le premier bloc vaut, par définition de l'estimateur sans biais de la variance des ε_i :

$$E\left(\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2\right) = (n-1)\sigma^2 \tag{3.24}$$

Pour le second bloc, puisque par définition de la variance $E\left((\hat{\beta}_1 - \beta_1)^2\right) = \mathrm{Var}(\hat{\beta}_1)$:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 E\left((\hat{\beta}_1 - \beta_1)^2\right) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \mathrm{Var}(\hat{\beta}_1) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sigma^2 \tag{3.25}$$

En combinant les résultats :

$$E\left(\sum_{i=1}^{n} \hat{e}_i^2\right) = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2 \tag{3.26}$$

Proposition 3.2 *L'estimateur sans biais de la variance des perturbations σ^2, noté s^2 (ou $\hat{\sigma}^2$), est défini par :*

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{SCR}{n-2} \tag{3.27}$$

Il s'agit bien d'une variable aléatoire, comme Y_i et $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

En substituant σ^2 inconnu par son estimation s^2 dans les expressions théoriques (équations 3.13 et 3.14), on obtient les estimateurs des écarts-types des coefficients, notés $\hat{\sigma}_{\hat{\beta}_1}$ et $\hat{\sigma}_{\hat{\beta}_0}$.

Remarque. Les résidus $\hat{e}_i$ (dits **résidus ordinaires**) sont de moyenne nulle (car $E(Y_i - \hat{Y}_i) = E(Y_i) - \beta_0 - \beta_1 x_i = 0$), mais de variance exprimée (après calculs) par $\mathrm{Var}(\hat{e}_i) = E(\hat{e}_i^2) = \sigma^2(1 - h_{ii})$ où $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$. Les résidus ordinaires $\hat{e}_i$ ne sont donc pas distribués de manière homoscédastique, contrairement aux termes d'erreur ε_i. On voit qu'une observation éloignée du barycentre de l'ensemble des observations ($x_i - \bar{x}$ grand) donnera lieu à une variance plus grande qu'une observation proche du barycentre.

Les **résidus de Pearson** $\hat{r}_i$ sont définis par :

$$\hat{r}_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}} \tag{3.28}$$

Les résidus de Pearson sont des versions standardisées des résidus ordinaires, globalement distribués entre -2 et 2, de manière à pouvoir les comparer entre eux. Un résidu supérieur à 2 en valeur absolue sera signe d'une observation aberrante.

3.3 Inférence statistique sous hypothèse de normalité

3.3.1 Ajout d'une hypothèse de normalité

Afin de mener à bien l'inférence (construction de tests et d'intervalles de confiance), on adjoint aux hypothèses de Gauss-Markov (**H1** à **H3**) une hypothèse sur la loi des perturbations :

H4 (normalité). Les perturbations ε_i suivent des lois normales indépendantes et identiquement distribuées : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Sous **H1–H4**, les variables aléatoires Y_i sont normales par linéarité (voir remarque à la fin de la section 3.2.4). Les estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$, formés par combinaison linéaire des Y_i, suivent également des lois normales caractérisées par les moyennes et variances établies en section 3.2.4 :

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1,\ \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \tag{3.29}$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0,\ \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \tag{3.30}$$

Comme la variance σ^2 est inconnue et substituée par son estimateur s^2, la « Studentisation »[1] de ces variables aléatoires conduit aux statistiques fondamentales suivantes :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T(n-2) \quad \text{et} \quad \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim T(n-2) \tag{3.31}$$

où $T(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté.

Le justification est que la SCR suit la loi du χ^2 à $n-2$ degrés de liberté. C'est une conséquence admise du théorème de Cochran déjà évoqué dans ce cours.

Remarque. Sous l'hypothèse **H4**, les observations $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ suivent une loi normale de moyenne $\beta_0 + \beta_1 x_i$ et de variance σ^2. La log-vraisemblance des observations est donc :

$$-\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n \left(\frac{Y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2 \tag{3.32}$$

En annulant les dérivées de la log-vraisemblance par rapport à β_0 et β_1, on trouve après calcul que les **estimateurs du maximum de vraisemblance** de ces deux paramètres sont les estimateurs des MCO. De la même manière, on trouve que l'estimateur du maximum de vraisemblance de σ^2 est $\sum_i e_i^2 / n$, dont on a vu qu'il était biaisé (voir proposition 3.2).

1. Cela désigne l'opération consistant à centrer et réduire une variable aléatoire avec une estimation de la moyenne et une estimation de l'écart-type, ce qui conduit, sous hypothèse de normalité, à une loi de Student. Plusieurs exemples de ce type ont été vus dans le cours.

3.3.2 Notion d'intervalle de confiance

Considérons un paramètre déterministe θ (potentiellement multidimensionnel) dont dépend la loi du n-échantillon considéré. On s'intéresse à une certaine quantité d'intérêt $g(\theta) \in \mathbb{R}$, et on cherche un intervalle I, dit **intervalle de confiance** de $g(\theta)$ tel que $P(g(\theta) \in I) = 1-\alpha$. Ici, $g(\theta)$ est un réel déterministe, ce sont les bornes de l'intervalle I qui sont aléatoires et dépendent du n-échantillon. Le paramètre α est le **niveau de risque** de l'intervalle.

Cela signifie qu'avec une probabilité $1-\alpha$, le n-échantillon produit un intervalle I contenant la valeur du paramètre cherché.

Exemple

Considérons un n-échantillon $(X_i)_{1\leqslant i\leqslant n}$ est i.i.d. selon une loi normale de moyenne μ et de variance σ^2. On suppose que σ est connu et on cherche un intervalle de confiance pour μ (avec les notations précédentes, $\theta = g(\theta) = \mu$). On estime μ par la moyenne empirique $\bar{X}$ que l'on sait distribuée selon la loi normale de moyenne μ et de variance σ^2/n. Ainsi $(\bar{X}-\mu)/\sigma$ suit la loi normale $\mathcal{N}(0,1)$. On sait donc que :

$$P\left((\bar{X}-\mu)/\sigma \in [-1{,}96\,;\,-1{,}96]\right) = 0{,}95 \tag{3.33}$$

L'intervalle $[-1{,}96\,;\,1{,}96]$ est déterministe : il s'agit de ce qu'on appelle l'**intervalle de fluctuation** au niveau de risque α de la variable aléatoire $(\bar{X}-\mu)/\sigma$.

On déduit :

$$P\left(\mu \in [\bar{X}-1{,}96\sigma\,,\ \bar{X}+1{,}96\sigma]\right) = 0{,}95 \tag{3.34}$$

Cela signifie que $[\bar{X}-1{,}96\sigma\,,\ \bar{X}+1{,}96\sigma]$ est un intervalle de confiance au risque 5% (ou au **niveau de confiance** 95%) de la vraie moyenne μ.

Remarque. En toute généralité, il n'y a pas unicité de l'intervalle de confiance au risque α. Dans cet exemple, on a utilisé la statistique (dite **pivotale**) $(\bar{X}-\mu)/\sigma$ qui dépend de la loi du n-échantillon, mais dont la loi (la loi normale centrée réduite) ne dépend pas de ce n-échantillon, ce qui permet de déterminer l'intervalle de fluctuation. Trouver un intervalle de fluctuation pour une statistique pivotale puis en déduire l'intervalle de confiance est la démarche standard. La loi normale a la particularité d'être symétrique (c'est le cas aussi de la loi de Student). Dans ce cas, on choisit l'intervalle de fluctuation centré sur 0. De cette manière, l'intervalle de confiance que l'on déduit est équilibré : la probabilité α d'être en dehors de l'intervalle est répartie équitablement à gauche et à droite de l'intervalle.

Pour plus d'information sur les intervalles de confiance, voir l'annexe A du polycopié électronique (sur Arche) de la première partie du cours d'inférence statistique.

3.3.3 Intervalles de confiance des estimateurs et test de significativité des paramètres

Cherchons un intervalle de confiance pour le paramètre β_1 de régression. On sait d'après l'équation 3.31 que :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T(n-2) \tag{3.35}$$

ce qui permet de définir une statistique pivotale de loi de Student à $n-2$ d.d.l.

Ainsi :

$$P\left(\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \in \left[-t^{n-2}_{1-\alpha/2}\,;\, t^{n-2}_{1-\alpha/2}\right]\right) = 1-\alpha \tag{3.36}$$

où $t^{n-2}_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.

L'intervalle $[-t^{n-2}_{1-\alpha/2}\,;\, t^{n-2}_{1-\alpha/2}]$ est l'intervalle de fluctuation au niveau de risque α de la variable aléatoire $\left(\hat{\beta}_1 - \beta_1\right) / \hat{\sigma}_{\hat{\beta}_1}$.

