UFR IM²AG

UNIVERSITÉ Grenoble Alpes

# Polycopié pour les UE *Statistique Inférentielle* et *Tests*.

Vincent Brault

# Introduction

Ce cours repose principalement sur des notes de cours de Gérard Biau (Université Pierre et Marie Curie), Anatoli Juditsky (Université Grenoble Alpes), Jean-François Le Gall (Université Paris-Saclay) et Marie-Anne Poursat (Université Paris-Saclay).

Par défaut, les corrections des exercices ne sont pas fournies : soit votre raisonnement est juste du début jusqu'à la fin et, dans ce cas, votre résultat est bon ; soit vous hésitez sur l'une des justifications et même si le résultat est bon, il ne vaut rien. En revanche, lorsque des astuces existent, elles sont mises en fin de chapitre.

Enfin, merci aux étudiants qui me font des retours sur mes coquilles ou fautes d'inattention.

Pour les lecteurices qui souhaiteraient plus d'exercices ou une vision différente de ce qui est présenté dans ce manuscrit, voici quelques références :

- *Mathématiques et statistique pour les sciences de la nature* de Biau et al. (2010) pour des exercices appliqués (Chapitres 5 et 6 principalement).
- *Probabilités, analyse des données et Statistique* de Saporta (2006) pour des exercices théoriques (Chapitre 13 et 14 principalement).
- *Statistique inférentielle : cours et exercices corrigés* de Fourdrinier (2002) pour des exercices théoriques (Chapitres 1, 4, 5 et 6 principalement).
- Pour des exercices en R, il y a les livres où François Husson est co-auteur [1] où vous retrouverez aussi des jeux de données sur son site. En particulier :
  - *Statistique avec R, 3ème édition revue et augmentée* de Cornillon et al. (2012).
  - *R pour la statistique et la science des données* de Husson et al. (2018).
  - *Statistiques générales pour utilisateurs. 2 - Exercices et corrigés* de Husson et Pagès (2013).

1. Retrouvez la liste sur sa page web `https://husson.github.io/books.html`

# Planning

Le planning mis ici est à titre indicatif et sera régulièrement mis à jour en fonction de l'avancée.

| Cours | Date | Estimation |
|---|---|---|
| Stat Inf 1 | 26-09-2025 | Cours statistiques descriptives et modélisation. |
| Stat Inf 2 | 01-10-2025 | TD/TP statistiques descriptives et modélisation. |
| Stat Inf 3 | 03-10-2025 | Cours sur les estimateurs. |
| Stat Inf 4 | 08-10-2025 | TD/TP sur les estimateurs. |
| Stat Inf 5 | 10-10-2025 | Cours sur les propriétés d'un estimateur. |
| Test 1 | 13-10-2025 | Cours sur l'introduction aux tests. |
| Stat Inf 6 | 15-10-2025 | TD/TP sur les propriétés d'un estimateur. |
| Stat Inf 7 | 22-10-2025 | TD/TP sur les intervalles de confiance et l'introduction aux tests. |
| Stat Inf 8 | 24-10-2025 | **Examen Stat Inf**. |
| Test 2 | 05-11-2025 | TD/TP |
| Test 3 | 07-11-2025 | Cours |
| Test 4 | 12-11-2025 | TD/TP |
| Test 5 | 14-11-2025 | Cours |
| Test 6 | 19-11-2025 | TD/TP |
| Test 7 | 26-11-2025 | TD/TP |
| Test 8 | 28-11-2025 | **Examen Test** |

# Table des matières

# Première partie

# UE Statistique inférentielle

# Chapitre 1

# Statistique descriptive

*"Quels que soient les progrès des connaissances humaines, il y aura toujours place pour l'ignorance et par suite pour le hasard et la probabilité."*

Émile Borel

## 1.1 Avant-propos

Le but du chapitre de ***Statistique Descriptive*** est de comprendre les fondements de la ***Statistique***. La statistique est une branche des mathématiques qui étudie les statistiques c'est-à-dire les résultats de calculs statistiques réalisés à partir des jeux de données.

Il faut faire attention à différencier le domaine des ***probabilités*** où nous nous donnons un modèle dans lequel nous effectuons des calculs et la ***statistique*** où nous observons des données et nous tentons de les expliquer en faisant éventuellement appel à un modèle probabiliste.

**Attention au piège**

En tant que branche des mathématiques, la statistique est une science rigoureuse qui impose un certain nombre de règles et d'obligations. Il n'est donc pas possible de faire tout ce que nous souhaitons juste parce que ce serait *joli* et que *cela montrerait mieux que nous avons raison.*

Nous verrons que la statistique est souvent mal utilisée ce qui la décrédibilise parfois aux yeux des personnes qui n'ont pas eu de formations poussées en statistique (comme la citation de Winston Churchill en témoigne). Le rôle des statisticien·ne·s est de s'assurer que les conclusions obtenues à partir de jeux de données proviennent d'études rigoureuses.

Pour conclure cette introduction, nous vous invitons à faire vôtre le serment d'Hippocrate du *Data Scientist*[1] :

1. **Intégrité scientifique et rigueur** : J'exploiterai les données avec toute la rigueur requise et en conformité avec les meilleurs standards de ma profession.
2. **Transparence** : J'informerai de manière compréhensible et précise toutes les parties prenantes sur les finalités, les modalités et les implications potentielles de mon utilisation des données.
3. **Equité** : Je veillerai à toujours m'assurer que des individus ou des groupes ne soient pas discriminés par rapport à des critères illégaux ou illégitimes, de manière directe ou indirecte, sur la base de mes travaux sur les données.
4. **Respect** : J'exercerai mon activité professionnelle en respectant la vie privée et la dignité des personnes dans toutes leurs dimensions.
5. **Responsabilité et indépendance** : J'assumerai mes responsabilités en cas de manquement ou de conflit d'intérêt et je donnerai l'alerte si des actes illégaux liés à des données sont constatés.

1. un complément peut être obtenu sur la page suivante : `https://hippocrate.tech/`

## 1.2 Rappels/Pré-requis

Ce chapitre est essentiellement un rappel de notions généralement vues en lycée ou dans les cours de statistique [2] (comme dans la formation STID par exemple).

## 1.3 Généralité sur la démarche scientifique

Une étude statistique a pour but de répondre à une **question** dans un domaine d'application particulier à partir d'un **jeu de données**.

Une étude statistique se déroule en plusieurs étapes :

1. La réflexion sur le **protocole** à suivre pour le recueil des données (plan d'expériences, plan de sondage, élaboration d'un questionnaire ...) ;
2. Le **recueil** et le **codage** des données (fichier informatique, base de données) ;
3. L'**exploration** des données (statistique descriptive, analyse des données, fouille de données, ...), sans chercher à les modéliser ;
4. Éventuellement, **prétraitement des données** (recodage, agrégation, transformations, création de nouvelles données...) ;
5. Si les données sont issues d'un échantillon, **modélisation statistique** : statistique inférentielle, appel à un modèle probabiliste ;
6. **Prévision** et/ou **prise de décision** (réponse à la question initiale).

Chacune de ces étapes est importante pour que les résultats de l'étude répondent correctement à la question posée initialement.


**La (mauvaise) utilisation de la statistique à travers les âges**

Lors de la pandémie du virus appelé *covid19*, une molécule, appelée *hydroxychloroquine*, a été mise en avant médiatiquement : essentiellement deux camps se sont opposés ; l'un disant que ce médicament soignait la maladie et l'autre qu'il était inefficace voire toxique. Nous allons reprendre deux études présentées par chacun des camps pour montrer que le protocole ne permet pas de conclure ni dans un sens, ni dans l'autre.

La première étude est celle faite par Gautret et al. (2020) sur l'effet jugé positif de la molécule ; elle était appelée dans les médias comme *la première étude du professeur Raoult*. Dans cette étude, plusieurs patients ont été suivis durant 6 jours dans plusieurs établissements : l'Institut Hospitalier Universitaire Infection Méditerranée à Marseille où ils ont pris un traitement chloroquine+azythromicine et un groupe n'ayant eu aucun traitement provenant de différents établissements (Nice, Avignon et Besançon). Les différents reproches peuvent être regroupés par thématiques :

- Le premier problème fut l'évaluation de l'état des patients. En effet, dans chacun des groupes, les patients ont été testés tous les jours et, pour un même patient, il arrivait que les résultats changent entre positif et négatif chaque jour. Les tests n'étaient donc pas fiables à 100%.
- Le deuxième reproche fut le faible nombre de personnes impliquées dans l'essai clinique (à savoir 26 patients traités) et la diversité des profils. En effet, comme il y avait quasiment unicité des profils, il n'était pas possible de différencier clairement l'effet du traitement par rapport aux autres co-variables potentielles (comme l'âge). Au moment où ce polycopié est rédigé, une estimation est de 2,3% de morts. Ceci veut dire que sur 26 individus, on peut s'attendre à n'avoir que 0,6 mort ce qui est en dessous d'un mort. En conclusion, s'il n'y a pas de mort dans l'étude, ceci peut être indépendant du traitement.
- Le plus gros reproche fut l'exclusion de 6 des 26 patients pour *non-suivis*. Ceci fut problématique car si 2 patients ont choisi d'arrêter d'eux-mêmes (donc nous ne pouvons critiquer ce choix), pour les 4 autres, ils ont été sortis car 3 sont partis en soin intensif et 1 est décédé. Or, écarter les patients qui ne répondent pas positivement au traitement ne peut qu'augmenter le taux de réussite du reste de la cohorte.

2. En particulier, ce polycopié est une synthèse de celui distribué aux étudiants de STID (voir Brault (2021b) disponible à l'adresse `https://www-ljk.imag.fr/membres/Vincent.Brault/Cours/Cours_1A.pdf`) et celui donné aux étudiants du M1 de Mathématiques Générales (voir Brault (2021a) disponible à l'adresse `https://www-ljk.imag.fr/membres/Vincent.Brault/Cours/CoursStatistique.pdf`).

- Le quatrième reproche fut également le suivi du groupe *placebo* (c'est-à-dire le groupe qui n'avait pas pris le traitement) car il a été fait de manière plus légère (les charges virales n'ont pas été faites quotidiennement par exemple) dans des CHU différents et par des équipes différentes ; il existe donc un biais non négligeable.

La deuxième étude, dite *du Lancet* du nom du journal qui la publia, regroupait cette fois 96 032 patients atteints du covid19 et traités avec de la chloroquine ou de l'hydroxychloroquine et révélait un important sur-risque de mortalité. Après avoir été étudiées de plus près, il s'est finalement avéré que les données étaient certainement fausses et/ou corrigées arbitrairement. Nous mettons ici quelques éléments qui ont permis de mettre en évidence le problème :

- Le taux de mortalité était très supérieur à tout ce qui avait été enregistré jusqu'alors. Même si la conclusion de l'étude était la sur-mortalité du traitement, des conclusions aussi flagrantes auraient mérité une meilleure critique de la part des auteurs.
- Le nombre de morts attribués pour l'Australie était supérieur (73 patients morts dans 5 hôpitaux) que toutes les données remontées par le pays entier (68 morts en tout). De plus, un hôpital fut surpris de se retrouver dans l'étude alors qu'il n'avait jamais transmis ses données.
- Le descriptif des patients (la prise en charge autre que le traitement par exemple) était très flou. Il n'était pas possible de savoir si des co-variables auraient expliqué certaines morts.
- Les doses données ont également surpris puisque certains patients d'Amérique auraient eu des doses de 600mg alors qu'aucun hôpital n'avait donné de doses supérieures à 500mg.

Après toute cette polémique, l'article fut finalement retiré car la société qui avait fourni les données n'était pas capable de prouver leur exactitude. Nous renvoyons vers le site de Libération les statisticien·ne· intéressé·e·s : `https://www.liberation.fr/checknews/2020/06/02/pourquoi-l-etude-du-lancet-sur-l-hydroxychloroquine-est-elle-sous-le-feu-des-critiques_1789844`

Nous voyons bien par ces deux exemples contraires l'importance d'un protocole correctement établi. L'une de ces deux affirmations est peut-être juste, toutefois, sans une étude rigoureuse, aucune des deux n'est acceptée par la totalité de la communauté scientifique.

**Attention au piège**

Il arrive parfois que les statisticien·ne·s ne soient associé·e·s qu'à partir de l'étape 3. Ceci est souvent trop tard car les données récoltées (parfois des plans à plusieurs millions d'euros) ne permettent pas de répondre à la question posée.

La réflexion sur le protocole doit également être l'occasion de considérer les erreurs potentielles. Par exemple, si nous récupérons des données de manière manuscrite, il y a un risque d'erreur humaine ou de mauvaise lecture de ce qui est marqué. A l'opposé, mettre des garde fous sur des recueils virtuels peut empêcher d'avoir certains résultats. Les étapes d'exploration et de prétraitement des données sont également importantes pour repérer les erreurs potentielles.


**La (mauvaise) utilisation de la statistique à travers les âges**

Un exemple classique d'erreur qui a encore la vie dure dans notre inconscient collectif est le fait que les épinards soient remplis de fer :

- En 1870, le biochimiste allemand nommé E. von Wolf découvre que les épinards contiennent environ 2,7mg de fer pour 100g. L'histoire raconte que, quand sa secrétaire recopia la valeur, elle oublia la virgule ce qui multiplia par 10 la teneur.
- En 1881, un chercheur nommé Gustav von Bunge réévalue la teneur en fer mais se trompe entre le poids des épinards frais et des épinards déshydratés. Comme les épinards sont constitués à 90% d'eau, l'erreur était à nouveau de 1 pour 10.
- Dans les années 1930 à 1937, la communauté scientifique réévalua cette teneur et découvrit des deux erreurs mais elle ne réussit pas à convaincre le public. Il fallut attendre l'article de Hamblin (1981) de 1981, soit plus d'un siècle après, pour que les médias reprennent l'information.

Nous conseillons le site *Science & fourchette* recensant les nombreuses erreurs commises au fil des ans sur les épinards : `http://sciencefourchette.com/2014/04/11/popeye-est-une-supercherie/` dont nous avons tiré une partie des informations.

## 1.4 Individus et variables statistiques

Les deux éléments importants à maîtriser dans une étude statistique sont les individus et les variables statistiques.