On déduit :

$$P\left(\beta_1 \in \left[\hat{\beta}_1 - t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_1}\,;\, \hat{\beta}_1 + t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_1}\right]\right) = 1-\alpha \tag{3.37}$$

Nous avons donc déterminé un intervalle de confiance pour le paramètre β_1 à un niveau de confiance de $1-\alpha$.

On raisonne de même pour β_0, d'où la proposition suivante.

Proposition 3.3 *L'intervalle de confiance au niveau* $1-\alpha$ *pour le paramètre de pente* β_1 *est donné par :*

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 - t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_1}\,;\, \hat{\beta}_1 + t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_1}\right] \tag{3.38}$$

et l'intervalle de confiance pour la constante β_0 *s'écrit :*

$$IC_{1-\alpha}(\beta_0) = \left[\hat{\beta}_0 - t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_0}\,;\, \hat{\beta}_0 + t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_0}\right] \tag{3.39}$$

où $t^{n-2}_{1-\alpha/2}$ *représente le quantile d'ordre* $1-\alpha/2$ *d'une loi de Student à* $n-2$ *degrés de liberté.*

Les intervalles de confiance sont estimés sur une réalisation des (x_i, Y_i) : concrètement, face à un jeu de données $(x_i, y_i)_{1\leqslant i\leqslant n}$, on calcule les réalisations de $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_{\hat{\beta}_0}$, et $\hat{\sigma}_{\hat{\beta}_1}$, et on en déduit la réalisation de l'intervalle de confiance.

Il est également possible de tester l'hypothèse nulle $\mathscr{H}_0 : \beta_0 = 0$ à partir de l'intervalle de confiance sur β_0 au risque α. Autrement dit, on se demande si la valeur $\hat{\beta}_0$ est possiblement due aux aléas et reste compatible avec une « vraie valeur » β_0 qui serait nulle. D'après l'équation 3.31, cela revient à former la statistique de test $T = \hat{\beta}_0 / \hat{\sigma}_{\hat{\beta}_0}$ et à faire un test bilatéral selon la loi de Student à $n-2$ degrés de liberté : si $|T|$ est supérieure à la valeur critique $t^{n-2}_{1-\alpha/2}$, alors on rejette $\mathscr{H}_0$, et $|T| < t^{n-2}_{1-\alpha/2}$, on ne rejette pas $\mathscr{H}_0$. Cette dernière condition se traduit par $|\hat{\beta}_0| < t^{n-2}_{1-\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_0}$. Autrement dit, cela revient exactement à regarder si la valeur 0 est dans l'intervalle de confiance $IC_{1-\alpha}(\beta_0)$: si ce n'est pas le cas, on rejette $\mathscr{H}_0$. On dit alors que la valeur de l'ordonnée à l'origine du modèle est significative.

Le raisonnement est le même pour $\hat{\beta}_1$.

3.3.4 Intervalle de confiance d'une prédiction

Soit x_0 une nouvelle valeur de la variable explicative. La prédiction ponctuelle fournie par le modèle estimé est $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Deux cas de figure se présentent selon la nature de l'inférence désirée :

1. **Intervalle de confiance de la valeur moyenne $E(\hat{Y}_0)$.** Il s'agit d'encadrer la vraie position de la droite de régression en x_0 (car $E(\hat{Y}_0) = \beta_0 + \beta_1 x_0$, où β_0 et β_1 sont les « vraies » valeurs, inconnues). La variance de cet estimateur $\hat{Y}_0$ vaut $\text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + x_0 \text{Covar}(\hat{\beta}_0, \hat{\beta}_1)$ (rappelons que les deux estimateurs des MCO des paramètres ne sont pas indépendants). On trouve :
$$\text{Var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{3.40}$$
En estimant σ^2 par s^2, l'intervalle de confiance à un niveau de confiance $1-\alpha$ pour un point de la droite de régression est défini par :
$$IP_{1-\alpha}\left(E(\bar{Y}_0)\right) = \left[\hat{Y}_0 \pm t_{1-\alpha/2}^{n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] \tag{3.41}$$
où $t_{1-\alpha/2}^{n-2}$ représente toujours le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.
2. **Intervalle de prédiction d'une observation individuelle Y_0.** Il s'agit d'anticiper la valeur future prise par un nouvel individu de variable explicative x_0. D'une part, $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ par définition du modèle, et d'autre part, la prédiction est donnée par $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. L'erreur de prédiction e_0 est l'écart entre $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (point de la droite de régression) et Y_0 (valeur réelle). On a $e_0 = Y_0 - \hat{Y}_0$, d'où $E(e_0) = \beta_0 + \beta_1 x_0 - (\beta_0 + \beta_1 x_0) = 0$. Par ailleurs Y_0 et $\hat{Y}_0$ sont indépendantes, comme ε_0 d'une part et $\hat{\beta}_0$ et $\hat{\beta}_0$ d'autre part. Donc
$$\text{Var}(e_0) = \text{Var}(Y_0) + \text{Var}(\hat{Y}_0) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{3.42}$$
En estimant σ^2 par s^2, l'intervalle de prédiction à un niveau de confiance $1-\alpha$ pour un individu est défini par :
$$IP_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \pm t_{1-\alpha/2}^{n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right] \tag{3.43}$$

En toute rigueur, un intervalle de prédiction d'une observation n'est pas un intervalle de confiance, puisqu'il concerne une variable aléatoire (Y_0) et pas une valeur déterministe (comme $E(\bar{Y}_0)$).

On voit que l'amplitude de ces intervalles est minimale au centre de gravité des données ($\bar{x}$) et s'accroît de façon hyperbolique à mesure que l'on s'éloigne du point moyen, mettant en évidence la perte de précision du modèle en extrapolation.

Remarque. Lorsque x_0 est « proche » de $\bar{x}$ (on néglige alors $(x_0-\bar{x})^2$ devant $\sum_i (x_i-\bar{x})^2$) et n est grand, l'intervalle de confiance de la prévision devient négligeable et les observations individuelles sont réparties dans une bande de part et d'autre de la droite de régression d'épaisseur $2t_{1-\alpha/2}s$.

3.4 Discussion d'un exemple

En pratique, on observe une réalisation $(x_i, y_i)_{1\leqslant i\leqslant n}$ du n-échantillon (x_i, Y_i), à partir duquel on calcule des estimations numériques des paramètres de la droite de régression, de la variance du bruit, des intervalles de confiance des prévisions... vues comme des réalisations des variables aléatoires introduites précédemment.

La figure 3.2 montre une droite de régression calculée sur des données, et les deux types d'intervalles de confiance.

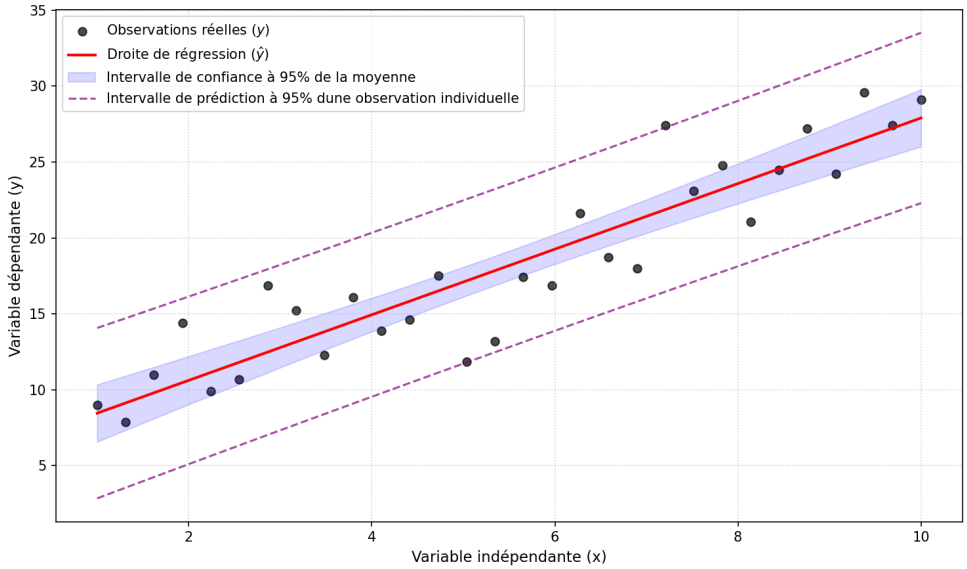


FIGURE 3.2 – *Régression linéaire : intervalles de confiance des points de la droite de régression et intervalles de prédiction d'une observation individuelle.*

L'intervalle de confiance de la moyenne (zone bleue). Il entoure la droite de régression rouge. Il représente l'incertitude liée à l'estimation de la vraie relation moyenne entre x et y. Si l'on répétait l'expérience à l'infini en collectant de nouveaux échantillons de données, la « vraie » droite théorique se trouverait dans cette zone bleue avec une probabilité de 0,95. La zone bleue est légèrement plus resserrée au centre qu'aux extrémités : c'est parce que le modèle est toujours plus précis autour de la moyenne des données observées. Les frontières de cette zone sont des arcs d'hyperbole.