### 1.4.1 Individus statistiques

Débutons par introduire deux définitions.

**Définitions 1 (Population et individus statistiques)**

La **population statistique** est l'ensemble concerné par une étude statistique. L'**individu statistique** est un élément unique de cette population.

La population statistique doit être correctement choisie en fonction de la question posée.

**Exemple fil rouge**

Dans l'étude fil rouge provenant de Micheaux et al. (2011) que nous allons utiliser tout au long de cette partie, nous analyserons le comportement alimentaire des personnes âgées de la région de Bordeaux en 2000. Il faut alors correctement caractériser ce qu'est *une personne âgée* (plus de 60 ans ? Plus de 80 ans ? Ou alors plus de 40 ans ?) et la *région de Bordeaux* (est-ce la ville ? L'agglomération proche ? Le département qui contient la ville de Bordeaux ?).

**Attention au piège**

Les constatations faites durant l'étude portent uniquement sur la population étudiée (en tenant compte de ses spécificités).

Par exemple, si nous étudions la survie des passagers du Titanic, la population est l'ensemble des passagers et tous les passagers. Il ne faut donc pas étudier uniquement les passagers de première classe ou ceux d'un autre paquebot et extrapoler les résultats.


**La (mauvaise) utilisation de la statistique à travers les âges**

Dans un tweet du 28 juin 2021, le compte @CNEWS publiait les conclusions d'une étude sur le fait que 40% des nouveaux cas de Covid en Israël étaient vaccinées (voir la gauche de la figure 1.1) avec une introduction laissant penser que la *couverture vaccinale importante* n'empêchait une large propagation du virus. Or, comme le rappelle le compte @sc_cath, malgré 87% de la population vaccinée, nous ne retrouvons que 40% dans les nouveaux cas. Ceci est dû à une plus faible contamination des personnes vaccinées (environs 10%) que des personnes non vaccinées (environ 100%).

En poussant le raisonnement à l'extrême limite, si 100% des israéliens étaient vaccinés alors le moindre cas serait forcément un vacciné et le titre serait alors *100% des nouveaux cas sont vaccinés*. La question est ici de savoir si le fait de vacciner largement la population diminue globalement le nombre de nouveaux cas ou pas par rapport à une population non vaccinée.

**Remarque**

Les termes *population* et *individus* font naturellement penser à des êtres humains. Néanmoins, l'individu peut également être un groupement d'êtres humains (par exemple si nous étudions le comportement de plusieurs villes, ce sera la ville l'individu statistique) ou des objets (par exemple un capteur de température).

**Exemple**

Si nous étudions les personnages de Star Wars, ce sont des personnes fictives qui peuvent également être des robots par exemple.


FIGURE 1.1 – Exemple d'étude où la conclusion (tweet de gauche) fut reprise sans recul sur le fait que la population était largement vaccinée et donc, sans avoir pris en compte les probabilités conditionnelles (rappelées par le tweet de droite).

**Attention au piège**

Le fait qu'il y ait un être humain dans une étude ne veut pas dire que l'individu statistique sera un être humain. Par exemple, si nous observons le suivi d'un traitement donné à des patients et que nous faisons des prélèvements à différents moments, l'individu statistique sera alors le couplage *patient + date du prélèvement*.

**Point méthode**

Un individu statistique doit **être unique**. Si vous avez un doute sur la qualification de votre individu statistique, demandez-vous si un même individu se retrouve plusieurs fois. Dans l'exemple précédent, nous voyons que si nous prenons uniquement le patient alors il sera associé à plusieurs prélèvements.

**Exemple**

Si nous étudions les vidéos de la plateforme *Youtube*, la caractérisation de l'unicité se fait par l'identifiant de la vidéo. En effet, une même vidéo (par exemple un clip de musique) peut être déposée par plusieurs youtubeurs et chacune de ces vidéos est pourtant unique.

Une fois la population statistique définie, le mieux serait de récupérer les informations pour tous les individus présents. En pratique, c'est bien sûr compliqué pour des questions de temps et d'argent. Nous introduisons alors de nouvelles notions :

**Définitions 2 ((Sous-)échantillon)**

Un **échantillon (statistique)** est un sous-ensemble de la population statistique. On appelle **sous-échantillon** une partie de cet échantillon.

**Remarque**

Pour plus d'information sur la façon de recueillir un échantillon, nous vous recommandons le cours (optionnel) sur les techniques de sondage du M2 SSD.

**Attention au piège**

Il est important de s'assurer que ce sous-ensemble soit vraiment représentatif de la population. Par exemple, si nous étudions la proportion de cancers en France dans la population et que l'échantillon ne contient que des hommes, nous aurons une sur-représentation des cancers des testicules et une sous-représentation des cancers du sein.

### 1.4.2 Variables statistiques

Une fois les individus correctement définis, nous devons chercher quelles seront les informations importantes à analyser pour répondre à la question.

**Définition 3 (Variable statistique)**

Nous appelons **variable statistique** toute information recueillie sur les individus statistiques.

**Exemple**

Dans le cadre d'une enquête sur la consommation des personnes âgées, nous pouvons leur demander leur âge, leur poids, leurs habitudes alimentaires (boivent-elles du café ou pas ? Si oui, combien de tasses par jour ?), la ville dans laquelle elles vivent...

Les variables statistiques peuvent être classées en deux types contenant chacun deux sous-catégories.

**Définitions 4 (Types de variables)**

Le premier type est celui des **variables qualitatives** qui contient toutes les variables pour lesquelles aucune opération mathématique n'est possible. Nous distinguons deux sous-catégories :

- Les variables qualitatives **nominales** lorsqu'il n'y a pas d'ordre sur ses modalités.
- Les variables qualitatives **ordinales** lorsqu'un ordre accepté par tout le monde peut être fait.

Le deuxième type est celui des **variables quantitatives** qui contient toutes les variables pour lesquelles il est possible de faire des opérations. Nous distinguons deux sous-catégories :

- Les variables quantitatives **discrètes** lorsqu'il y a un faible nombre de valeurs possibles, relevant essentiellement d'un comptage.
- Les variables quantitatives **continues** dans les autres cas.

**Exemples**

- Genre d'une personne : variable qualitative nominale car il n'est pas possible d'additionner *un homme + une femme* et qu'il n'y a pas d'ordre naturel et accepté par tous entre les hommes et les femmes.
- La situation familiale (célibataire, marié·e, veuf·ve, en couple...) : variable qualitative nominale car il n'est pas possible d'additionner *célibataire + en couple* par exemple et qu'il n'y a pas d'ordre naturel et accepté par tous.
- "Êtes-vous *totalement d'accord*, *plutôt d'accord*, *plutôt pas d'accord* ou *pas du tout d'accord* sur le fait que le professeur de statistique est vraiment pédagogue ?" est une question renvoyant une réponse qui est une variable qualitative ordinale puisqu'un ordre naturel est présent mais on ne peut pas faire d'opérations arithmétiques sur les modalités de réponse.
- Le nombre d'enfants d'une personne : variable quantitative discrète puisque l'essentiel de la population se concentre sur un faible nombre d'enfants et qu'on peut calculer le *nombre d'enfants moyens par personne*.

- La pression atmosphérique : variable quantitative continue puisque nous pouvons calculer une pression moyenne et que, si on a un appareil suffisamment précis, il est possible de relever un ensemble continu de valeurs.

Généralement, il y a deux types d'erreurs dans la classification des variables.

**Attention au piège (Variables qualitatives)**

La première difficulté rencontrée est pour différencier une variable qualitative ordinale d'une variable qualitative nominale. Il est important de se souvenir que *l'ordre doit être accepté par tout le monde*. Par exemple, nous pourrions demander la couleur des cheveux de plusieurs personnes et proposer un ordre suivant l'intensité de la couleur. Mais est-ce que tout le monde serait d'accord sur cet ordre ? Par exemple, est-ce que la couleur *roux* est plus intense ou moins intense que *châtain* ou *blond* ?

**Attention au piège (Variables quantitatives)**

La frontière entre *discret* et *continu* est quelque fois floue et il est possible qu'une même donnée puisse être considérée comme continue ou discrète suivant la situation.

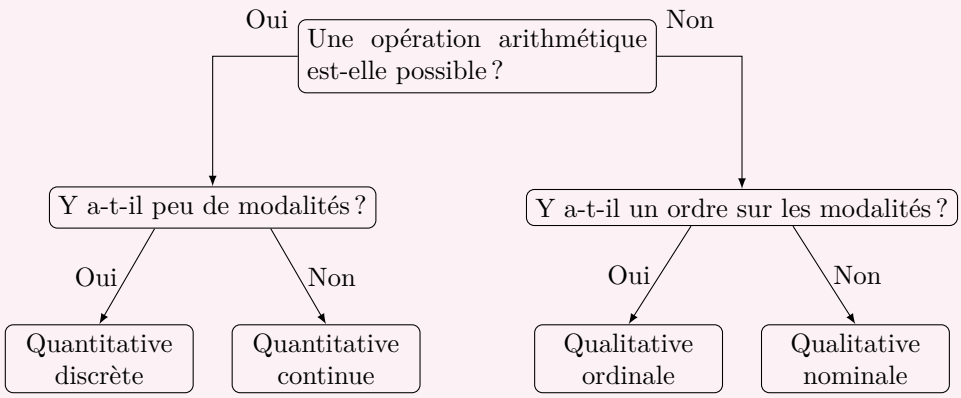
Par exemple, si nous regardons les poids de personnes avec un outil suffisamment précis, nous pourrons avoir autant de poids que d'individus (donc une variable quantitative continue). A l'opposé, si nous regardons les poids de personnes ayant entre 50 et 55 kilogrammes avec un outil ne permettant d'avoir que des valeurs entières, nous n'aurons que 6 valeurs (donc une variable quantitative continue).

Un autre exemple peut être le nombre de **j'aime** sur un réseau social (twitter, facebook, instagram...) : c'est un nombre entier et, généralement, les utilisateurs ont peu de notifications (donc une variable quantitative discrète) mais pour les influenceur·se·s, ce nombre peut être très grand et surtout très différent d'une publication à une autre.

Pour vous aider, voici un arbre de décision :

**Point méthode (Type de variable)**

Pour vous aider à décider du type d'une variable, deux questions sont nécessaires :




Indépendamment de leur type, il faut également différencier les variables qui nous permettront de répondre à la question.

**Définitions 5 (Variables explicatives et à expliquer)**

Nous appelons **variable à expliquer** les variables que nous cherchons à décrire, expliqué ou prédire afin de répondre à la question posée en début d'étude. Nous appelons **variable explicative** les variables utilisées pour expliquer, décrire ou prédire les variables à expliquer.

**Exemple**

Dans le cas des passagers du Titanic, nous cherchons à expliquer la variable *survie* en fonction de différentes variables explicatives (âge, prix du billet, port d'embarquement...).


**La (mauvaise) utilisation de la statistique à travers les âges**

Avec l'avènement des algorithmes automatiques et du *Big Data*, de nouvelles questions se posent quant au choix et à l'utilisation de certaines variables explicatives. Par exemple, nous nous apercevons que, mal utilisés, certains algorithmes reproduisent, voire accentuent, les inégalités déjà présentes. En 2015, des chercheurs et chercheuses ont montré que des profils masculins avaient plus de chances de se voir proposer des offres à hauts salaires que des profils féminins. Ceci est dû au fait que jusqu'à présent, les hommes ont des salaires plus élevés et, comme les concepteurs n'ont pas corrigé ce biais, les algorithmes ont cru que c'était la norme. Nous recommandons la vidéo *Algocratie : L'inégalité programmée - #DATAGUEULE 84* de la chaîne *Data Gueule* proposée par Goetz et al. (2014).

## 1.5 Statistiques descriptives

Dans la démarche énoncée dans la partie 1.3, la statistique descriptive prend sa place dans la partie d'exploration et de **prétraitement des données**. Ce sera à partir des résultats présentés que nous pourrons proposer une modélisation statistique (qui sera abordée dans tout le reste de la formation) qui mènera à une prévision et/ou à une prise de décision. Dans ce cadre, trois grandes catégories existent :

- Univarié.
- Bivarié.
- Multivarié.

La statistique descriptive désigne un ensemble de techniques dont le but est de

- explorer, découvrir l'information contenue dans les données,
- les représenter graphiquement,
- détecter des premières tendances.

A chacun de ces buts correspond une ou plusieurs techniques (voir le tableau 1.1).

TABLE 1.1 – Mise en relation des objectifs (à gauche) avec les outils possibles (au milieu). La dernière colonne correspond au cours dans lequel nous verrons ces notions.

| Objectif | Techniques | Cours |
|---|---|---|
| Explorer, découvrir les données, résumer l'information | tableaux statistiques<br>résumés statistiques | Univarié |
| Représenter graphiquement | graphiques | Univarié |
| Détecter des tendances | indicateurs de liaison | Bivarié et multivarié |

Il est important de ne pas utiliser une technique pour un autre objectif. De même, nous verrons par la suite que chaque outil correspond à un ou plusieurs objectifs précis : leur choix est donc primordial.

La suite de cette partie consiste à présenter ces outils suivants les catégories (univariée, bivariée puis multivariée).