L'intervalle de prédiction des observations (courbes pointillées). Il est beaucoup plus large et englobe la quasi-totalité des points. Il représente l'incertitude liée à la prédiction d'une valeur individuelle future. Si on choisit une nouvelle valeur de x, il y a 95% de chances que l'observation réelle y se situe entre ces deux lignes violettes. Il est plus large que l'intervalle précédent car il doit cumuler deux types d'incertitudes : l'imprécision de la droite elle-même (la zone bleue) et la variabilité naturelle / le bruit aléatoire de chaque individu (les fluctuations des points noirs autour de la droite).

3.5 Qualité de l'ajustement et diagnostics avancés

Une fois les calculs effectués, le statisticien doit valider la qualité globale de l'ajustement (les tests de Student précédents sont une manière de s'assurer que le modèle est statistiquement significatif), et s'assurer que les hypothèses théoriques imposées au modèle (hypothèses **H1** à **H4**) se vérifient empiriquement.

3.5.1 Coefficient de détermination R^2

Le **coefficient de détermination**, noté R^2, est défini comme :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3.44}$$

On démontre aussi (c'est un calcul semblable à celui qu'on fera pour l'analyse de la variance au chapitre suivant) que :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3.45}$$

Le coefficient R^2 mesure donc la proportion de la variance totale de la variable cible expliquée par le régresseur du modèle.

Le coefficient de détermination R^2 est un indicateur compris entre 0 et 1. Plus il est proche de 1, meilleur est l'ajustement global du modèle aux données. Le cas où $R^2 = 1$ correspond à $\forall\, i,\ Y_i = \hat{Y}_i$: le modèle linéaire prédit parfaitement les observations.

Remarque. Dans le cas d'autres modèles de régression, la définition (équation 3.44) n'implique pas la propriété de l'équation 3.45. On peut alors avoir des valeurs de R^2 négatives (pour un modèle très mauvais), mais on a toujours $R^2 \leqslant 1$.

3.5.2 Analyse des résidus et tests de validation

La validation d'un modèle de régression linéaire passe par l'étude des propriétés des résidus observés $e_i = Y_i - \hat{Y}_i$. Les diagnostics suivants seront vus dans des cours spécialisés.

— **Diagnostic d'homoscedasticité**. On trace graphiquement les résidus en fonction des valeurs prédites $\hat{Y}$. Si le nuage de points présente une forme d'entonnoir, l'hypothèse de variance constante est violée. Différents tests statistiques sont disponibles.

— **Diagnostic de normalité**. L'analyse visuelle s'appuie sur un graphique Quantile-Quantile QQ-plot. Les points doivent s'aligner sur une droite. Les tests statistiques de Shapiro-Wilk ou de Jarque-Bera permettent de tester la normalité des résidus.
— **Diagnostic d'indépendance**. En présence de données chronologiques (l'indice i désigne alors le temps), les erreurs peuvent être corrélées dans le temps. On utilise le test de Durbin-Watson pour détecter une éventuelle autocorrélation d'ordre 1.

3.6 Exercices

3.6.1 Un exercice d'application directe avec des calculs simples

Un professeur souhaite analyser la relation entre le nombre d'heures de révision de ses étudiants et la note qu'ils ont obtenue (sur 10) lors d'un test. Il a sélectionné un échantillon représentatif de 5 étudiants.

Voici les données récoltées :

Étudiant	**Heures de révision** (x_i)	**Note obtenue** (y_i)
A	1	3
B	2	4
C	3	6
D	4	8
E	5	9

Travail à faire

1. **Droite de régression :** Déterminez l'équation de la droite de régression linéaire de y en x, sous la forme $y = ax + b$.
2. **Intervalle de confiance de la pente :** Sachant que pour un niveau de confiance de 95% et avec 3 degrés de liberté ($n-2$), la valeur critique de la loi de Student est de $t = 3{,}182$, calculez l'intervalle de confiance à 95% pour la pente a. La relation entre heures de révision et note est-elle significative ?
3. **Prédiction :** Si un étudiant révise pendant 3,5 heures, quelle note peut-il espérer obtenir selon ce modèle ?
4. **Intervalle de prédiction :** Calculez l'intervalle de prédiction à 95% pour la note exacte de cet étudiant ayant révisé 3,5 heures.

3.6.2 Étude sur ordinateur de ce jeu de données réduit

Le carnet Jupyter sur Arche donne le code permettant de faire une régression linéaire et de valider le modèle par les statistiques du cours et la visualisation de différents graphiques. Il est basé sur la bibliothèque Python `statsmodels`.

Interprétez les sorties du carnet.

3.6.3 Une étude à l'aide de la bibliothèque statsmodels

Voir sur Arche.

Chapitre 4

Analyse de la variance

4.1 Introduction

L'**analyse de la variance** (*analysis of variance*, **ANOVA**), développée par le statisticien britannique Ronald A. Fisher dans les années 1920, est une méthode statistique permettant de comparer simultanément les moyennes de plusieurs populations (ou groupes). Cette procédure peut être vue comme une extension à $k > 2$ groupes de la comparaison de moyennes entre deux groupes à l'aide d'un test de Student (voir section 1.1 du second chapitre sur les tests : comparaison de deux échantillons gaussiens indépendants).

Exemple 1

Connaissant les notes d'examen des dix groupes de TD du cours *Inférence statistique*, peut-on dire si les notes des groupes sont équivalentes ? Plus précisément, les différences observées entre les moyennes des différents groupes sont-elles compatibles avec les fluctuations aléatoires d'échanillonnage, ou bien ces différences sont-elles statistiquement significatives ?

Dans cet exemple 1, on cherche à mettre en évidence l'influence d'un **facteur de variabilité** (le groupe de TD) sur la note obtenue. C'est ce qui est appelé **analyse de variance à un facteur**. Des facteurs supplémentaires peuvent être envisagés : homme / femme, langue vivante étudiée, etc. Ce sera l'objet de l'**analyse de variance multifactorielle**. Les facteurs sont des variables qualitatives qui peuvent prendre un nombre fini k de **modalités** (on parle aussi de **niveaux**). Ici, le « nom » du groupe de TD (A1, A2...) est une modalité, et le facteur de variabilité a dix modalités.

Pour que les conclusions tirées d'une ANOVA soient statistiquement valides, les données doivent respecter trois hypothèses fondamentales :

1. **indépendance des observations :** les individus de chaque modalité doivent être choisis de manière aléatoire et indépendante.

2. **normalité des distributions :** dans chacune des k modalités, la variable étudiée doit suivre la loi normale $\mathcal{N}(\mu_i, \sigma_i^2)$, pour $1 \leqslant i \leqslant k$.
3. **homoscédasticité (homogénéité des variances) :** la variance de la variable étudiée doit être la même pour chacune des k modalités étudiées ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$).

Dans la suite, ces hypothèses seront supposées vérifiées.

Dans l'exemple 1, cela veut dire que les élèves sont répartis aléatoirement dans les groupes, indépendamment l'un de l'autre (on ne fait pas des groupes de niveau, ou des groupes par filière d'origine). Cela semble raisonnable de supposer une répartition normale des notes. Supposer les variances toutes égales est une hypothèse plus forte qui mériterait d'être testée. Cela veut dire qu'on suppose que l'effet du facteur « groupe de TD » ne se traduit que par un décalage des moyennes.

4.2 ANOVA à un facteur

On considère un facteur explicatif qualitatif (appelée **variable indépendante**) présentant k modalités. On cherche à savoir si ce facteur influence une variable quantitative (appelée **variable dépendante**).

L'objectif est de tester l'hypothèse $\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

Il serait possible d'effectuer des tests de Student d'égalité deux-à-deux : l'inconvénient est que cela représente un grand nombre de tests et que les risques d'erreur de type 1 s'« ajoutent ». Par exemple, si $k = 10$, on est amené à faire $k \times (k-1)/2 = 45$ tests de Student. Si le risque est $\alpha = 0,05$, dans 5% des cas le test de Student rejette l'hypothèse d'égalité de deux moyennes « par hasard ». Donc dans une proportion $1-(1-\alpha)^{(k(k-1)/2}$, au moins une des égalités est rejetée par hasard, consuisant à rejeter l'hypothèse $\mathcal{H}_0$ [1]. Lorsque $k = 10$, cette proportion atteint 90% ! La stratégie des tests multiples n'est donc pas envisageable. Il faut un test « global ».

4.2.1 Modèle statistique

Soit y_{ij} la j-ème observation de la i-ème modalité, considérée comme une variable aléatoire. On suppose que la i-ème modalité comprend n_i observations (donc $1 \leqslant j \leqslant n_i$ pour $1 \leqslant i \leqslant k$), de manière à ce que le nombre total d'observations soit $N = \sum_{i=1}^{k} n_i$. Le modèle s'écrit de la manière suivante :

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

où :

— μ représente la moyenne générale de la population globale ;
— α_i représente l'effet spécifique de la modalité i, en supposant $\sum_{i=1}^{k} n_i \alpha_i = 0$ (l'effet moyen est nul) ;

1. En effet, la probabilité que toutes les hypothèses $\mu_i = \mu_j$ ($1 \leqslant i < j \leqslant k$) soient rejetées par hasard est $(1-\alpha)^{k(k-1)/2}$. Notons qu'à l'ordre 1 pour α petit, $1-(1-\alpha)^{k(k-1)/2} = k(k-1)/2\,\alpha + o(\alpha)$: en ce sens, les risques d'erreur de type 1 s'ajoutent.

— ε_{ij} représente l'erreur aléatoire (résidu), supposée suivre la loi normale de moyenne nulle et d'écart-type constant σ. Les ε_{ij} sont supposés indépendants.

Ainsi, les observations y_{ij} respectent bien les trois hypothèses fondamentales de l'ANOVA.

4.2.2 Décomposition de la variance

La somme des carrés totale T est décomposée en deux parties : la variation due au facteur étudié A et la variation résiduelle R. Notons

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2, \quad A = \sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2, \quad R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$$

où $\bar{y}_{..}$ est la moyenne générale et $\bar{y}_{i.}$ est la moyenne du groupe i.

On a :

$$T = A + R$$

D'où vient cette relation ? On calcule :

$$\begin{aligned} T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{..})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..}) \end{aligned}$$

Par ailleurs :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$$

et :

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..}) &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}\bar{y}_{i.} - y_{ij}\bar{y}_{..} - \bar{y}_{i.}^2 + \bar{y}_{i.}\bar{y}_{..}\right) \\ &= \sum_{i=1}^{k} n_i\bar{y}_{i.}^2 - n\bar{y}_{..}^2 - \sum_{i=1}^{k} n_i\bar{y}_{i.}^2 + \sum_{i=1}^{k} n_i\bar{y}_{i.}\bar{y}_{..} \\ &= 0 \end{aligned}$$

D'où la relation cherchée.

En notant $S_T^2 = T/N$, $S_A^2 = A/N$, et $S_R^2 = R/N$, S_T^2 est la variance totale, S_A^2 est la variance expliquée (par le facteur étudié), et S_R^2 est la variance résiduelle.

On a donc la relation :

$$S_T^2 = S_A^2 + S_R^2$$

4.2.3 Analyse des variances

Remarque. La justification des lois peut être omise.

4.2.3.1 Loi de la variance résiduelle S_R^2

Comme les observations de la modalité i sont indépendantes et suivent la loi normale $\mathcal{N}(\mu_i, \sigma^2)$, la statistique

$$S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

est telle que $n_i S_i^2/\sigma^2$ suit la loi du χ^2 à $n_i - 1$ degrés de liberté[2].

Les observations étant indépendantes et la variance résiduelle vérifiant :

$$S_R^2 = \frac{1}{N} \sum_{i=1}^{k} n_i S_i^2$$

S_R^2 est telle que NS_R^2/σ^2 suit la loi du χ^2 à $\sum_{i=1}^{k}(n_i - 1) = N - k$ degrés de liberté[3].

4.2.3.2 Loi de la variance expliquée S_A^2

Sous l'hypothèse $\mathcal{H}_0$, $\bar{y}_{i.}$ suit la loi $\mathcal{N}(\mu, \sigma^2/n_i)$, dont $\sqrt{n_i}(\bar{y}_{i.} - \mu)/\sigma$ suit la loi $\mathcal{N}(0,1)$. Ainsi, $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \mu)^2/\sigma^2$ suit la loi du χ^2 à k d.d.l.

On ne connaît pas la moyenne μ, mais :

$$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \mu)^2 = \sum_{i=1}^{k} n_i((\bar{y}_{i.} - \bar{y}_{..}) - (\mu - \bar{y}_{..}))^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + N(\bar{y}_{..} - \mu)^2$$

Comme $N(\bar{y}_{..} - \mu)^2/\sigma^2$ suit la loi du χ^2 à un degré de liberté, le théorème de Cochran (voir la première partie du cours) permet de conclure que NS_A^2/σ^2 suit la loi du χ^2 à $k-1$ degrés de liberté.

4.2.3.3 Loi de la variance totale S_T^2

Sous l'hypothèse $\mathcal{H}_0$, NS_T^2/σ^2 suit la loi du χ^2 à $N-1$ degrés de liberté[2].

Notons que les lois de NS_A^2/σ^2 et NS_T^2/σ^2 données ici ne sont valables que sous hypothèse $\mathcal{H}_0$ d'égalité des moyennes (μ_i), contrairement à celle de NS_R^2/σ^2.

4.2.4 Test de Fisher

L'hypothèse nulle $(\mathcal{H}_0)$ postule qu'il n'y a aucune différence entre les moyennes des groupes :

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Sous l'hypothèse alternative $(\mathcal{H}_1)$, au moins deux moyennes sont différentes.

2. Rappelons la propriété (conséquence du théorème de Cochran) : si $(X_1, \ldots, X_n)$ est un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ et S^2 est la variance empirique (non corrigée), alors nS^2/σ^2 suit la loi du χ^2 à $n-1$ d.d.l.

3. La somme de deux variables aléatoires indépendantes suivant les lois du χ^2 à a et b d.d.l. suit la loi du χ^2 à $a+b$ d.d.l.

On admet que les statistiques NS_A^2/σ^2 et NS_R^2/σ^2 sont indépendantes (c'est encore une conséquence du théorème de Cochran).

La statistique de test est donnée par le rapport :

$$F = \frac{S_A^2/(k-1)}{S_R^2/(N-k)} = \frac{A/(k-1)}{R/(N-k)}$$

Sous $\mathscr{H}_0$, cette statistique suit la loi de Fisher-Snedecor à $(k-1)$ et $(N-k)$ degrés de liberté (voir annexe A.4). Pour un risque α donné, la règle de décision est donc : si la valeur F calculée est supérieure à la valeur critique F_α, on rejette $\mathscr{H}_0$.

Bien que le nom *Analysis of variance* suggère une analyse des variances, l'objectif principal de l'ANOVA est d'inférer une propriété sur les moyennes. Le principe central consiste à comparer la variance estimée *entre* les différents groupes (variabilité expliquée par le facteur étudié) à la variance estimée *à l'intérieur* des groupes (variabilité résiduelle ou bruit).

4.2.5 Le tableau récapitulatif de l'ANOVA

Les résultats d'une ANOVA sont classiquement présentés sous la forme du tableau synthétique suivant.

Source de variation	**Somme des carrés**	**ddl**	**Carré moyen**	$\boldsymbol{F}_{\text{obs}}$
Facteur (Inter-groupes)	A	$k-1$	$\frac{A}{k-1}$	$\frac{A/(k-1)}{R/(N-k)}$
Résiduelle (Intra-groupe)	R	$N-k$	$\frac{R}{N-k}$	
Total	T	$N-1$		

Les logiciels ou bibliothèques Python donnent une colonne supplémentaire avec la p-valeur de F_{obs} qui permet de prendre une décision selon le risque α assumé.

4.2.6 Comparaisons multiples (tests *post-hoc*)

Le test de l'ANOVA est dit *omnibus* (« pour tous » en latin) : il teste les moyennes dans leur ensemble. Lorsqu'il est significatif (rejet de $\mathscr{H}_0$), il indique qu'au moins une moyenne diffère des autres, mais il ne précise pas laquelle. Pour identifier où se situent ces différences, on réalise des tests de comparaisons multiples des moyennes (c'est-à-dire, d'après la première séance sur les tests, des tests de Student). On partle de test *post-hoc* (« après cela » en latin) : ce sont des tests réalisés après analyse de la variance. On peut être amené à réaliser un grand nombre de tests deux-à-deux : si le facteur a 10 modalités, il faudra procéder à $m = 10 \times 9/2 = 45$ tests.

On a déjà discuté du problème méthodologique posé par un grand nombre de tests sur le même ensemble d'observations (cf. première séance sur les tests et remarque en section 4.2) : multiplier les tests augmente le risque de première espèce (conclure à tort à une différence). On peut être amené à utiliser la correction de Bonferroni, qui consiste à diviser le risque α par le nombre de comparaisons m. D'autres méthodologies de tests *post-hoc* existent.

4.2.7 Cas pratique

Imaginons une étude où nous testons l'efficacité de 3 méthodes d'apprentissage (A, B, C) sur les scores d'un test (sur 20) obtenu par 15 étudiants (5 par méthode).