### 1.5.1 Statistiques descriptives univariées

Pour la description, les statistiques descriptives univariées permettent d'observer les variables indépendamment des autres. Nous pouvons les décomposer en plusieurs grandes catégories :

- Tableau
- Résumés statistiques
  - ♦ de position
  - ♦ de dispersion
  - ♦ de forme

- Graphique

TABLE 1.2: Tableau récapitulatif des outils et de quand les appliquer. Le symbole ✍ signifie qu'il faut faire attention pour l'utiliser (par exemple, que sur les variables positives) ou qu'il y a une différence vis-à-vis de la définition usuelle (comme pour les fractiles dans le cas discret).

| Nom | Qualitative | | Quantitative | |
|---|---|---|---|---|
| | Nominale | Ordinale | Discrète | Continue |
| **Tableaux** | | | | |
| Tri à plat | ✓ | ✓ | ✓ | |
| **Résumés statistiques de position** | | | | |
| Centile | | | ✍ | ✓ |
| Classe modale | | | | ✓ |
| Décile | | | ✍ | ✓ |
| Fractile | | | ✍ | ✓ |
| Médiane | | | ✍ | ✓ |
| Maximum | | | ✓ | ✓ |
| Minimum | | | ✓ | ✓ |
| Mode | ✓ | ✓ | ✓ | |
| Moyenne | | | ✓ | ✓ |
| Quartile | | | ✍ | ✓ |
| **Résumés statistiques de dispersion** | | | | |
| Écart-type | | | ✓ | ✓ |
| Étendue | | | ✓ | ✓ |
| Longueur interquartile | | | ✓ | ✓ |
| Rapport interdécile | | | ✍ | ✍ |
| Variance | | | ✓ | ✓ |
| **Résumés statistiques de forme** | | | | |
| Kurtosis | | | ✓ | ✓ |
| Skewness | | | ✓ | ✓ |
| **Graphiques** | | | | |
| Boxplot ou boîte à moustaches | | | ✓ | ✓ |
| Diagramme circulaire | ✍ | ✍ | | |
| Diagramme de Pareto | ✓ | | | |
| Diagramme empilé | ✓ | ✓ | | |
| Diagramme en barres ou tuyaux d'orgues | ✓ | ✓ | | |
| Diagramme en bâtons | | | ✓ | |
| Fonction de répartition empirique | | | ✓ | ✓ |
| Histogramme | | | | ✓ |
| Polygone régulier | | | | ✓ |

Dans le tableau 1.2, tous les outils ont été récapitulés. Le symbole ✍ signifie qu'il y a une particularité vis-à-vis de l'utilisation ou de la définition classique. Notamment, il est important de noter :

- Dans le cas discret, les fractiles (médiane, quartiles, déciles et centiles notamment) doivent appartenir à l'ensemble de définition (par exemple, un nombre entier si on parle du nombre d'enfants dans un foyer). Les logiciels utilisent par défaut la version continue qui peut faire des pondérations si le fractile se trouve entre deux valeurs distinctes.
- Le rapport interdécile ne s'utilise que sur des variables strictement positives.
- Le diagramme circulaire ne s'utilise que quand il y a un nombre très réduit (typiquement quatre ou moins) de modalités effectives (c'est-à-dire de modalités prises par au moins un individu).

Une autre façon de résumer l'information est de faire un diagramme en mettant les outils dans la zone qui correspond aux types de variables sur lesquelles les outils s'appliquent (voir la figure 1.2).

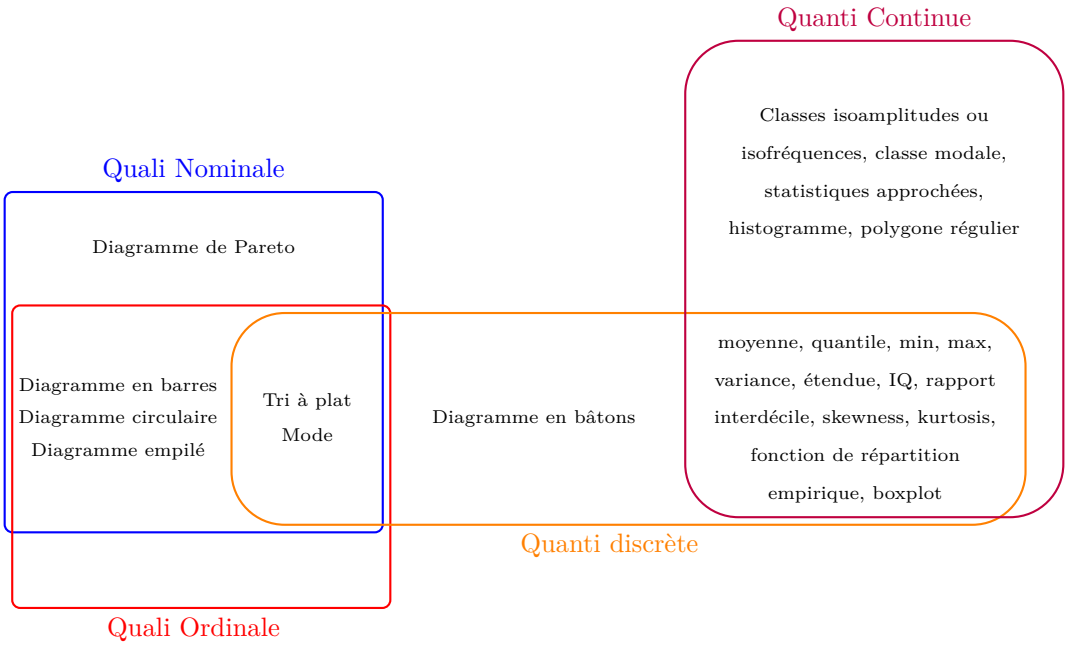



FIGURE 1.2 – Diagramme des outils en fonctions des types de variables.

### Tableaux

Pour présenter le tri à plat, nous avons besoin de quelques définitions.

**Définitions 6 (Effectifs et fréquences)**

Etant donnée une modalité $a_k$ de la variable qualitative **X**, nous appelons **effectif de la modalité**, noté $N_k$, le nombre d'individus prenant la modalité $a_k$. Nous appelons **fréquence de la modalité**, notée $f_k$, la proportion d'individus prenant la modalité $a_k$ parmi ceux qui ont donné une réponse. Nous avons donc les relations suivantes :

$$f_k = \frac{N_k}{N} \text{ et } N_k = N f_k$$

sous la condition que $N > 0$.

Dans le cas de variables qualitatives ordonnées et de quantitatives discrètes, nous introduisons le principe de fréquence cumulée :

**Définition 7 (Fréquence cumulée pour variable qualitative ordonnée)**

Étant donnée une modalité $a_k$ de la variable qualitative **ordonnée X**, nous appelons **fréquence cumulée de la modalité**, notée $F_k$, la proportion d'individus prenant la modalité $a_k$ ou une modalité inférieure. Nous avons donc les relations suivantes :

$$F_k = \sum_{\ell=1}^{k} f_\ell = \frac{1}{N} \sum_{\ell=1}^{k} N_\ell.$$

Une autre façon de l'écrire est :

$$F_k = \sum_{\ell \leq k} f_\ell = \frac{1}{N} \sum_{\ell \leq k} N_\ell.$$

Avec ces notations, nous pouvons introduire le tableau de tri à plat.

**Définition 8 (Tri à plat)**

Un **tri à plat** est un tableau récapitulant toutes les informations précédentes sous la forme suivante :

Qualitative nominale

| Modalités | Effectifs | Fréquences |
|---|---|---|
| $a_1$ | $N_1$ | $f_1$ |
| $a_2$ | $N_2$ | $f_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $a_k$ | $N_k$ | $f_k$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $a_K$ | $N_K$ | $f_K$ |
| Total | $N$ | 1 |

Qualitative ordinale ou quantitative discrète

| Modalités | Effectifs | Fréquences | Cumulées |
|---|---|---|---|
| $a_1$ | $N_1$ | $f_1$ | $F_1 = f_1$ |
| $a_2$ | $N_2$ | $f_2$ | $F_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a_k$ | $N_k$ | $f_k$ | $F_k$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a_K$ | $N_K$ | $f_K$ | $F_K = 1$ |
| Total | $N$ | 1 | |

Si la variable contient des données manquantes (c'est-à-dire si $N < n$), nous préciserons le pourcentage de répondants lorsque nous présenterons le tableau (généralement, nous le mettons dans la légende du tableau ; voir le chapitre **??**).

### Résumés statistiques de dispersion

Dans cette partie, nous présentons quelques résumés statistiques de dispersion.

**Définition 9 (Étendue)**

L'**étendue** (***range*** en anglais) de la distribution $x_1, \dots, x_n$, notée $\boldsymbol{W}$, est définie comme l'écart (positif) entre la plus grande et la plus petite valeur :

$$W = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i.$$

Nous introduisons maintenant la longueur de l'intervalle inter-quartile.

**Définition 10 (Intervalle interquartile)**

L'**intervalle interquartile** de la distribution $x_1, \dots, x_n$ est l'intervalle de bornes $Q_1$ et $Q_3$ ; c'est-à-dire $[Q_1; Q_3]$. La **longueur de cet intervalle** est un indicatif de dispersion, noté $\boldsymbol{IQ}$, valant :

$$IQ = Q_3 - Q_1.$$

**Attention au piège**

L'erreur la plus commune faite par les étudiants est de dire que *l'intervalle interquartile vaut 5* par exemple : il y a confusion entre l'intervalle (c'est-à-dire l'objet $[Q_1; Q_3]$) et sa longueur.

**Point méthode (Vérifications)**

Comme nous avons les relations suivantes

$$\min_{1 \leq i \leq n} x_i \leq Q_1 \leq Q_3 \leq \max_{1 \leq i \leq n} x_i$$

alors nous devons vérifier les deux points suivants :

1. $IQ \geq 0$ (comme pour $W$).
2. $IQ \leq W$.

### Résumés statistiques de forme

Les résumés statistiques de forme sont basés sur la forme de la distribution d'une gaussienne.

**Définition 11 (Coefficient d'asymétrie)**

Le **coefficient d'asymétrie** (ou ***skewness*** en anglais) mesure la dissymétrie de la distribution. Sa formule vaut :

$$G_1(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma_x^3}$$

où $\overline{x}$ est la moyenne empirique des valeurs des observations de la variable $X$ et $\sigma_x$ est son écart-type.

**Remarque**

Plus la distribution sera symétrique, plus le coefficient sera proche de 0.

**Définition 12 (Coefficient d'aplatissement)**

Le **coefficient d'aplatissement** (ou ***kurtosis***) s'utilise dans le cadre d'une distribution symétrique (donc si le précédent coefficient est proche de 0) et mesure la répartition des poids. Sa formule vaut :

$$G_2(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma_x^4} - 3.$$

**Remarque**

Il sera proche de 0 si la distribution ressemble à une loi gaussienne et négatif si les points sont répartis uniformément.

### Graphiques

Dans cette partie, nous présentons les graphiques.

**Boxplot ou boîte à moustaches :** La boîte à moustaches est un diagramme généralement présenté horizontalement dans les cours de lycée. Nous faisons le choix de le représenter verticalement comme le font de nombreux logiciels.

**Définition 13 (Boxplot ou boîte à moustaches (variante))**

La variante du **boxplot** ou **boîte à moustaches** est la suivante :

1. Nous traçons 3 traits horizontaux de mêmes longueurs placés aux niveaux de chacun des 3 quartiles centraux : $Q_1$, médiane et $Q_3$.
2. Deux traits relient verticalement les extrémités du premier ($Q_1$) et du troisième quartiles ($Q_3$). Ce trait est censé passer par les extrémités du trait horizontal symbolisant la médiane.
3. Nous calculons la longueur interquartile $IQ$.
4. Nous regardons si le maximum est plus grand que $Q_3 + 1,5IQ$ :
   - Si la réponse est oui :
     (a) Nous traçons un trait horizontal à l'ordonnée $Q_3 + 1,5IQ$ (plus petit que les autres traits horizontaux).
     (b) puis nous relions le centre de ce trait avec le centre du trait symbolisant le troisième quartile $Q_3$.
     (c) Nous mettons une croix rouge à l'emplacement du maximum et dans l'alignement des centres des traits horizontaux.

(d) (Facultatif) Enfin, nous mettons des croix bleues pour chaque modalité effective (c'est-à-dire ayant un effectif non nul) entre $Q_3 + 1,5IQ$ et le maximum (également dans l'alignement des centres des traits horizontaux).

- Si la réponse est non (et que le maximum est donc plus petit que $Q_3 + 1,5IQ$), nous procédons de la sorte :
  (a) Nous traçons le trait horizontal correspondant au maximum.
  (b) Nous relions ce trait au trait symbolisant le troisième quartile.

5. Nous faisons la procédure symétrique avec le bas du graphique en comparant le minimum avec la valeur $Q_1 - 1,5IQ$.

Dans le cas où le maximum serait plus grand que $Q_3 + 1,5IQ$ et/ou que le minimum serait plus petit que $Q_1 - 1,5IQ$, nous devons préciser combien d'individus se trouvent dans les intervalles $\left[\min_{1\leq i\leq n} x_i; Q_1 - 1,5IQ\right[$ et $\left]Q_3 + 1,5IQ; \max_{1\leq i\leq n} x_i\right]$ ; nous appelons ces points des ***outsiders***.

**Remarque**

Il existe une autre variante consistant à ne pas dépasser $Q_1 - 3IQ$ et $Q_3 - 3IQ$ : si le maximum (resp. le minimum) est plus grand que $Q_3 - 3IQ$ (resp. plus petit que $Q_1 - 3IQ$), nous mettons la croix rouge au niveau de $Q_3 + 3IQ$ (resp $Q_1 - 3IQ$) et nous ne mettons pas de croix bleues au delà de cette limite.

**Point logiciel**

Dans le logiciel `Excel`, il n'existe pas de fonctions pour en construire simplement ; nous verrons en TP une procédure pour le faire à partir de nuage de points. Dans le langage `R`, nous pouvons utiliser la fonction `boxplot` ou le package `ggplot2` par exemple.

**Diagramme circulaire :** Ce diagramme est connu aussi sous le nom de *camembert*.

**Définition 14 (Diagramme circulaire)**

Le **diagramme circulaire** consiste à découper un disque en plusieurs zones de telle sorte que :

- Chaque zone est délimitée par deux rayons et une partie du cercle.
- Chaque angle (et donc chaque zone) est proportionnel à la fréquence de la modalité associée.

**Attention au piège**

Bien que très visuel, le diagramme circulaire rend les comparaisons entre fréquences des modalités plus difficiles à faire car nous ne sommes pas vraiment habitué·e·s à comparer des angles (par opposition à la comparaison des fréquences dans le cas des diagrammes en tuyaux d'orgue). En particulier, il est fortement déconseillé d'utiliser ce diagramme si :

- Les fréquences sont assez proches.
- Il y a beaucoup de modalités.

**Point méthode (Calcul de l'angle)**

Comme un cercle fait au total 360° et que chaque région est proportionnelle à la fréquence de la modalité, nous calculons l'angle de chaque région par la formule $360° \times f_k$ où $f_k$ est la fréquence de la modalité $a_k$.