Données :

— Méthode A, notes obtenues : 10, 12, 11, 14, 13 → $\bar{y}_A = 12$

— Méthode B, notes obtenues : 16, 15, 18, 17, 19 → $\bar{y}_B = 17$

— Méthode C, notes obtenues : 8, 9, 11, 10, 7 → $\bar{y}_C = 9$

On calcule alors :

Moyenne générale $\bar{y}_{..} = 12.67$

Calcul des sommes des carrés :

$$A = 5 \times [(12-12.67)^2 + (17-12.67)^2 + (9-12.67)^2] = 163.33$$
$$R = \sum_{i,j} (\bar{y}_{ij} - \bar{y}_{i.})^2 = (10-12)^2 + \cdots + (7-9)^2 = 30.00$$

Tableau de l'ANOVA :

Source	**SC**	**ddl**	**Carrés moyens**	F_{obs}
Méthode	163.33	$3-1=2$	$163.33/2 = 81.67$	$81.67/2.5 = 32.67$
Erreur	30.00	$15-3=12$	$30.00/12 = 2.50$	
Total	193.33	14		

Conclusion. La valeur critique de F pour $\alpha = 0.05$ avec 2 et 12 degrés de liberté est d'environ 3.89 (voir la table en annexe A.4). Puisque $F_{\text{obs}} = 32.67 > 3.89$, on rejette l'hypothèse nulle. La méthode d'apprentissage a un effet statistiquement très significatif sur le score des étudiants.

4.3 ANOVA à deux facteurs

Lorsqu'on étudie l'influence simultanée de deux variables qualitatives (facteurs A et B) sur une variable quantitative X, on utilise une ANOVA à deux facteurs.

Exemple 2

On souhaite tester l'influence du groupe de TD et du choix de Langue Vivante 2 sur la note au cours *Inférence statistique*.

Notons p et q le nombre de modalités respectives des facteurs A et B, et n_{ij} ($1 \le i \le p$ et $1 \le j \le q$) le nombre d'observations de X pour le couple (i, j) de modalités de A et B. On suppose disposer d'un modèle **équilibré** : $n_{ij} = r \ge 1$. À tout couple de modalités correspond le même nombre d'observations. L'intérêt de cette hypothèse est de simplifier les calculs dans la suite. De cette manière, le nombre total d'observations est $N = pqr$.

Dans l'exemple 2, cela signifie qu'il y a le même nombre r d'élèves de chaque langue vivante dans chaque groupe de TD.

4.3.1 Effet principal et interaction

Ce modèle permet de tester trois hypothèses :

1. L'effet principal du facteur A est-il significatif?
2. L'effet principal du facteur B est-il significatif?
3. Y a-t-il une interaction entre les facteurs A et B?

Exemple 3

Nous voulons mesurer la croissance d'une plante (variable dépendante) en fonction de deux facteurs prenant deux modalités chacun :

— ensoleillement (facteur A : Faible ou Fort)
— arrosage (facteur B : Hebdomadaire ou Quotidien)

L'ANOVA va tester trois hypothèses distinctes :

1. l'effet principal du facteur A : l'ensoleillement a-t-il un impact significatif sur la croissance, indépendamment de l'arrosage?
2. l'effet principal du facteur B : l'arrosage a-t-il un impact significatif sur la croissance, indépendamment de l'ensoleillement?
3. l'effet d'interaction : l'effet de l'ensoleillement sur la croissance dépend-il de la fréquence d'arrosage? C'est souvent l'effet le plus intéressant.

4.3.2 Le modèle statistique avec interaction

Soit un facteur A à p niveaux et un facteur B à q niveaux, avec r répétitions par combinaison du facteur A et du facteur B. On supposera $r > 1$. Les données recueillies sont les y_{ijk} pour ($1 \leq i \leq p$, $1 \leq j \leq q$, et $1 \leq k \leq r$).

Explicitons cette notations dans le cas de l'exemple 2 : pour tout groupe de TD i il y a $r = 4$ étudiants qui suivent la LV2 j. Dans le groupe de TD 2, les notes des 4 étudiants qui suivent la LV2 3 sont : $y_{2,3,1}, y_{2,3,2}, y_{2,3,3}, y_{2,3,4}$.

On suppose que y_{ijk} suit la loi $\mathcal{N}(\mu_{ij}, \sigma^2)$, pour $1 \leq k \leq r$.

On peut alors décomposer :

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

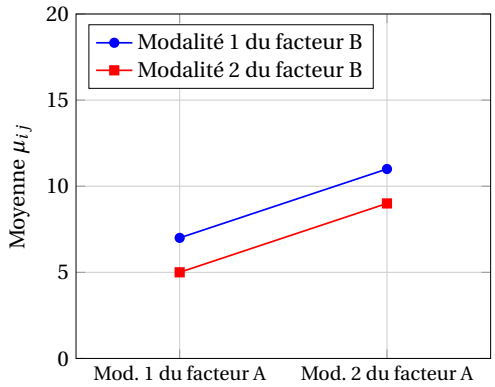
où μ est l'effet moyen, α_i l'effet principal de la modalité i du facteur A (on suppose $\sum_i \alpha_i = 0$), β_j l'effet principal de la modalité j du facteur B (on suppose $\sum_j \beta_j = 0$), γ_{ij} l'effet d'interaction causé par la combinaison du niveau i de A et j de B (on suppose $\sum_i \gamma_{ij} = 0$ et $\sum_j \gamma_{ij} = 0$), de manière à ce que les effets représentent un écart à la moyenne μ. Par ailleurs, ε_{ijk} est l'effet résiduel : les ε_{ijk} sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$.

Ainsi :

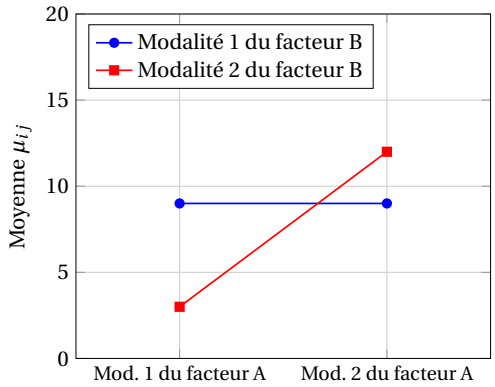
$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Dans le cas de deux facteurs à deux modalités chacun, le **graphique d'interaction** peut prendre les allures de l'exemple 4 suivant.

Exemple 4

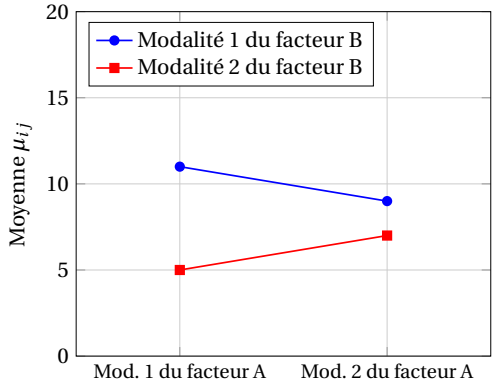


$\mu = 8$, $\alpha = (-2;2)$, $\beta = (1,-1)$, $\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$

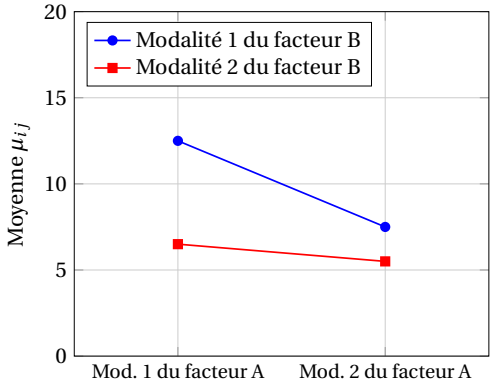


$\mu = 8$, $\alpha = (-2;2)$, $\beta = (1,-1)$, $\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{2} & \mathbf{-2} \\ \mathbf{-2} & \mathbf{2} \end{pmatrix}$

En l'absence d'interaction (ci-dessus) les modalités de B sont translatées l'une par rapport à l'autre.



$\mu = 8$, $\boldsymbol{\alpha = (0;0)}$, $\beta = (2,-2)$, $\gamma = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$



$\mu = 8$, $\boldsymbol{\alpha = (1.5;-1.5)}$, $\beta = (2,-2)$, $\gamma = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

En présence d'interaction (ci-dessus), difficile de juger l'effet de A et B à partir des graphiques.

Remarque. Il est tout à fait possible d'avoir interaction sans effet principal des deux facteurs. Par exemple, imaginons une étude sur l'efficacité de deux types de médicaments (facteur A) sur deux groupes de patients (facteur B : Hommes et Femmes).

— Médicament 1 : très efficace sur les hommes (+10 points de santé), mais réduit la santé des femmes (-10 points).

— Médicament 2 : réduit la santé des hommes (-10 points), mais très efficace sur la santé des femmes (+10 points).

Effet principal du facteur A : si nous faisons la moyenne globale, les +10 et -10 s'annulent. La moyenne du Médicament 1 est 0, celle du Médicament 2 est 0, l'effet principal est non significatif.

Effet principal du facteur B : la moyenne des hommes est 0, celle des femmes est 0, l'effet principal est non significatif.

L'interaction est, pour sa part, hautement significative. L'efficacité du médicament dépend totalement de si vous êtes un homme ou une femme.

Il est souvent erroné (ou du moins incomplet) de conclure qu'un facteur n'a « pas d'effet » simplement parce que son effet principal n'est pas significatif. Dans le cas d'une interaction significative, le facteur a bien un impact, mais cet impact est conditionnel à la modalité.

Bien entendu, la difficulté est qu'en pratique, la situation dans laquelle on se trouve est difficile à identifier à cause des fluctuations aléatoires sur les observations.

Selon des calculs semblables à ceux de la section 4.2.2 pour l'ANOVA à un facteur, la décomposition de la variance devient dans le cas de deux facteurs :

$$S_T^2 = S_A^2 + S_B^2 + S_I^2 + S_R^2$$

avec $NS_A^2 = qr\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$, $NS_B^2 = pr\sum_i(\bar{y}_{.j.} - \bar{y}_{...})^2$, $NS_I^2 = r\sum_i\sum_j(\bar{y}_{ij.} - \bar{y}_{.j.} - \bar{y}_{i..} + \bar{y}_{...})^2$, et enfin $NS_R^2 = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij.})^2$.

Si on suppose que les moyennes « annulent » les termes aléatoires ε_{ijk} dans l'expression des y_{ijk}, on obtient, grâce aux hypothèses $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$, les approximations suivantes : $\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2 \simeq \sum_i \alpha_i^2$, $\sum_i(\bar{y}_{.j.} - \bar{y}_{...})^2 \simeq \sum_j \beta_j^2$, $\sum_i\sum_j(\bar{y}_{ij.} - \bar{y}_{.j.} - \bar{y}_{i..} + \bar{y}_{...})^2 \simeq \sum_i\sum_j \gamma_{ij}^2$, et enfin $\sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij.})^2 \simeq N\sigma^2$. Ceci explique (en divisant ces quantités par $N = pqr$), les dénominations suivantes : les quantités S_T^2, S_A^2, S_B^2, S_I^2, et S_R^2 sont appelées variances, respectivement, totale, expliquée par le facteur A, expliquée par le facteur B, expliquée par l'interaction entre A et B, et résiduelle.

On calcule les sommes de carrés NS_A^2, NS_B^2, NS_I^2 à l'aide d'un tableau des moyennes et des moyennes marginales analogue au tableau suivant (pour $p = 2$ et $q = 3$). La valeur de NS_R^2 ne peut, pour sa part, pas se calculer juste avec ce tableau.

Facteur A \ Facteur B	B_1	B_2	B_3	*Moy. factorielle A*
A_1	$\bar{y}_{11.}$	$\bar{y}_{12.}$	$\bar{y}_{13.}$	$\bar{y}_{1..}$
A_2	$\bar{y}_{21.}$	$\bar{y}_{22.}$	$\bar{y}_{23.}$	$\bar{y}_{2..}$
Moy. factorielle B	$\bar{y}_{..1}$	$\bar{y}_{..2}$	$\bar{y}_{..3}$	$\bar{y}_{...}$

De la même manière que pour l'ANOVA à un facteur, sous l'hypothèse $\mathscr{H}_0$ selon laquelle tous les μ_{ij} sont égaux, on forme des statistiques qui suivent des lois du χ^2 indépendantes, dont les rapports normalisés par les nombres de d.d.l. suivent des lois de Fisher-Snedecor. On remplit alors le tableau suivant.

Source de variation	**Somme des carrés**	**ddl**	**Carré moyen**	F_{obs}
Facteur A	$A = NS_A^2$	$p-1$	$\frac{A}{p-1}$	$\frac{A/(p-1)}{R/(pq(r-1))}$
Facteur B	$B = NS_B^2$	$q-1$	$\frac{B}{q-1}$	$\frac{B/(q-1)}{R/(pq(r-1))}$
Interaction A/B	$I = NS_I^2$	$(p-1)(q-1)$	$\frac{I}{(p-1)(q-1)}$	$\frac{I/((p-1)(q-1))}{R/(pq(r-1))}$
Résiduelle	$R = NS_R^2$	$pq(r-1)$	$\frac{R}{pq(r-1)}$	
Total	$T = NS_T^2$	$pqr-1$		

Les logiciels ou les bibliothèques Python donnent une colonne supplémentaire avec les p-valeurs des statistiques de test calculées sur les observations.

Il n'est pas facile de comprendre intuitivement ce que signifie « rejeter » ou « accepter » l'hypothèse $\mathscr{H}_0$ (tous les μ_{ij} sont égaux à μ) selon un des trois tests du tableau précédent. Notons que sous $\mathscr{H}_0$, $\forall i, j, \alpha_i+\beta_j+\gamma_{ij} = 0$ par définition. Ainsi (en sommant sur j) $\forall i, \alpha_i = 0$, (en sommant sur i) $\forall j, \beta_j = 0$, et donc $\forall i, j, \gamma_{ij} = 0$.

C'est la raison pour laquelle certains auteurs parlent de trois hypothèses $\mathscr{H}_0$ distinctes pour F_A (qui serait $\mathscr{H}_0 : \forall i, \alpha_i = 0$), F_B ($\mathscr{H}_0 : \forall j, \beta_j = 0$), et F_{AB} ($\mathscr{H}_0 : \forall i, j, \gamma_{ij} = 0$).

L'implication importante de cette remarque est que rejeter $\mathscr{H}_0$ avec F_{AB} par exemple veut dire que les γ_{ij} ne sont pas tous nuls (ce qui est plus facile à interpréter que « il existe une différence entre deux moyennes μ_{ij} et $\mu_{i'j'}$ »). Certes, accepter $\mathscr{H}_0$ avec F_{AB} permet d'accepter la proposition équivalente $\forall i, j, \alpha_i = \beta_j = \gamma_{ij} = 0$, mais cela ne veut pas dire que $\mathscr{H}_0$ est vraie. Dans la statistique de test de la troisième ligne, la seule information mesurée par $I = NS_I^2$ porte sur les interactions, donc il est raisonnable de se contenter d'accepter $\forall i, j, \gamma_{ij} = 0$ et de conclure seulement à l'absence d'interaction.

À un niveau de risque α fixé, les tests à effectuer sont les suivants :

— l'interaction n'est pas significative si $\frac{I/((p-1)(q-1))}{R/(pq(r-1))}$ est inférieure au quantile $1-\alpha$ de la loi de Fisher-Snedecor à $(p-1)(q-1)$ et $pq(r-1)$ degrés de liberté (on accepte $\forall i, j, \gamma_{ij} = 0$) ;

— le facteur A n'est pas significatif si $\frac{A/(p-1)}{R/(pq(r-1))}$ est inférieure au quantile $1-\alpha$ de la loi de Fisher-Snedecor à $p-1$ et $pq(r-1)$ degrés de liberté (on accepte $\forall i, \alpha_i = 0$) ;

— le facteur B n'est pas significatif si $\frac{B/(q-1)}{R/(pq(r-1))}$ est inférieure au quantile $1-\alpha$ de la loi de Fisher-Snedecor à $q-1$ et $pq(r-1)$ degrés de liberté (on accepte $\forall i, \beta_i = 0$) ;

Autrement dit, l'interaction, le facteur A, ou le facteur B ne sont pas significatifs si l'hypothèse H_0 reste plausible face à des valeurs, respectivement, des sommes des carrés expliqués par l'interaction (I), expliqués par A (A), ou expliqués par B (B) qui ne sont pas assez grandes par rapport aux carrés expliqués par les résidus (R).

D'après la discussion à la fin de l'exemple 4, ces tests sont à faire dans cet ordre. Si l'interaction est significative, l'effet des facteurs A et B devient difficile à interpréter. Il faudrait alors comparer l'effet de A pour chaque niveau de B par exemple (tests d'effet simple). Si l'interaction n'est pas significative, on teste l'effet des facteurs A et B.

Si les trois effets ne sont pas significatifs (aucun des trois tests ne permet de rejeter $\mathscr{H}_0$), alors la population est homogène et distribuée selon la loi $\mathcal{N}(\mu, \sigma^2)$.

Remarque. Si $r = 1$ (une seule observation par couple de modalités), on a un problème car dire que R suit la loi du χ^2 à $pq(r-1) = 0$ d.d.l. n'a pas de sens. Il faut alors adapter l'ANOVA (voir un livre de statistique).