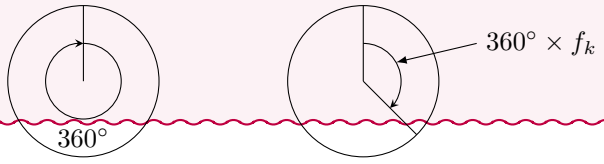


**Diagramme en barres et diagramme de Pareto :** Pour présenter ces diagrammes, nous commençons par celui en barres.

**Définition 15 (Diagramme en tuyaux d'orgue ou en barres)**

Le **diagramme en tuyaux d'orgue** ou **en barres** consiste à représenter les fréquences des modalités d'une variable qualitative à l'aide de barres rectangulaires verticales parallèles de telle sorte que :

- Toutes les barres ont la même largeur à la base.
- La hauteur de chaque barre représente la valeur de la fréquence (l'axe des ordonnées est gradué suivant les valeurs des fréquences).
- Les barres sont toutes espacées de la même distance (elles ne se collent pas).
- Le nom de chaque modalité doit apparaître en dessous de chaque barre.
- Dans le cas de variables qualitatives ordinales, il faut conserver l'ordre naturel des modalités.

Le diagramme de Pareto est un diagramme en tuyaux d'orgue particulier.

**Définition 16 (Diagramme de Pareto)**

Le **diagramme de Pareto** est un diagramme en tuyaux d'orgue où les modalités sont réorganisées par ordre décroissant de fréquences.

**Attention au piège**

Comme nous intervenons sur l'ordre des modalités, le diagramme de Pareto ne s'utilise jamais pour des variables qualitatives ordonnées.

**Diagramme empilé :** Le diagramme empilé permet une représentation rapide des fréquences cumulées.

**Définition 17 (Diagramme empilé)**

Le **diagramme empilé** consiste à représenter les fréquences des modalités à l'aide de rectangles empilés les uns au-dessus des autres de telle sorte que :

- Tous les rectangles ont la même largeur.
- La hauteur de chaque rectangle correspond à la fréquence de chaque modalité.
- Les rectangles sont empilés les uns au-dessus des autres de telle sorte que la hauteur totale fasse 100%.

**Diagramme en bâtons :** Le diagramme en bâtons est spécifique aux variables quantitatives discrètes.

**Définition 18 (Diagramme en bâtons)**

Le **diagramme en bâtons** consiste à représenter les fréquences des modalités d'une variable quantitative discrète à l'aide de segments verticaux de coordonnées $(a_k, 0)$ et $(a_k, f_k)$ :

- Le segment est parallèle à l'axe des ordonnées (ou perpendiculaire à l'axe des abscisses suivant le point de vue).
- Il se situe en abscisse au niveau de la modalité $a_k$.
- Il va de 0 à $f_k$.

Éventuellement, on peut rajouter un point en haut de chaque segment.

**Fonction de répartition empirique :** La fonction de répartition empirique est une fonction dont le graphique est peu utilisé alors qu'il contient un très grand nombre d'informations.

**Définition 19 (Fonction de répartition empirique)**

La **fonction de répartition empirique** est une fonction définie pour tout $t \in \mathbb{R}$ par :

$$\widehat{F}(t) = \sum_{\substack{k=1 \\ \text{tel que } a_k \leq t}}^{K} f_k = \sum_{k=1}^{K} f_k \mathbb{1}_{\{a_k \leq t\}}$$

où $\mathbb{1}_{\{a_k \leq t\}}$ vaut 1 si et seulement si $a_k \leq t$ et 0 sinon.

**Proposition 1 (Propriété de la fonction de répartition empirique)**

Les propriétés de la fonction de répartition empirique sont les suivantes :

- Elle est croissante et constante par morceaux (nos disons aussi qu'elle est *en escalier croissant*).
- Chaque saut se fait au moment d'une modalité avec une fréquence non nulle.
- La hauteur de chaque saut correspond à la fréquence de la modalité associée.
- Elle vaut 0 avant la première modalité et 1 après la dernière modalité.
- Elle est continue à droite ; c'est-à-dire que, au moment de la modalité $a_k$ la fonction vaut toujours $F_k$ qui est la valeur du plateau du-dessus.

### 1.5.2 Statistiques descriptives bivariées

Le principe des croisements est de regarder les corrélations possibles entre deux variables. Ainsi, il est intéressant comment une variable quantitative se comporte vis-à-vis d'une autre variable quantitative.

**Attention au piège**

Attention, tous les croisements ne se représentent pas (voir le tableau 1.6). Par exemple, il est possible de regarder les salaires de personnes en fonction de leur sexe (une variable quantitative en fonction d'une variable qualitative). L'autre sens est plus compliqué à représenter. Dans le cas de la description, cela n'a pas d'importance car seule une corrélation est recherchée (pas une causalité).

Nous voyons dans le tableau 1.6 qu'il y a un indicateur particulier pour le cas d'un croisement entre deux ordinales. Si l'une des deux variables n'est pas une ordinale (par exemple, le croisement d'une nominale avec une ordinale), il n'existe plus d'indicateurs particuliers : il faut se contenter de considérer que l'ordinale est une qualitative (presque) comme une autre.

TABLE 1.3 – Résumé des possibilités d'expliquer une variable (colonnes) en fonction d'une variable (lignes) : ✓ s'il existe un indicateur de liaison, ✗ s'il n'y a pas de théorie et ✍ si c'est possible mais qu'il faut prendre un indicateur plus général (dans le cas où il faut se contenter qu'une ordinale est une qualitative).

| en fonction de \ à expliquer | Qualitative | Quantitative | Ordinale |
|---|---|---|---|
| Qualitative | ✓ | ✓ | ✍ |
| Quantitative | ✗ | ✓ | ✗ |
| Ordinale | ✍ | ✍ | ✓ |

Comme précédemment, il y a trois groupes d'outils qui peuvent être utilisés :

- Tableau (voir le tableau 1.4)

- Graphique (voir le tableau 1.5)
- Indicateurs de liaisons (voir le tableau 1.6).

TABLE 1.4 – Tableaux qui peuvent être calculés pour chaque croisement.

| à expliquer / en fonction de | Qualitative | Quantitative | Ordinale |
|---|---|---|---|
| Qualitative | Contingence, distribution conjointe, distribution conditionnelle, effectifs théoriques | Tri à plat par sous populations | |
| Quantitative | | Tableau des variables centrées | |
| Ordinale | | | Tableau des concordances |

TABLE 1.5 – Graphiques qui peuvent être calculés pour chaque croisement.

| à expliquer / en fonction de | Qualitative | Quantitative |
|---|---|---|
| Qualitative | Diagrammes empilés, Diagrammes en barres | Boxplots |
| Quantitative | | Nuages de points |

TABLE 1.6 – Indicateurs de liaisons qui peuvent être calculés pour chaque croisement.

| à expliquer / en fonction de | Qualitative | Quantitative | Ordinale |
|---|---|---|---|
| Qualitative | Khi2 $\chi^2_n$, Phi2 $\phi^2$, $V^2$ de Cramer | Variances intra et inter, rapport de corrélation $\eta^2$ | |
| Quantitative | | Covariance, coefficient de corrélation linéaire (de Pearson) | |
| Ordinale | | | Tau de Kendall |

Il est important de souvenir que les indicateurs du tableau 1.6 ne permettent qu'une description : il faudra confirmer par un test ensuite (voir le cours de *Compléments en tests statistiques*).

**Attention au piège**

Attention, une corrélation ne veut pas dire qu'il y a une causalité. Par exemple, il est observé une corrélation entre les coups de soleil et une augmentation dans les ventes de glaces. Même si cette corrélation peut être intéressante mais il ne faut pas voir une causalité : par exemple, en tant que vendeur de glaces, je peux me dire qu'il faut que j'ouvre mon échoppe si je vois qu'il y a beaucoup de coups de soleil, cela ne veut pas dire que les coups de soleil impliquent que je vais vendre plus de glaces.

En particulier, il arrive que certaines co-variables n'aient pas été prises en compte dans le cadre de l'étude et viennent fausser les résultats. Nous renvoyons par exemple les statisticien·ne·s intéressé·e·s à la vidéo *Chocolat, corrélation et moustaches de chats* [3] de la chaîne Youtube *la statistique expliquée à mon chat* [4] proposée par Uyttendaele et al. (2016).

3. Disponible à l'url suivante : `https://www.youtube.com/watch?v=aOX0pIwBCvw`
4. L'url de la chaîne est la suivante : `https://www.youtube.com/channel/UCWty1tzwZW_ZNSp5GVGteaA`

**La (mauvaise) utilisation de la statistique à travers les âges**

Pour sensibiliser à la différence entre corrélation et causalité, *les décodeurs* du journal *Le monde* proposent un générateur de corrélation aléatoire :
`https://www.lemonde.fr/les-decodeurs/article/2019/01/02/correlation-ou-causalite-brillez-en-societe-avec-notre-generateur-aleatoire-de-comparaisons-absurdes_5404286_4355770.html`
N'hésitez pas à vous amuser avec.

### 1.5.3 Statistiques descriptives multivariées

Les statistiques descriptives multivariées portent sur des techniques impliquant plus de deux variables. Par exemple, l'*Analyse en Composantes Principales* (ou ACP) ou la *Classification Ascendante Hiérarchique* (CAH) peuvent être citées. Toutes ces méthodes sont abordées dans le cours d'*analyse de données* ou encore dans des livres tels que Husson et al. (2016) ou sur le site *Wikistat* de Besse.

# Chapitre 2

# Modélisation

*"Le loto, c'est un impôt sur les gens qui ne comprennent pas les statistiques."*

## 2.1 Objectifs

L'objectif de ce court chapitre est de comprendre les fondements de la statistique, la philosophie et le principe de modélisation.

## 2.2 Introduction

Commençons par quelques exemples d'application de la statistique :

**Exemple**

Adapté d'un exemple tiré du livre de Robert (2006). Considérant la proportion proportion de naissances masculines à Paris, Pierre-Simon de Laplace s'intéresse à savoir s'il y a plus de naissances d'hommes que de femmes. Pour cela, il va recenser 251 527 naissances masculines et 241 945 naissances féminines en 1785.
Une naissance peut être considérée comme le résultat d'une variable $X_i$ de Bernoulli de paramètre $\theta \in [0;1]$ :

$$X_i = \begin{cases} \text{homme} & \text{avec probabilité } \theta, \\ \text{femme} & \text{avec probabilité } 1-\theta. \end{cases}$$

$X_1, \ldots, X_n$ est alors l'échantillon des $n = 493\,472$ résultats de naissances observées. La question qu'on peut alors se poser est de savoir si $\theta = 1/2$, $\theta < 1/2$ ou $\theta > 1/2$.

**Exemple**

Adapté d'une histoire sur la *cryptographie automatique* de Ycart (2017). Une ancienne technique de cryptographie consiste à associer chaque lettre de l'alphabet à une autre de façon unique : nous obtenons ainsi une bijection entre un texte lisible et un texte codé. Pour décoder, il suffit de faire la transformation inverse. Par exemple, avec le tableau de correspondance suivant :

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K | R | D | U | C | W | M | G | B | E | F | P | N | H | I | Y | O | S | X | J | A | V | L | T | Q | Z |
| | | | | | | | | | | | | | | | | | | | | | | | | | |

nous pouvons décrypter le texte :

**KYYSCHUSC PK XJKJBXJBOAC CXJ SCYIXKHJ** [1]

Une technique de *cryptanalyse* pourrait être d'essayer toutes les combinaisons possibles mais cela représente $26! \approx 4 \times 10^{26}$ possibilités pour la langue française (à la raison d'une possibilité toutes les secondes, cela prendrait un peu moins de $10^{10}$ fois l'âge de l'univers). Une autre idée a été proposée dès le $9^{\text{ème}}$ siècle par Al-Kindi basée sur une méthode statistique : en fait, il s'est aperçu que toutes les lettres ne revenaient pas de la même façon. Par exemple, dans la langue française, le $e$ est celui qui revient le plus souvent (avec une fréquence de $12,10\%$ pour le corpus de wikipédia en 2008 par exemple

si nous séparons les $e$ sans accent de ceux avec accent [2]).
L'idée derrière est de dire que chaque lettre est le résultat d'une variable aléatoire $X_i$ prenant ses valeurs dans $\{A, B, \dots, Z\}$ avec une probabilité $(p_A, p_B, \dots, p_Z)$. En calculant les fréquences dans le texte, nous pouvons essayer de retrouver le codage.
Notons toutefois que ce système demande des textes codés relativement longs. Pour accélérer l'estimation, il est conseillé de coupler avec les probabilités conditionnelles d'une lettre sachant une autre ; par exemple, la probabilité d'avoir deux $w$ successifs est quasiment nulle.

En particulier, la procédure proposée par Alan Turing (1912-1954) pour décrypter Enigma fonctionna car les allemands envoyés régulièrement les bulletins météorologiques avec une chaîne de caractère régulière et assez longue à chaque fois (voir par exemple Lehning (2006)).

**Exemple**

Dans le film de présentation [3], nous nous étions intéressés à la hauteur d'un fleuve en fonction des pluies tombées dans les montagnes. Pour cela, nous supposons que chaque observation est le résultat d'une variable aléatoire $Y_i$ qui dépend d'une co-variable $x_i$ (la quantité de pluie tombée dans les montagnes) de façon polynomiale ; c'est-à-dire qu'il existe un polynôme $P_m$ de degrés $m$ inconnu et de coefficients inconnus tels que pour tout $i \in \mathbb{N}$ :

$$Y_i = P_m(x_i) + \varepsilon_i \text{ avec } \varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$$

avec $\sigma$ également inconnu. La question est alors d'estimer au mieux chacun des paramètres notamment le degré $m$ du polynôme car la théorie de la sélection de modèle nous permet de voir que le polynôme qui collerait le mieux aux données serait un polynôme de degrés $n$ (passant par toutes les observations) mais ce serait également le plus éloigné du vrai polynôme [4]. Ce phénomène est appelé du *surapprentissage*.

## 2.3 Modèle statistique

Commençons par quelques définitions.

**Définitions 20 (Modèle statistique)**

Un **modèle statistique** est la donnée d'un espace mesurée $(E, \mathcal{E})$ et d'une famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$ de mesures de probabilité d'indexation $\Theta$. Le modèle associé est noté $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$.
Quand il existe $d \in \mathbb{N}^\star$ tel que $\Theta \subset \mathbb{R}^d$, nous disons que le modèle est **paramétrique**. Sinon, il est dit **non-paramétrique**.

**Remarque**

$\theta$ est inconnu mais son ensemble d'appartenance $\Theta$ est connu.

**Exemple**

Pour la loi de Bernoulli, nous savons que $\theta \in [0; 1]$.

**Attention au piège**

Le plus souvent, l'indexation se fait sur une seule famille de probabilités. Dans de très rares cas, nous pouvons imaginer une concurrence entre plusieurs familles de lois. Par exemple, sur l'ensemble fini $\mathbb{N}$, nous pouvons mettre en concurrence des lois de Poisson de paramètres $\lambda \in \mathbb{R}^+$ et des lois géométriques de paramètres $p \in [0; 1]$.

1. La solution étant **APPRENDRE LA STATISTIQUE EST REPOSANT** .
2. Voir par exemple `https://fr.wikipedia.org/wiki/Fr%C3%A9quence_d%27apparition_des_lettres_en_fran%C3%A7ais`
3. disponible ici : `https://www.youtube.com/watch?v=wQ-1Zj9sBlM`
4. Une vidéo pour illustrer ce problème est disponible ici : `https://www.youtube.com/watch?v=YuP1dOIbWfQ`

La démarche statistique consiste à donner de l'information sur le paramètre $\theta$ en s'appuyant sur la notion d'observation ; on parle alors d'**inférence** ou de **statistique inférentielle**.

**Définition 21 (Observation)**

Une **observation** $X$ est une variable aléatoire à valeur dans $E$ et dont la loi appartient à la famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

**Point méthode (Hypothèse fondamentale)**

L'hypothèse fondamentale de la statistique est de supposer qu'une observation $X$ suit une loi $\mathbb{P}_X$ et qu'il existe un paramètre $\theta \in \Theta$ tel que $\mathbb{P}_X = \mathbb{P}_\theta$.

**Remarque**

Dans le langage courant, nous confondons souvent l'observation $X$ qui est une variable aléatoire et sa réalisation $x$ qui est fixée.

**Définition 22 ($n$-échantillon)**

Étant donné $n \in \mathbb{N}^\star$, un **$n$-échantillon de loi** $\mathbb{P}_X$ est la donnée de $n$ variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon la loi $\mathbb{P}_\theta$.

**Remarque**

Lorsque l'observation $X$ a la structure d'un $n$-échantillon de loi $\mathbb{P}_\theta$ alors $\mathbb{P}_{\mathbf{X}} = (\mathbb{P}_\theta)^{\otimes n}$.

**Exercices 2.1**

Reprendre les exercices en début de cours et dire si les modèles sont paramétriques ou non-paramétriques. Dans le cas de modèles paramétriques, donnez l'espace des paramètres. Dans tous les cas, donner l'espace des observations.

**Exemple**

Une entreprise de téléphonie pense que le nombre d'appels journaliers passés par chaque client suit une loi de Poisson et souhaite tester cette hypothèse pour pouvoir fixer les tarifs. Pour ce faire, elle va choisir aléatoirement des clients et récupérer le nombre d'appels.
Formellement, les observations $X_1, \dots, X_n$ sont supposées indépendantes et de même loi et l'entreprise se demande donc si celle-ci suit une loi de Poisson et, dans ce cas, quel est le paramètre ?

**Remarque**

Il peut arriver que l'exigence d'indépendance soit affaiblie voir totalement abandonnée. C'est par exemple le cas dans la plupart des analyses de série chronologique (Voir par exemple le cours de *Données dépendantes 1 : séries temporelles*).

**Point méthode (Quelle loi choisir ?)**

La problématique du choix de la loi dans le cas de modèles non-paramétriques peut être simplifiée en regardant l'ensemble de définition. Voici quelques exemples classiques :

- Si l'ensemble est dénombrable :
  - S'il n'a que deux éléments : loi de Bernoulli.
  - Si c'est $\{0, \dots, n\}$ : loi binomiale, loi uniforme, loi hypergéométrique.
  - Si c'est $\mathbb{N}$ ou $\mathbb{N}^\star$ : loi de Poisson, loi géométrique, loi binomiale négative.

- Si l'ensemble est non dénombrable :
  - Si c'est l'intervalle $[0,1]$ : loi uniforme, loi bêta.
  - Si c'est un intervalle borné : loi uniforme.
  - Si c'est $\mathbb{R}^+$ : loi exponentielle, loi gamma, loi inverse gamma.
  - Si c'est $[a,+\infty[$ avec $a>0$ : loi de Pareto
  - Si c'est $\mathbb{R}$ : loi gaussienne ou normale, loi de Cauchy, loi de Laplace.

Une autre solution consiste à utiliser les liens entre les lois. Pour cela, nous renvoyons à la figure 2.1.

**Définition 23 (Identifiabilité)**

Étant donné un modèle paramétré par $\theta$, nous dirons que le modèle est **identifiable** si l'application $\theta \mapsto \mathbb{P}_\theta$ est injective.

**Contre-exemple**

Le contre-exemple le plus simple est celui de la loi gaussienne centrée de variance inconnue $\mathcal{N}(0,\theta^2)$ :

- Si $\Theta=\mathbb{R}_+^\star$, le modèle est identifiable.
- Si $\Theta=\mathbb{R}^\star$, le modèle n'est pas identifiable car pour tout $\theta\neq 0$, les paramètres $\theta$ et $-\theta$ fournissent la même loi.

**Exercices 2.2**

Pour les lois mentionnées dans le point méthode précédent, donner l'espace de paramètre $\Theta$ de telle sorte que les lois correspondent à des modèles identifiables. Pour rappel, nous mettons les densités :

| Nom | Abréviation | Support | Densité $f(x)$ ou $\mathbb{P}(X=x)$ |
|---|---|---|---|
| Bernoulli | $\mathcal{B}(p)$ | $\{0,1\}$ | $p^x(1-p)^{1-x}$ |
| Bêta | $\mathcal{B}e(a,b)$ | $]0,1[$ | $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ |
| Binomiale | $\mathcal{B}in(n;p)$ | $\{0,1,\dots,n\}$ | $\binom{n}{x}p^x(1-p)^{n-x}$ |
| Binomiale négative | $\mathcal{NB}in(r;p)$ | $\mathbb{N}$ | $\binom{x+n-1}{x}p^n(1-p)^x$ |
| Cauchy | $\mathcal{C}au(\mu,a)$ | $\mathbb{R}$ | $\frac{\pi}{1+\left(\frac{x-\mu}{a}\right)^2}$ |
| Exponentielle | $\mathcal{E}(\lambda)$ | $\mathbb{R}^+$ | $\lambda e^{-\lambda x}$ |
| Gamma | $\mathcal{G}a(\lambda)$ | $\mathbb{R}_+^\star$ | $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ |
| Gaussienne ou normale | $\mathcal{N}(\mu,\sigma^2)$ | $\mathbb{R}$ | $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ |
| Géométrique | $\mathcal{G}éo(p)$ | $\mathbb{N}^\star$ | $(1-p)^{x-1}p$ |
| Hypergéométrique | $\mathcal{H}(n,p,A)$ | $\{0,1,\dots,n\}$ | $\frac{\binom{pA}{x}\binom{(1-p)A}{n-x}}{\binom{A}{x}}$ |
| Inverse gamma | $\mathcal{IG}a(\lambda)$ | $\mathbb{R}_+^\star$ | $\frac{\beta^\alpha}{\Gamma\alpha}x^{-\alpha-1}e^{\frac{-\beta}{x}}$ |
| Laplace | Laplace$(\mu,b)$ | $\mathbb{R}$ | $\frac{1}{2b}e^{-\frac{\lvert x-\mu\rvert}{b}}$ |
| Pareto | $\mathcal{P}a(a,k)$ | $[a,+\infty[$ | $\frac{ka^k}{x^{k+1}}$ |
| Poisson | $\mathcal{P}(\lambda)$ | $\mathbb{N}$ | $e^{-\lambda}\frac{\lambda^x}{x!}$ |
| Uniforme | $\mathcal{U}(\{0,1,\dots,n\})$ | $\{0,1,\dots,n\}$ | $\frac{1}{n+1}$ |
| Uniforme | $\mathcal{U}([a,b])$ | $[a,b]$ | $\frac{1}{b-a}$ |

**Hypothèse**

Dans la suite de ce cours, nous supposerons généralement avoir un modèle statistique identifiable $(E,\mathcal{E},\mathbb{P}_\theta,\theta\in\Theta)$. De plus, nous noterons souvent $\mathbf{X}=(X_1,\dots,X_n)$ pour $n$-échantillon.

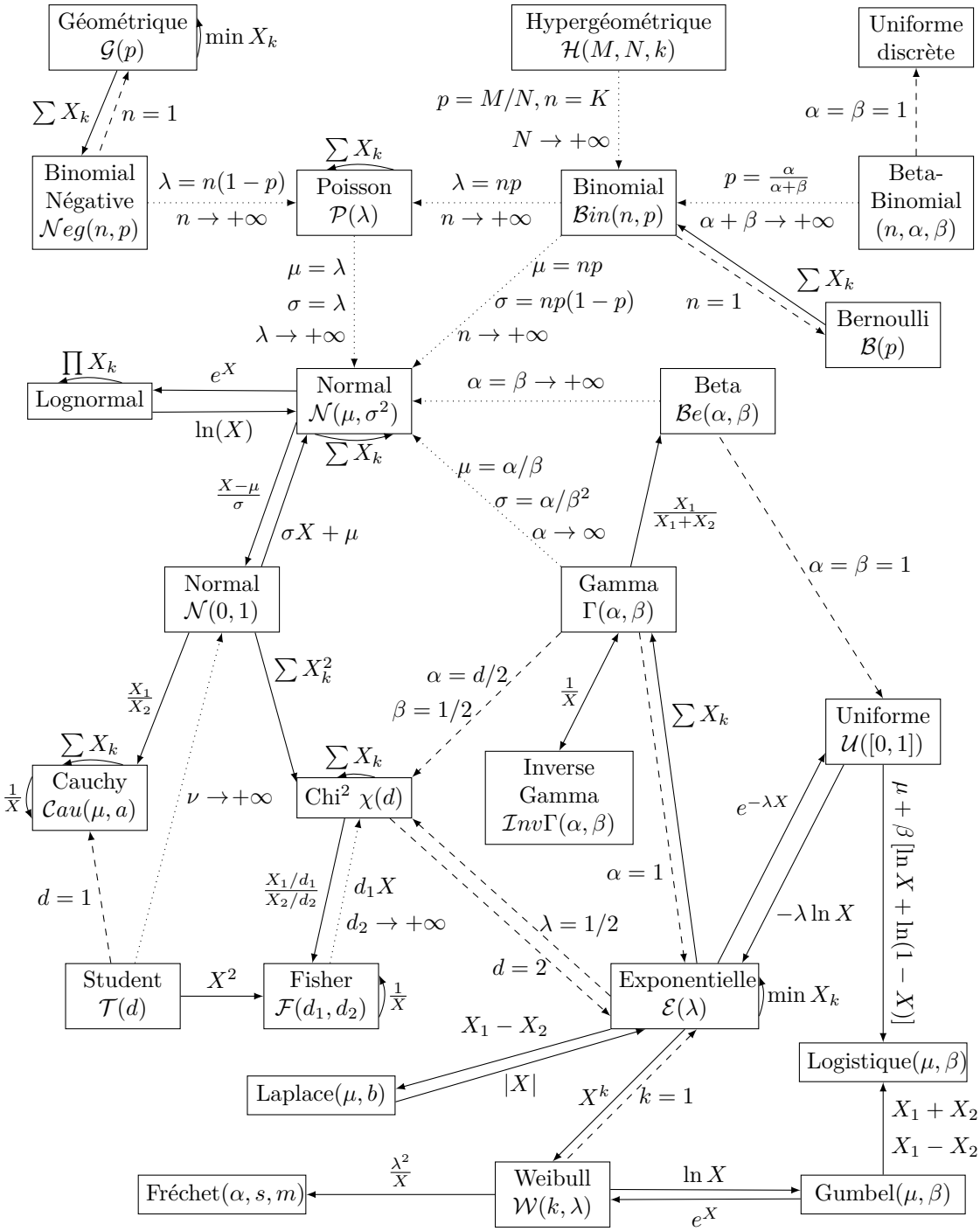



FIGURE 2.1 – Graphe de liens entre différentes lois : les flèches en traits pleins représentent une transformation de la loi du vecteur $\mathbf{X}$ (début de la flèche) vers la variable $Y=f(\mathbf{X})$ (bout de la flèche), les flèches en pointillés longs indiquent un cas particulier et les flèches en pointillés courts un cas limite ou une approximation. Ce graphe est adapté du travail de Leemis (1986).

## 2.4 Solutions des exercices

**Exercices 2.1**

Pour la première, c'est une loi de Bernoulli (donc paramétrique), la deuxième peut être supposée paramétrique (puisque nous sommes dans un simplexe de $\mathbb{R}^d$) et la dernière n'est pas paramétrique si nous

ne connaissons pas le degrés.