4.3.3 Cas pratique

On souhaite étudier l'influence de deux facteurs sur les notes obtenues par des étudiants à un examen de statistiques :

1. *le Facteur A (Méthode) :* Classique (A_1) vs Interactive (A_2) ;
2. *le Facteur B (Moment) :* Matin (B_1) vs Après-midi (B_2).

Pour chaque combinaison de facteurs, on a mesuré les notes de 3 étudiants (répétitions). Les résultats sont consignés dans le tableau ci-dessous :

Méthode \ Moment	*Matin* (B_1)	*Après-midi* (B_2)
Classique (A_1)	11 ; 12 ; 13	8 ; 9 ; 10
Interactive (A_2)	13 ; 14 ; 15	15 ; 16 ; 17

Pour l'ANOVA à deux facteurs, la première étape consiste à calculer les moyennes par cellule (combinaisons), les moyennes marginales (par facteur) et la moyenne générale. Notons $Y_{ij} = \bar{y}_{ij.}$ la moyenne des observations de modalités A_i et B_j.

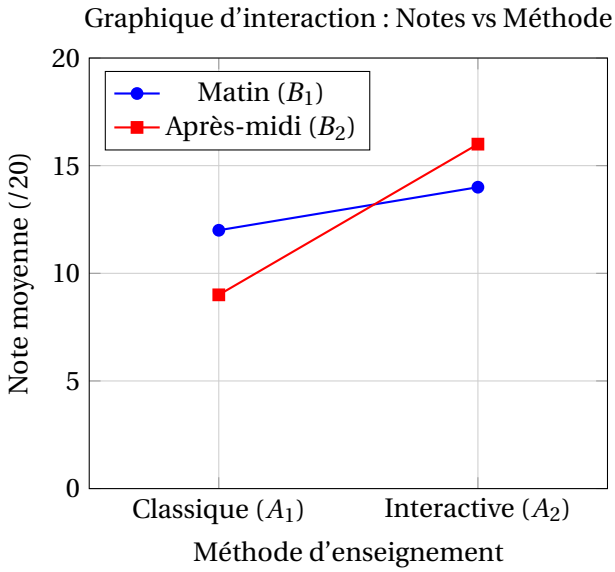
Méthode \ Moment	*Matin* (B_1)	*Après-midi* (B_2)	*Moy. Factorielle A*
Classique (A_1)	12,0	9,0	$\bar{Y}_{1.} =$ **10,5**
Interactive (A_2)	14,0	16,0	$\bar{Y}_{2.} =$ **15,0**
Moy. Factorielle B	$\bar{Y}_{.1} =$ **13,0**	$\bar{Y}_{.2} =$ **12,5**	$\bar{Y}_{..} = 12,75$

On trace le graphique d'interaction suivant, et on observe que :

— Pour la méthode *Classique*, la note diminue de 3 points l'après-midi.

— Pour la méthode *Interactive*, la note augmente de 2 points l'après-midi.

Les segments ne sont pas parallèles et se croisent. Cela suggère visuellement une forte interaction entre la méthode d'enseignement et le moment de la journée.



Voici le tableau de l'ANOVA rempli avec les calculs des carrés moyens et des rapports de Fisher-Snedecor F_{obs}.

Source	**Somme des carrés**	**ddl**	**Carré moyen**	F_{obs}
Facteur A (Méthode)	80,67	$2-1=1$	80,67	**40,34**
Facteur B (Moment)	0,67	$2-1=1$	0,67	**0,34**
Interaction (A × B)	24,00	$(2-1)(2-1)=1$	24,00	**12,00**
Résiduelle	16,00	$12-4=8$	2,00	
Total	121,34	11		

Interprétation statistique ($\alpha = 5\%$). La valeur critique donnée est $F_{\text{crit}}(1,8) = 5,32$. Il se trouve que, dans cet exemple, la valeur critique est la même pour les trois tests car les degrés de liberté sont les mêmes. Cela n'a bien entendu aucune raison d'être le cas en général.

1. *Interaction :* $F_{\text{obs}} = 12,00 > 5,32$. L'interaction est significative.
2. *Effet Méthode :* $F_{\text{obs}} = 40,34 > 5,32$. L'effet est très significatif. La méthode interactive est globalement supérieure à la méthode classique (moyenne de 15,0 pour la première, et 10,5 pour la seconde).
3. *Effet Moment :* $F_{\text{obs}} = 0,34 < 5,32$. L'effet n'est pas significatif. Pris isolément, le moment de la journée n'influence pas la note moyenne (moyenne de 13,0 le matin, et 12,5 l'après-midi).

Même si le facteur Moment n'a pas une influence globale significative, comme l'effet de l'interaction entre Moment et Méthode est significative, l'étude montre que l'efficacité d'une méthode dépend sans doute du moment où elle est appliquée.

— Pour les cours du *matin*, les deux méthodes sont relativement proches, bien que la méthode interactive soit légèrement devant.
— Pour les cours de l'*après-midi*, la méthode classique provoque une chute des résultats (fatigue/passivité), tandis que la méthode interactive stimule les étudiants et améliore leurs performances.

La préconisation à la suite de l'étude pourrait être de privilégier la méthode interactive pour les créneaux de fin de journée. Néanmoins, la puissance du test est sans doute assez faible, les échantillons étant de taille limitée.

4.4 Exercices

4.4.1 Rendement agricole

Un agronome souhaite comparer l'effet de trois types d'engrais (A, B et C) sur le rendement du blé. Il a sélectionné 9 parcelles de terrain identiques et a appliqué chaque engrais sur 3 parcelles choisies au hasard. Les rendements obtenus (en quintaux par hectare) sont les suivants :

— **Engrais A :** 20 ; 22 ; 24
— **Engrais B :** 28 ; 30 ; 32
— **Engrais C :** 18 ; 20 ; 16

1) Analyse préliminaire

Calculer la moyenne de chaque groupe ($\bar{x}_A, \bar{x}_B, \bar{x}_C$) ainsi que la moyenne générale de l'échantillon ($\bar{x}_{\text{total}}$).

2) Sommes des Carrés

Calculer la Somme des Carrés Inter-groupes (SSB) et la Somme des Carrés Intra-groupe (SSW).
Rappel : $SSB = \sum n_i(\bar{x}_i - \bar{x}_{\text{total}})^2$ et $SSW = \sum\sum(y_{ij} - \bar{x}_i)^2$

3) Tableau de l'ANOVA

Compléter le tableau d'analyse de la variance ci-dessous :

Source de variation	**Somme des carrés**	**ddl**	**Carré moyen**	**F observé**
Inter-groupes (Facteur)	.....	.....	.....	.....
Intra-groupe (Résidus)	.....	.....	.....	
Total	.....	.....		

4) Prise de décision

On donne la valeur critique de Fisher-Snedecor au seuil $\alpha = 5\%$ pour les degrés de liberté correspondants : $F_{\text{critique}}(2,6) = 5,14$.

1. Formuler l'hypothèse nulle $\mathscr{H}_0$ et l'hypothèse alternative $\mathscr{H}_1$.
2. Comparer F_{obs} et F_{critique}.
3. Peut-on conclure que le type d'engrais a un impact significatif sur le rendement ?

4.4.2 Café et sommeil

On étudie l'influence de deux facteurs sur le score de concentration (sur 100) d'un groupe d'étudiants :

— **Facteur A (Sommeil) :** Court (A_1) ou Long (A_2).
— **Facteur B (Café) :** Sans café (B_1) ou Avec café (B_2).

On dispose de 2 répétitions par condition (8 observations au total). Les scores sont les suivants :

Sommeil \ Café	**Sans Café** (B_1)	**Avec Café** (B_2)
Court (A_1)	40 ; 50	70 ; 80
Long (A_2)	80 ; 90	85 ; 95

1) Moyennes des cellules

Calculer la moyenne de score pour chacune des 4 combinaisons (cellules), ainsi que les moyennes marginales pour les facteurs A et B, et enfin la moyenne générale G.

2) Analyse de la Variance

On donne les sommes des carrés suivantes :

— $SS_{\text{Sommeil}} = 1512,5$
— $SS_{\text{Café}} = 612,5$
— $SS_{\text{Résiduelle}} = 200,0$
— $SS_{\text{Total}} = 2637,5$

Question : par déduction, calculez la somme des carrés de l'interaction ($SS_{A\times B}$).

3. Tableau de l'ANOVA

Compléter le tableau d'analyse de la variance ci-dessous :

Source	**Somme des carrés**	**ddl**	**Carré moyen**	**F observé**
Sommeil (A)	1512,5	...	...	...
Café (B)	612,5	...	...	...
Interaction (AxB)	...	...	...	...
Résiduelle	200,0	...	...	
Total	2637,5	7		

4) Interprétation

Au seuil $\alpha = 5\%$, la valeur critique de Fisher-Snedecor $F(1,4)$ est de **7,71**. Quelles sont vos conclusions concernant l'effet du sommeil, l'effet du café et l'existence d'une interaction ?