**Exercices 2.2**

| Nom | Abréviation | Support | Densité $f(x)$ ou $\mathbb{P}(X=x)$ | |
|---|---|---|---|---|
| Bernoulli | $\mathcal{B}(p)$ | $\{0,1\}$ | $p^x(1-p)^{1-x}$ | $p \in [0,1]$ |
| Bêta | $\mathcal{B}e(a,b)$ | $]0,1[$ | $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ | $(a,b) \in \mathbb{R}_+^\star \times \mathbb{R}_+^\star$ |
| Binomiale | $\mathcal{B}in(n;p)$ | $\{0,1,\ldots,n\}$ | $\binom{n}{x}p^x(1-p)^{n-x}$ | $(n,p) \in \mathbb{N}^\star \times [0;1]$ |
| Binomiale négative | $\mathcal{NB}in(n;p)$ | $\mathbb{N}$ | $\binom{x+n-1}{x}p^n(1-p)^x$ | $(n,p) \in \mathbb{N}^\star \times [0;1]$ |
| Cauchy | $\mathcal{C}au(\mu,a)$ | $\mathbb{R}$ | $\frac{\pi}{1+\left(\frac{x-\mu}{a}\right)^2}$ | $(\mu,a) \in \mathbb{R} \times \mathbb{R}_+^\star$ |
| Exponentielle | $\mathcal{E}(\lambda)$ | $\mathbb{R}^+$ | $\lambda e^{-\lambda x}$ | $\lambda \in \mathbb{R}_+^\star$ |
| Gamma | $\mathcal{G}a(\lambda)$ | $\mathbb{R}_+^\star$ | $\frac{\beta^\alpha}{\Gamma\alpha}x^{\alpha-1}e^{-\beta x}$ | $(\alpha,\beta) \in \mathbb{R}_+^\star \times \mathbb{R}_+^\star$ |
| Gaussienne ou normale | $\mathcal{N}(\mu,\sigma^2)$ | $\mathbb{R}$ | $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ | $(\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+^\star$ |
| Géométrique | $\mathcal{G}éo(p)$ | $\mathbb{N}^\star$ | $(1-p)^{x-1}p$ | $p \in ]0,1]$ |
| Hypergéométrique | $\mathcal{H}(n,p,A)$ | $\{0,1,\ldots,n\}$ | $\frac{\binom{pA}{x}\binom{(1-p)A}{n-x}}{\binom{A}{x}}$ | $A \in \mathbb{N}$<br>$n \in \{0,\ldots,A\}$<br>$p \in \{p \in [0,1] \mid pA \text{ et } (1-p)A \text{ entiers}\}$ |
| Inverse gamma | $\mathcal{IG}a(\alpha,\beta)$ | $\mathbb{R}_+^\star$ | $\frac{\beta^\alpha}{\Gamma\alpha}x^{-\alpha-1}e^{\frac{-\beta}{x}}$ | $(\alpha,\beta) \in \mathbb{R}_+^\star \times \mathbb{R}_+^\star$ |
| Laplace | Laplace$(\mu,b)$ | $\mathbb{R}$ | $\frac{1}{2b}e^{-\frac{\lvert x-\mu\rvert}{b}}$ | $(\mu,b) \in \mathbb{R} \times \mathbb{R}_+^\star$ |
| Pareto | $\mathcal{P}a(a,k)$ | $[a,+\infty[$ | $\frac{ka^k}{x^{k+1}}$ | $(a,k) \in \mathbb{R}_+^\star \times \mathbb{R}_+^\star$ |
| Poisson | $\mathcal{P}(\lambda)$ | $\mathbb{N}$ | $e^{-\lambda}\frac{\lambda^x}{x!}$ | $\mathbb{R}^+$ |
| Uniforme | $\mathcal{U}(\{0,1,\ldots,n\})$ | $\{0,1,\ldots,n\}$ | $\frac{1}{n+1}$ | $n \in \mathbb{N}^\star$ |
| Uniforme | $\mathcal{U}([a,b])$ | $[a,b]$ | $\frac{1}{b-a}$ | $(a,b) \in \left\{(a,b) \in \mathbb{R}^2 \middle\vert a<b\right\}$ |

# Chapitre 3

# Estimation

*"Vous [les statisticien.ne.s] n'êtes jamais contents : avant, nous n'avions pas assez de données et les résultats obtenus n'étaient pas souvent concluants ; maintenant, nous en avons trop et nous ne pouvons pas avoir de résultats dans des temps corrects."*

Discussion avec un biologiste.

## 3.1 Objectifs

Le but de ce chapitre est de comprendre le principe d'estimation et ce qui caractérise un *bon* estimateur. Nous allons donc introduire un certain nombre de notions et présenterons les estimateurs les plus connus.

## 3.2 Rappels/Pré-requis

Dans ce chapitre, nous aurons besoin des notions suivantes.

**Probabilités** : Différentes convergences, loi forte des grands nombres

## 3.3 Introduction

Le but de ce chapitre est d'estimer au mieux une quantité dépendante du paramètre inconnu $\theta$. Pour bien différencier les outils que nous utilisons, nous allons commencer par un certain nombre de notations.

**Définitions 24 (Paramètre d'intérêt)**

Étant donnée une famille de probabilité $(\mathbb{P}_\theta)_{\theta\in\Theta}$ et un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$, nous notons $\theta^\star$ le paramètre *inconnu* de la loi de $X$ ; c'est-à-dire que $\mathbb{P}_X = \mathbb{P}_{\theta^\star}$.
La **quantité à estimer** ou **paramètre d'intérêt** est notée $g(\theta^\star)$ avec $g : \Theta \to \mathbb{R}^d$ où $d \in \mathbb{N}^\star$.

**Remarque**

Souvent, la quantité à estimer est simplement $\theta^\star$ c'est-à-dire pour $g$ valant l'identité.

Maintenant, nous allons parler de l'estimation :

**Définition 25 (Estimateur)**

Étant donné un paramètre d'intérêt $g(\theta^\star)$, nous appelons un **estimateur** de $g(\theta^\star)$ toute application, notée $\widehat{g}$, mesurable en l'observation $X$ et indépendante de $\theta^\star$.

**Remarque**

Tout estimateur est donc de la forme $\widehat{g} = h(X)$ avec $h$ une application mesurable $E \to \mathbb{R}^d$.

**Exemple**

Par exemple, lorsque nous cherchons à estimer $\theta^\star$, nous proposons un estimateur noté $\widehat{\theta}$. Cet estimateur est donc une fonction des observations.

**Attention au piège**

La notion *d'indépendance de* $\theta^\star$ peut paraître étrange mais il arrive parfois que des auteurs proposent un estimateur directement lié à ce dernier. Par exemple, un estimateur de la forme $2\theta^\star$ ne fonctionne pas.

**Remarque**

On peut trouver toute sorte d'estimateurs. Un exemple naïf serait l'estimateur qui renvoie toujours la valeur 0 par exemple. Il faut donc réfléchir à obtenir un *bon* estimateur et en quel sens ?

## 3.4 Consistance d'un estimateur

La première propriété que nous voulons pour un estimateur est que son estimation soit relativement proche de la vraie valeur.

**Définition 26 (Consistance d'un estimateur)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ et un estimateur $\widehat{g}_n = h_n(X_1, \ldots, X_n)$, nous dirons que l'estimateur est

- **consistant** si pour tout $\theta^\star$ de $\Theta$, $\widehat{g}_n$ converge en $\mathbb{P}_{\theta^\star}$-probabilité vers $g(\theta^\star)$ lorsque $n$ tend vers $+\infty$. C'est-à-dire que :

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\|g(\theta^\star) - \widehat{g}_n\|_{+\infty} > \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$

  où $\|\cdot\|_{+\infty}$ est la **norme infinie** sur $\mathbb{R}^d$ c'est-à-dire définie par :

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \|x\|_{+\infty} = \max_{1\le j\le d} |x_j|.$$

- **fortement consistant** si pour tout $\theta^\star$ de $\Theta$ l'estimateur $\widehat{g}_n$ converge $\mathbb{P}_{\theta^\star}$-presque sûrement vers $g(\theta^\star)$ lorsque $n$ tend vers $+\infty$. C'est-à-dire que :

$$\mathbb{P}\left(\lim_{n\to+\infty} \widehat{g}_n = g(\theta^\star)\right) = 1.$$

**Attention au piège**

En dimension finie, toutes les normes sont équivalentes donc nous pouvons choisir celle que nous préférons. En dimension infinie, il faut préciser par rapport à quelle norme nous avons la consistance.

**Remarque**

Par abus de langage, on parle souvent de consistance d'un estimateur là où on devrait plutôt parler de la consistance de la suite d'estimateurs.

**Exemple**

Dans l'exemple de la proportion de naissances masculines dans Paris, il est naturel d'estimer le paramètre $\theta^\star$ par la moyenne empirique

$$\widehat{\theta}_n(X_1, \ldots, X_n) = \overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Alors, par la loi forte des grands nombres, nous savons que c'est un estimateur fortement consistant du paramètre.

D'autres propriétés peuvent être utilisées pour caractériser un bon estimateur, nous les verrons un peu plus tard.

## 3.5 Construction d'un estimateur

Dans cette section, nous présenterons deux types d'estimateurs : celui des moments et celui du maximum de vraisemblance.

### 3.5.1 Méthode des moments

Commençons par un exercice :

**Exercices 3.1**

Nous avons lancé un dé 1000 fois et nous avons obtenus 237 fois un chiffre inférieur ou égal à 3. Pensez-vous que le dé est équilibré ? Pourquoi ?

Modéliser l'expérience : comment d'écrire la situation ? Quelle loi est derrière ? Quel estimateur pourrions-nous proposer ? Quelle est sa loi exacte ? Sa loi asymptotique ? Que nous dit la loi forte des grands nombres ?

Le principe de la méthode des moments repose simplement sur la loi forte des grands nombres. Pour ce faire, nous supposons posséder un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi $\mathbb{P}_{\theta^\star}$ et une fonction $\phi : E \to \mathbb{R}$ mesurable telle que $\phi(X_1)$ possède un moment d'ordre 1.
Pour estimer la valeur $g(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi(X_1)]$, nous nous appuyons sur la moyenne empirique :

$$\widehat{g}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$$

qui, d'après la loi forte des grands nombres, converge presque sûrement vers $g(\theta^\star)$. Par conséquent, la suite d'estimateurs est fortement consistante.

**Remarque**

Cette méthode fonctionne dès que les paramètres d'intérêt peuvent s'exprimer comme des fonctions des moments de la loi des observations $X_i$.

**Exemple**

Nous estimons par exemple l'espérance $\mu_1 = \mathbb{E}[X]$ par la **moyenne empirique** $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Si les variables $X_i$ sont à valeurs réelles et possèdent un moment d'ordre 2 alors nous pouvons estimer $\mu_2 = \mathbb{E}\left[X^2\right]$ par $\widehat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

**Définition 27 (Méthode des moments (MM))**

Étant donnés un $n$-échantillon $(X_1, \dots, X_n)$ de $\mathbb{R}^d$, $q$ fonctions $\phi_1, \dots, \phi_q$ allant respectivement de $\mathbb{R}^d$ dans $\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_q}$ ayant chacune un moment d'ordre 1 ; c'est-à-dire que $\mathbb{E}_{\theta^\star}[\|\phi(X_1)\|]$ est finie et une fonction $\Psi$ allant de $\prod_{j=1}^{q} \mathbb{R}^{n_j}$ dans $\mathbb{R}^m$, la **méthode des moments** ou **(MM)** consiste à estimer $g(\theta^\star) = \Psi(g_1(\theta^\star), \dots, g_q(\theta^\star))$ où $g_j(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi_j(\theta^\star)]$ par

$$\widehat{g}_n = \Psi\left(\frac{1}{n}\sum_{i=1}^{n}\phi_1(X_i), \dots, \frac{1}{n}\sum_{i=1}^{n}\phi_q(X_i)\right).$$

**Exemple**

Étant donné un $n$-échantillon $(X_1, \ldots, X_n)$ de loi admettant un moment d'ordre 2, la variance $\sigma^2 = \mathbb{V}_{\theta^\star}[X_1] = \mathbb{E}_{\theta^\star}\left[X_1^2\right] - \mathbb{E}_{\theta^\star}[X_1]^2$ est estimée par

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

appelée la **variance empirique**.

**Exemple**

Étant donné un $n$-échantillon $(X_1, \ldots, X_n)$ de loi de Poisson $\mathcal{P}(\theta^\star)$ de paramètre $\theta^\star$ alors nous avons deux estimateurs des moments possibles pour $\theta^\star$ :

- Comme $\mathbb{E}_{\theta^\star}[X_1] = \theta^\star$, nous pouvons prendre la moyenne empirique.
- Comme $\mathbb{V}_{\theta^\star}[X_1] = \theta^\star$, nous pouvons prendre la variance empirique.

**Exemple**

Étant donné un $n$-échantillon $(X_1, \ldots, X_n)$ de loi $\mathcal{N}(\mu, \sigma^2)$ alors :

- Si $\sigma^2$ est connu, la méthode des moments suggère d'estimer $\mu$ par la moyenne empirique.
- Si $\mu$ est connue, la méthode des moments suggère d'estimer $\sigma^2$ par

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

- Si $\mu$ et $\sigma^2$ sont tous les deux inconnus alors la méthode des moments suggère d'estimer $\mu$ par la moyenne empirique et $\sigma^2$ par la variance empirique.

### 3.5.2 Méthode du maximum de vraisemblance

**Attention au piège**

Dans cette partie, nous supposons que toutes les lois étudiées sont dominées presque sûrement par une mesure $\mu$. Typiquement, nous aurons :

- la mesure de Lebesgue pour les lois sur $\mathbb{R}^d$.
- la mesure de comptage pour les cas de variables à valeurs dans un espace dénombrable.

Avant d'introduire la méthode, nous avons besoin d'une notation et d'une définition.

**Notation**

Pour tout paramètre $\theta \in \Theta$, nous notons $f_\theta$ la densité de $\mathbb{P}_\theta$ par rapport à la mesure dominante $\mu$. En particulier, dans le cas discret, nous avons pour tout $x \in S$, $f_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}(X = x; \theta)$.

Certains auteurs notent la densité par $p_\theta(\cdot)$ ou $p(\cdot; \theta)$.

**Remarque**

L'idée de la méthode du maximum de vraisemblance est de partir du constat que si nous avons une observation $x$ et une famille de loi $(\mathbb{P}_\theta)_{\theta\in\Theta}$ alors nous pouvons calculer pour chaque $\theta$ la probabilité d'avoir eu l'observation. Par exemple, dans le cas discret, cela revient à calculer $\mathbb{P}_\theta(X = x) = \mathbb{P}(X = x; \theta)$. Une fois ces calculs faits, le $\theta$ le plus *vraisemblable* est simplement celui qui correspond à la plus forte probabilité d'avoir eu l'observation.