4.4.3 ANOVA en Python

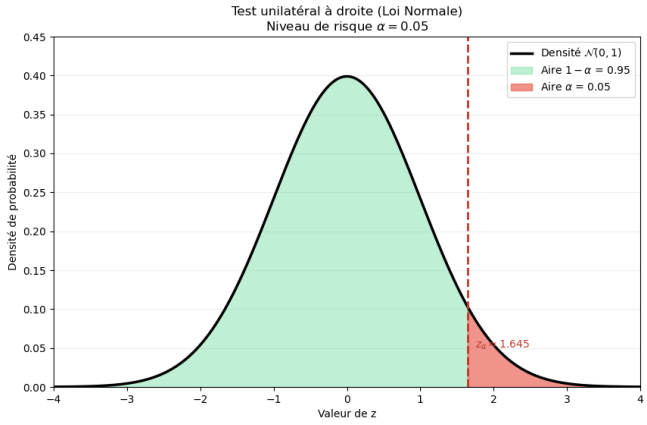
Voir sur Arche.

Chapitre A

Tables statistiques

Voici quelques tables statistiques qui seront utiles dans les exercices.

A.1 Loi normale centrée réduite

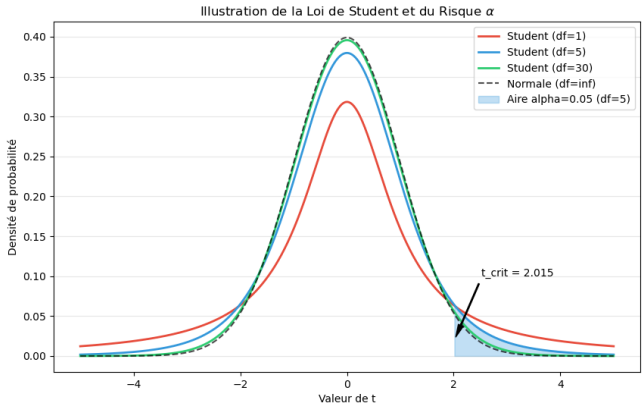


La table suivante représente des valeurs de la fonction de répartition de la loi normale centrée réduite. On peut y lire, pour Z suivant une loi normale centrée réduite, la valeur de $P(Z < z)$ pour des valeurs de z entre 0 et 3,49. On déduit les valeurs pour $z < 0$ par symétrie.

Exemple d'utilisation : on lit dans la table que $P(Z < 1,24) \simeq 0,8925$ (à la croisée de la ligne 1.2 et de la colonne 0.04). Si on cherche 0,95 dans la table, on trouve z_0 tel que $P(Z < z_0) = 0,95$: z_0 entre 1,64 et 1,65, donc par interpolation $z_0 \simeq 1,645$. Si on cherche $P(Z < -0,2)$, on écrit $P(Z < -0,2) = P(Z > 0,2) = 1 - P(Z < 0,2)$ et on lit $P(Z < 0,2)$ dans la table.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

A.2 Loi de Student



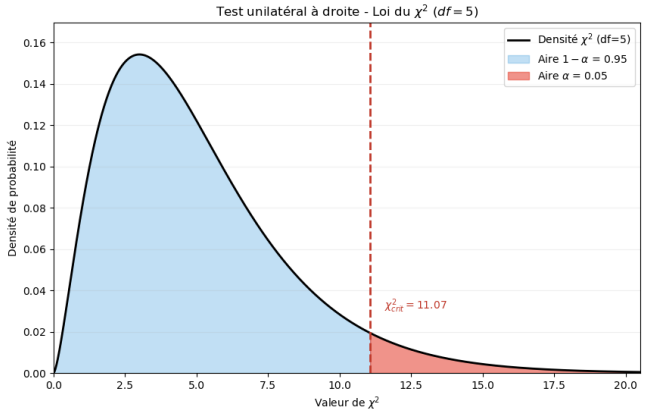
La table suivante donne, pour T suivant une loi de Student à n degrés de liberté, la valeur de t_{crit} tel que $P(T > t_{\text{crit}}) = \alpha$. Ces valeurs sont utiles pour un test à droite. On trouve les valeurs pour un test bilatéral en divisant le risque par 2 et en cherchant t_{crit} tel que $P(T > t_{\text{crit}}) = \alpha/2$.

Exemple d'utilisation : on lit dans la table que, si T est une variable aléatoire qui suit une loi de Student à 5 degrés de liberté, $P(T > 2{,}015) = 0{,}05$.

df / α	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787

26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

A.3 Loi du χ^2



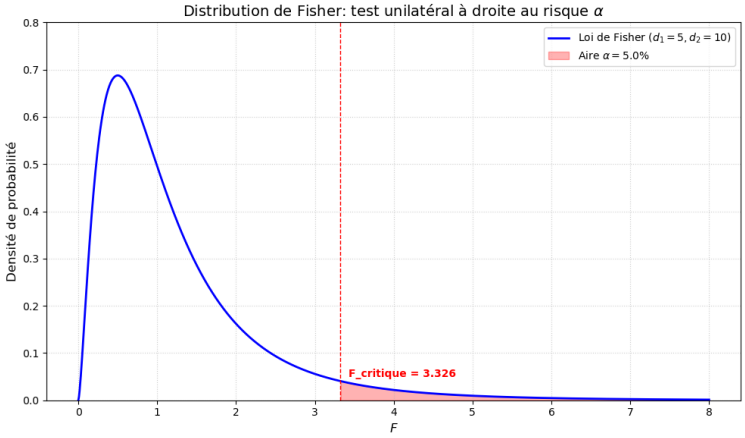
La table suivante donne, pour X suivant une loi du χ^2 à n degrés de liberté, la valeur de χ^2_{crit} tel que $P(X > \chi^2_{\text{crit}}) = \alpha$. Cette valeur est utile pour un test à droite. Pour un test à gauche : si on cherche la valeur de χ^2_{crit} tel que $P(X < \chi^2_{\text{crit}}) = \alpha$, on la lit dans la table en utilisant $P(X > \chi^2_{\text{crit}}) = 1 - \alpha$.

Exemple d'utilisation : on lit dans la table que, si X est une variable aléatoire qui suit une loi du χ^2 à 5 degrés de liberté, $P(X > 11,07) \simeq 0,05$. De même $P(X < 0,55) \simeq 0,01$.

df / α	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005	0.001
1	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95	26.12
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.59

11	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32	36.12
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27	39.25
17	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.31
21	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	49.73
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
26	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64	55.48
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70

A.4 Loi de Fisher-Snedecor



Si U_1 et U_2 sont de variables aléatoires indépendantes suivant des lois du χ^2 à d_1 et d_2 degrés de liberté,

$$F = \frac{U_1/d_1}{U_2/d_2}$$

est une variable aléatoire qui suit une **loi de Fisher-Snedecor** à d_1 et d_2 degrés de liberté (on dit aussi loi de Fisher, ou loi de Snedecor). C'est une loi classique dont on connaît un certain nombre de propriétés. Voir par exemple https://fr.wikipedia.org/wiki/Loi_de_Fisher.

La table suivante donne, pour une variable aléatoire F distribuée selon une loi de *Fisher-Snedecor* (aussi appelée loi de Fisher) à d_1 et d_2 degrés de liberté, la valeur de F_{crit} tel que $P(X > F_{\text{crit}}) = 0,05$. Cette valeur est utile pour un test à droite au risque $\alpha = 5\%$ illustré par la figure ci-dessus.

$d_2 \backslash d_1$	1	2	3	4	5	6	7
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103

$d_2 \backslash d_1$	8	9	10	20	30	50	100
1	238.883	240.543	241.882	248.013	250.095	251.774	253.041
2	19.371	19.385	19.396	19.446	19.462	19.476	19.486
3	8.845	8.812	8.786	8.660	8.617	8.581	8.554
4	6.041	5.999	5.964	5.803	5.746	5.699	5.664
5	4.818	4.772	4.735	4.558	4.496	4.444	4.405
6	4.147	4.099	4.060	3.874	3.808	3.754	3.712
7	3.726	3.677	3.637	3.445	3.376	3.319	3.275
8	3.438	3.388	3.347	3.150	3.079	3.020	2.975
9	3.230	3.179	3.137	2.936	2.864	2.803	2.756
10	3.072	3.020	2.978	2.774	2.700	2.637	2.588
20	2.447	2.393	2.348	2.124	2.039	1.966	1.907
30	2.266	2.211	2.165	1.932	1.841	1.761	1.695
50	2.130	2.073	2.026	1.784	1.687	1.599	1.525
100	2.032	1.975	1.927	1.676	1.573	1.477	1.392