**Définition 28 (Vraisemblance)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de densité $f_{\theta^\star}$, nous appelons **vraisemblance du $n$-échantillon $\mathbf{X}$**, notée $V_{\mathbf{X}}$, la fonction :

$$\begin{array}{rcl} V_{\mathbf{X}} : \Theta & \to & \mathbb{R} \\ \theta & \mapsto & V_{\mathbf{X}}(\theta) = f_\theta(\mathbf{X}) \end{array}$$

**Définition 29 (Méthode du maximum de vraisemblance (MMV))**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de densité $f_{\theta^\star}$, la **méthode du maximum de vraisemblance (MMV)** consiste à estimer $g(\theta^\star) = \theta^\star$ par l'élément de $\Theta$ maximisant la vraisemblance $V_{\mathbf{X}}$.

*Si un tel point $\widehat{\theta}_n$ existe,* on l'appelle **l'estimateur du maximum de vraisemblance**.

**Attention au piège**

Nous pouvons faire trois remarques importantes :

- L'estimateur du maximum de vraisemblance n'existe pas toujours : c'est-à-dire qu'il n'y a pas toujours un point qui maximise la vraisemblance.
- S'il existe, il n'est pas obligatoirement unique : c'est-à-dire qu'il peut y avoir plusieurs points qui maximisent la vraisemblance.
- S'il existe et qu'il est unique, il n'est pas obligatoirement calculable : nous pouvons juste étudier la vraisemblance pour voir qu'elle doit admettre un maximum sans pouvoir trouver une formule explicite.

**Proposition 2**

Dans le cas d'un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune $\mathbb{P}_{\theta^\star}$, la vraisemblance prend la forme produit :

$$V_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

La conséquence est que nous préférons souvent maximiser la **log-vraisemblance** :

$$\log V_{\mathbf{X}}(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

**Preuve**

Le produit vient de l'indépendance des variables $(X_i)_{i\in\{1,\ldots,n\}}$ et la formule du fait que nous avons une simple densité.

**Remarque**

L'inconvénient du maximum de vraisemblance est que nous n'estimons que le paramètre $\theta^\star$ mais pas des fonctions de celui-ci. En revanche, nous verrons par la suite que certains estimateurs du maximum de vraisemblance ont des meilleures propriétés que l'estimateur des moments.

Un calcul assez simple permet d'obtenir les estimateurs suivants :

**Exercices 3.2**

Dans le cas d'un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi de Bernoulli $\mathcal{B}(\theta^\star)$ ou de Poisson $\mathcal{P}(\theta^\star)$, l'estimateur du maximum de vraisemblance est le même que celui des moments.

De même pour les modèles gaussiens $\mathcal{N}(\mu, \sigma^2)$ que nous connaissions $\mu$ ou $\sigma^2$ voir aucun des deux.

Un exemple assez intéressant d'estimateur du maximum de vraisemblance est le suivant :

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de même loi $\mathcal{U}([0, \theta^\star])$ avec $\theta^\star \in \mathbb{R}_+^\star$, nous avons :

$$\begin{aligned} V_{\mathbf{X}}(\theta) &= \prod_{i=1}^{n} \left( \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \right) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(X_i) \\ &= \theta^{-n} \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta) \text{ où } X_{(n)} = \max_{1 \leq i \leq n} X_i. \end{aligned}$$

Nous voyons que pour tout $n \in \mathbb{N}^\star$, la fonction $\theta \mapsto \theta^{-n}$ est décroissante et positive mais l'indicatrice $\theta \mapsto \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta)$ est nulle pour tout $\theta$ strictement plus petit que $X_{(n)}$. La valeur maximale de la fonction est donc atteinte pour $\widehat{\theta}_n = X_{(n)}$. Notons, que c'est bien $n$ qui est fixé et $\theta$ qui tend vers l'infini (si on commence à regarder dans l'autre sens, on pourrait croire que $0 < \theta < 1$ pose problème).

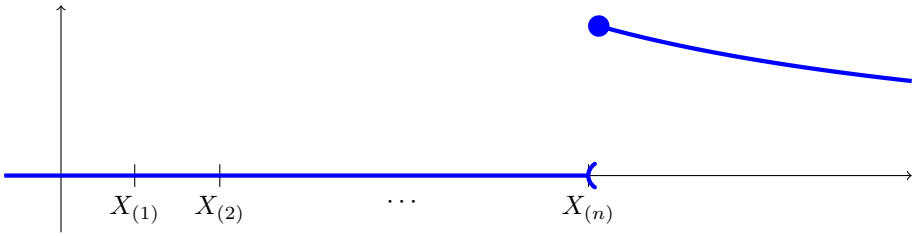



FIGURE 3.1 – Représentation en bleu de la fonction $\theta \mapsto \theta^{-n} \mathbb{1}_{\left[X_{(n)}, +\infty\right[}(\theta)$.

Enfin, voici un contre-exemple pour se rappeler qu'il n'est pas toujours possible de calculer le maximum de vraisemblance :

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de même loi $\mathcal{C}au(\theta^\star, 1)$ c'est-à-dire de densité

$$f_\theta(x) = \frac{1}{\pi} \times \frac{1}{1 + (x - \theta)^2}.$$

Le calcul de la log-vraisemblance est alors :

$$\log V_{\mathbf{X}}(\theta) = -n \log \pi - \sum_{i=1}^{n} \log \left[ 1 + (X_i - \theta)^2 \right].$$

Cette fonction est définie sur $\mathbb{R}$, dérivable en $\theta$, tend vers $-\infty$ lorsque $\theta$ tend vers $\pm\infty$ donc elle admet au moins un maximum qui est un point où la dérivée s'annule. En revanche, le calcul n'est pas explicite car la dérivée est une somme de fractions qui, une fois réduite au même dénominateur, est la somme de polynôme de degrés $2n-1$. Il y a donc beaucoup de racines qui n'ont pas forcément de formules explicites (exceptées dans certains rares cas).

## 3.6 Qualité d'un estimateur

Dans cette partie, nous discutons des différentes propriétés que peut avoir un estimateur. Nous avons déjà vu l'indispensable consistance d'un estimateur qui permet de garantir que ce dernier sera *suffisamment proche* lorsque nous aurons suffisamment d'observations mais nous pouvons être intéressés par la vitesse à laquelle il va converger ou encore par la qualité non-asymptotique de son estimation.

### 3.6.1 Loi asymptotique

Comme nous allons parler de convergence en loi, nous commençons par un petit rappel et un théorème que nous utiliserons.

**Définition 30 (Convergence en loi)**

Étant donnée une suite $(X_n)_{n\in\mathbb{N}^\star}$ de variables aléatoires d'un même espace de probabilité $(E,\mathcal{E},\mathbb{P})$, nous disons que la suite $(X_n)_{n\in\mathbb{N}^\star}$ **converge en loi** vers la variable $X$ si, pour toute fonction $\varphi$ *continue bornée* sur $E$ à valeurs dans $\mathbb{R}$, nous avons :

$$\lim_{n\to+\infty}\mathbb{E}\left[\varphi(X_n)\right]=\mathbb{E}\left[\varphi(X)\right].$$

Nous notons alors :

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X.$$

**Remarque**

Dans le cas de $\mathbb{R}^d$, nous pouvons montrer qu'il suffit de prendre des fonctions dans un espace dont l'adhérence englobe les fonctions continues à support compact sur $\mathbb{R}^d$ (voir par exemple le polycopié de Le Gall (2006), proposition 10.3.3). En particulier, nous pourrons être intéressés par prendre des fonctions lipschitzienne (voir la preuve du lemme 5 de Slutsky).

Dans notre cas, nous allons souvent utiliser la proposition suivante :

**Proposition 3 (Cas de variables réelles)**

Étant données une suite $(X_n)_{n\in\mathbb{N}^\star}$ de variables aléatoires d'un même espace de probabilité $(E,\mathcal{E},\mathbb{P})$ et une variable $X$ sur le même espace, nous avons l'équivalence des propositions suivantes :

(1) $(X_n)_{n\in\mathbb{N}^\star}$ converge en loi vers $X$ ;

(2) Pour tout $x\in\mathbb{R}$ où $F_X$ est continue (c'est-à-dire que $\mathbb{P}(X=x)=0$), nous avons la convergence simple :
$$\lim_{n\to+\infty}F_{X_n}(x)=F_X(x)$$
où $F_X(x)=\mathbb{P}(X\leq x)$ pour tout $x\in\mathbb{R}$.

**Preuve**

Une preuve complète peut être trouvée dans le polycopié de Le Gall (2006), proposition 10.3.2 et sa conséquence.

Dans un premier temps, nous allons nous concentrer sur l'estimation de $g(\theta^\star)=\theta^\star$. Le principe de normalité asymptotique va plus loin que la consistance puisque nous nous intéressons à la vitesse de convergence de $\widehat{\theta}_n$ vers $\theta^\star$.

**Définition 31 (Vitesse de convergence)**

Étant donné un estimateur $\widehat{g}_n$ du paramètre $g(\theta^\star)$ et une suite $(a_n)_{n\in\mathbb{N}^\star}$ tendant vers $+\infty$, nous disons que $(a_n)_{n\in\mathbb{N}^\star}$ est l'ordre de la **vitesse de convergence** de l'estimateur $\widehat{g}_n$ s'il

existe une variable aléatoire $Y$ de loi non triviale indépendante de $\widehat{g}_n$ telle que

$$a_n\left(\widehat{g}_n - g\left(\theta^\star\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Y.$$

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune admettant un moment d'ordre 2, d'espérance $\mu^\star$ et de matrice de variance covariance $\Sigma^\star$ alors, d'après le théorème de limite centrale, nous avons :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu^\star\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma^\star\right).$$

Sa vitesse de convergence est donc $\sqrt{n}$.

**Exercices 3.3**

Dans le cas d'un $n$-échantillon suivant une loi de Poisson $\mathcal{P}(\theta^\star)$, nous avons déjà qu'un estimateur possible est également $\widehat{\theta}_n = \overline{\mathbf{X}}_n$ et nous aurons dans ce cas :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \theta^\star\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \theta^\star\right).$$

Certains auteurs utilisent parfois la notion d'équivalence asymptotique :

$$\overline{\mathbf{X}}_n \underset{+\infty}{\overset{\mathcal{L}}{\sim}} \mathcal{N}\left(\theta^\star, \frac{\theta^\star}{n}\right).$$

**Définition 32 (Asymptotiquement normal)**

Un estimateur $\widehat{g}_n$ est dit **asymptotiquement normal** s'il converge vers une loi asymptotique normal à une vitesse de l'ordre de $\sqrt{n}$. Donc il faut vérifier les trois conditions suivantes :

- La vitesse de convergence est en $\sqrt{n}$.
- La convergence a lieu en loi.
- La loi limite est une loi normale non triviale.

**Remarque**

Dans les cas classiques et notamment pour les estimateurs des moments, ce sera bien souvent le cas.

**Contre-exemple**

Si $X_1, \ldots, X_n$ est un $n$-échantillon suivant une loi $\mathcal{U}([0, \theta^\star])$ alors, nous avons vu que l'estimateur du maximum de vraisemblance était $\widehat{\theta}_n = \max\{X_1, \ldots, X_n\} = X_{(n)}$ et calculons la loi de $n\left(\theta^\star - \widehat{\theta}_n\right)$. Comme, $\mathbb{P}_{\theta^\star}$-presque sûrement, $\theta^\star$ est plus grand que toutes les variables $X_i$, nous avons que $n\left(\theta^\star - \widehat{\theta}_n\right)$ est $\mathbb{P}_{\theta^\star}$-presque sûrement positif. Donc, pour tout $x \in \mathbb{R}^+$, nous avons :

$$\begin{aligned}
F_{n\left(\theta^\star - \widehat{\theta}_n\right)}(x) &= \mathbb{P}\left(n\left(\theta^\star - \widehat{\theta}_n\right) \leq x\right) \\
&= \mathbb{P}\left(\theta^\star - \widehat{\theta}_n \leq \frac{x}{n}\right) \\
&= \mathbb{P}\left(\theta^\star - \frac{x}{n} \leq \widehat{\theta}_n\right) \\
&= 1 - \mathbb{P}\left(\theta^\star - \frac{x}{n} > \widehat{\theta}_n\right) \\
&= 1 - \mathbb{P}\left(\max\{X_1, \ldots, X_n\} < \theta^\star - \frac{x}{n}\right)
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \mathbb{P}\left(\left(X_1 < \theta^\star - \frac{x}{n}\right) \text{ et } \left(X_2 < \theta^\star - \frac{x}{n}\right) \text{ et } \ldots \text{ et } \left(X_n < \theta^\star - \frac{x}{n}\right)\right) \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}\left(X_i < \theta^\star - \frac{x}{n}\right) \\
&= 1 - \left[\prod_{i=1}^{n}\left(\frac{\theta^\star - \frac{x}{n}}{\theta^\star}\right)\right] \mathbb{1}_{\{\theta^\star - \frac{x}{n} > 0\}} \\
&= 1 - \left(1 - \frac{x}{n\theta^\star}\right)^n \mathbb{1}_{]-\infty; n\theta^\star]}(x) \\
\underset{n\to+\infty}{\longrightarrow} \quad & 1 - e^{-\frac{1}{\theta^\star}x}
\end{aligned}
$$

et nous reconnaissons la fonction de répartition de la loi exponentielle de paramètre $1/\theta^\star$. Par conséquent, la loi asymptotique de $n\left(\theta^\star - \widehat{\theta}_n\right)$ est $\mathcal{E}\left(\frac{1}{\theta^\star}\right)$.

**Remarque**

Dans le contre-exemple précédent, nous voyons que la vitesse est en $n$ par opposition à l'estimateur des moments qui converge à la vitesse $\sqrt{n}$.

Dans cette partie, la proposition la plus importante est celle de la méthode Delta :

**Théorème 4 (Méthode Delta)**

Étant données une suite $(U_n)_{n\in\mathbb{N}^\star}$ de vecteurs aléatoires de $\mathbb{R}^d$ avec $d \in \mathbb{N}^\star$, une suite déterministe de réels $(a_n)_{n\in\mathbb{N}^\star}$ et une application $\ell : \mathbb{R}^d \to \mathbb{R}^p$ avec $p \in \mathbb{N}^\star$ telles que :

- $a_n$ tend vers $+\infty$.
- Il existe $U^\star \in \mathbb{R}^d$ un vecteur déterministe et $V \in \mathbb{R}^d$ un vecteur aléatoire tels que

$$a_n\left(U_n - U^\star\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} V.$$

- $\ell$ est différentiable en $U^\star$ de différentielle notée $D\ell\left(U^\star\right) \in \mathcal{M}_{p\times d}(\mathbb{R})$.

Alors, nous avons :

$$a_n\left(\ell(U_n) - \ell\left(U^\star\right)\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} D\ell\left(U^\star\right) V.$$

Pour la démonstration de ce théorème, nous avons besoin de rappeler le lemme de Slutsky.

**Lemme 5 (Slutsky)**

Étant données $(X_n)_{n\in\mathbb{N}^\star}$ et $(Y_n)_{n\in\mathbb{N}^\star}$ deux suites de variables aléatoires prenant leurs valeurs respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^p$ avec $d, p \in \mathbb{N}^\star$ éventuellement différents telles que $X_n$ converge en loi vers un vecteur aléatoire $X$ et $Y_n$ converge en probabilité vers un vecteur déterministe $C$ alors :

$$(X_n, Y_n) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} (X, C).$$

**Remarque**

Le plus souvent, nous appliquons à cette convergence jointe une fonction $g$ ce qui permet de déduire :

$$g\left(X_n, Y_n\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} g(X, C).$$

Les fonctions les plus utilisées sont par exemple la somme (si $d = p$), le produit (si $p = 1$) ou la division (si $C \neq 0$ et $p = 1$).

**Preuve**

Pour démontrer le lemme de Slutsky, nous prenons une fonction $\varphi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ lipshitzienne et bornée de constante de lipschitz $L$ et $M$ un majorant de la valeur absolue de la fonction. Commençons par utiliser l'inégalité triangulaire pour montrer que :

$$|\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X, C)]| \leq \underbrace{|\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]|}_{=(1)} + \underbrace{|\mathbb{E}[\varphi(X_n, C)] - \mathbb{E}[\varphi(X, C)]|}_{=(2)}.$$

Comme la fonction $x \mapsto \varphi(x, C)$ est bornée et que la suite $(X_n)_{n\in\mathbb{N}^\star}$ tend en loi vers $X$ alors, par le théorème du porte-manteau, le second terme (2) tend vers 0.

Pour le premier terme (1), nous prenons $\varepsilon > 0$ et nous utilisons le fait que la fonction soit lipshitzienne :

$$\begin{aligned}
(1) &= |\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]| \\
&\leq \mathbb{E}[|\varphi(X_n, Y_n) - \varphi(X_n, C)|] \\
&\leq \mathbb{E}\left[|\varphi(X_n, Y_n) - \varphi(X_n, C)| \,\middle|\, \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[|\varphi(X_n, Y_n) - \varphi(X_n, C)| \,\middle|\, \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq \mathbb{E}\left[|\varphi(X_n, Y_n)| + |\varphi(X_n, C)| \,\middle|\, \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[L\|Y_n - C\| \,\middle|\, \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq \mathbb{E}\left[2M \,\middle|\, \|Y_n - C\| > \varepsilon\right] \mathbb{P}(\|Y_n - C\| > \varepsilon) \\
&\quad + \mathbb{E}\left[L\varepsilon \,\middle|\, \|Y_n - C\| \leq \varepsilon\right] \mathbb{P}(\|Y_n - C\| \leq \varepsilon) \\
&\leq 2M\mathbb{P}(\|Y_n - C\| > \varepsilon) + L\varepsilon\mathbb{P}(\|Y_n - C\| \leq \varepsilon)
\end{aligned}$$

Comme $(Y_n)_{n\in\mathbb{N}^\star}$ converge en probabilité vers $C$, nous avons pour tout $\varepsilon > 0$ :

$$\limsup_{n\to+\infty} |\mathbb{E}[\varphi(X_n, Y_n)] - \mathbb{E}[\varphi(X_n, C)]| \leq L\varepsilon$$

ce qui permet de conclure.

À l'aide de ce lemme, nous pouvons maintenant démontrer la méthode Delta :

**Preuve**

Avant de commencer, rappelons que la définition de la différentiabilité est la suivante :

$$\forall M > 0,\ \forall \varepsilon > 0,\ \exists \delta > 0,$$
$$\forall U' \in \mathbb{R}^d, \|U' - U^\star\| \leq \delta \Rightarrow \|\ell(U') - \ell(U^\star) - D\ell(U^\star)(U' - U^\star)\| \leq \frac{M}{\varepsilon}\|U' - U^\star\|.$$

Commençons par décomposer la variable étudiée pour introduire la notion de différentiabilité :

$$a_n(\ell(U_n) - \ell(U^\star)) = \underbrace{a_n(\ell(U_n) - \ell(U^\star)) - D\ell(U^\star)\, a_n(U_n - U^\star)}_{=(1)} + \underbrace{D\ell(U^\star)\, a_n(U_n - U^\star)}_{=(2)}$$

et montrons que le premier terme converge vers 0 et en probabilité. Pour cela, nous fixons $\varepsilon > 0$ et $M > 0$ et nous choisissons $\delta > 0$ tel que

$$\begin{aligned}
\mathbb{P}(\|(1)\| > \varepsilon) &= \mathbb{P}(\|a_n(\ell(U_n) - \ell(U^\star)) - D\ell(U^\star)\, a_n(U_n - U^\star)\| > \varepsilon) \\
&= \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon) \\
&= \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon | \|U_n - U^\star\| \leq \delta)\,\mathbb{P}(\|U_n - U^\star\| \leq \delta) \\
&\quad + \mathbb{P}(a_n\|\ell(U_n) - \ell(U^\star) - D\ell(U^\star)(U_n - U^\star)\| > \varepsilon | \|U_n - U^\star\| > \delta)\,\mathbb{P}(\|U_n - U^\star\| > \delta) \\
&\leq \mathbb{P}(a_n\|U_n - U^\star\| > M | \|U_n - U^\star\| \leq \delta) + \mathbb{P}(a_n\|U_n - U^\star\| > a_n\delta) \\
&\leq \mathbb{P}(a_n\|U_n - U^\star\| > M) + \mathbb{P}(a_n\|U_n - U^\star\| > a_n\delta) \\
&\leq 2\mathbb{P}(a_n\|U_n - U^\star\| > \min(M, a_n\delta)).
\end{aligned}$$

Ceci étant vrai pour tout $n \in \mathbb{N}^\star$ et comme $a_n$ tend vers l'infini, nous avons pour $n$ assez grand :

$$\begin{aligned}\limsup_{n\to+\infty} \mathbb{P}\left(\|(1)\| > \varepsilon\right) &\leq \limsup_{n\to+\infty} 2\mathbb{P}\left(a_n \|U_n - U^\star\| > M\right)) \\ &\leq 2\mathbb{P}\left(\|V\| \geq M\right)\end{aligned}$$

car $a_n\left(U_n - U^\star\right)$ converge en loi vers $V$ et par le lemme de Porte-Manteau. Or, ceci est vrai pour tout $M$ donc nous pouvons faire tendre $M$ vers $+\infty$ et nous avons donc la convergence en probabilité vers 0.

Enfin, par les hypothèses, nous savons que la deuxième partie (2) converge en loi vers $D\ell\left(U^\star\right)V$ donc, en utilisant le lemme de Slutsky sur la somme, nous avons le résultat.

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ suivant une loi de Poisson $\mathcal{P}\left(\theta^\star\right)$, nous avons vu qu'un estimateur possible de $\theta^\star$ est la variance empirique $\widehat{\theta}_{n,2} = \overline{X_n^2} - \overline{\mathbf{X}}_n^2$. Par la méthode Delta, nous montrons que :

$$\sqrt{n}\left(\widehat{\theta}_{n,2} - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, 2\theta^{\star 2} + \theta^\star\right).$$

**Exercices 3.4**

Montrer le résultat de l'exemple.

**Corollaire 6 (Méthode Delta et estimateur des moments)**

Étant donnés un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de vecteurs de $\mathbb{R}^d$, de loi commune admettant un moment d'ordre 2, d'espérance $\mu$ et de matrice de variance-covariance $\Sigma$ et $\ell$ une fonction différentiable en $\mu$ de différentielle notée $D\ell(\mu)$ alors

$$\sqrt{n}\left(\ell\left(\overline{\mathbf{X}}_n\right) - \ell(\mu)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, D\ell(\mu)\Sigma D\ell(\mu)^T\right).$$

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ suivant une loi exponentielle $\mathcal{E}\left(\theta^\star\right)$, nous cherchons estimer la variance $1/\theta^{\star 2}$ par l'estimateur des moments

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \overline{\mathbf{X}^2} - \overline{\mathbf{X}}^2 = \ell\left(\overline{\mathbf{X}^2}, \overline{\mathbf{X}}\right) \text{ avec } \ell(x,y) = x - y^2.$$

Or, nous savons que, pour tout $m \in \mathbb{N}^\star$, nous avons $\mathbb{E}_{\theta^\star}\left[X^m\right] = \frac{m!}{\theta^{\star m}}$ et ainsi :

$$\begin{aligned}\mathbb{V}_{\theta^\star}\left[X_1\right] &= \mathbb{E}_{\theta^\star}\left[X_1^2\right] - \mathbb{E}_{\theta^\star}\left[X_1\right]^2 \\ &= \frac{2!}{\theta^{\star 2}} - \left(\frac{1!}{\theta^{\star 1}}\right)^2 \\ &= \frac{1}{\theta^{\star 2}}, \\ \mathrm{Cov}_{\theta^\star}\left(X_1, X_1^2\right) &= \mathbb{E}_{\theta^\star}\left[X_1^3\right] - \mathbb{E}_{\theta^\star}\left[X_1\right]\mathbb{E}_{\theta^\star}\left[X_1^2\right] \\ &= \frac{3!}{\theta^{\star 3}} - \frac{1!}{\theta^{\star 1}} \times \frac{2!}{\theta^{\star 2}} \\ &= \frac{4}{\theta^{\star 3}}, \\ \mathbb{V}_{\theta^\star}\left[X_1^2\right] &= \mathbb{E}_{\theta^\star}\left[X_1^4\right] - \mathbb{E}_{\theta^\star}\left[X_1^2\right]^2\end{aligned}$$

$$= \frac{4!}{\theta^{\star 4}} - \left(\frac{2!}{\theta^{\star 2}}\right)^2$$
$$= \frac{20}{\theta^{\star 4}}.$$

Par conséquent, nous avons par le théorème de limite centrale :

$$\sqrt{n}\left(\begin{pmatrix}\overline{\mathbf{X}^2}\\ \overline{\mathbf{X}}\end{pmatrix} - \begin{pmatrix}\frac{2}{\theta^{\star 2}}\\ \frac{1}{\theta^\star}\end{pmatrix}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{20}{\theta^{\star 4}} & \frac{4}{\theta^{\star 3}}\\ \frac{4}{\theta^{\star 3}} & \frac{1}{\theta^{\star 2}}\end{pmatrix}\right).$$

Pour appliquer la méthode Delta, nous devons calculer la différentielle :

$$D\ell(x,y) = (1, -2y).$$

De plus, nous avons :

$$\begin{aligned} D\ell\left(\frac{2}{\theta^{\star 2}}, \frac{1}{\theta^\star}\right)\Sigma D\ell\left(\frac{2}{\theta^{\star 2}}, \frac{1}{\theta^\star}\right)^T &= \begin{pmatrix}1 & -\frac{2}{\theta^\star}\end{pmatrix}\begin{pmatrix}\frac{20}{\theta^{\star 4}} & \frac{4}{\theta^{\star 3}}\\ \frac{4}{\theta^{\star 3}} & \frac{1}{\theta^{\star 2}}\end{pmatrix}\begin{pmatrix}1\\ -\frac{2}{\theta^\star}\end{pmatrix}\\ &= \begin{pmatrix}1 & -\frac{2}{\theta^\star}\end{pmatrix}\begin{pmatrix}\frac{12}{\theta^{\star 4}}\\ \frac{2}{\theta^{\star 3}}\end{pmatrix}\\ &= \frac{8}{\theta^{\star 4}}. \end{aligned}$$

Par la méthode Delta, nous avons donc :

$$\sqrt{n}\left(\widehat{s}_n^2 - \frac{1}{\theta^{\star 2}}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{8}{\theta^{\star 4}}\right)$$

c'est-à-dire que l'estimateur est asymptotiquement normal.

### 3.6.2 Estimation sans biais

Le principe d'estimation est de vérifier si on peut *espérer* avoir une estimation parfaite en moyenne.

**Définitions 33 (Biais)**

Étant donnés une observation $X$ et un estimateur $\widehat{g} = h(X)$ de $g(\theta^\star)$ ayant un moment d'ordre 1 ; c'est-à-dire tel que $\mathbb{E}_{\theta^\star}[\|\widehat{g}\|] < +\infty$, pour tout $\theta^\star \in \Theta$, nous appelons **biais** la fonction $b : \Theta \to \mathbb{R}^d$ définie pour tout $\theta^\star \in \Theta$ par :

$$b(\theta^\star) = \mathbb{E}_{\theta^\star}[\widehat{g}] - g(\theta^\star).$$

L'estimateur $\widehat{g}$ sera dit **sans biais** si $b(\theta^\star) = 0$ pour tout $\theta^\star \in \Theta$ ; ou encore si l'espérance $\mathbb{E}_{\theta^\star}[\widehat{g}] = g(\theta^\star)$.

**Exemple**

Lorsqu'un moment d'ordre 1 existe, la moyenne empirique est toujours un estimateur sans biais ; en revanche, celui de la variance est biaisé.

**Astuce algorithmique**

Généralement, nous trouvons l'estimateur non biaisé de la variance dans les logiciels par défaut. De plus en plus, nous avons aussi l'estimateur classique.

**Exemple**

Dans le cas d'un $n$-échantillon de loi $\mathcal{U}([0,\theta^\star])$, l'estimateur des moments est sans biais mais celui du maximum de vraisemblance est biaisé (il sous-estime la vraie valeur). Néanmoins, nous préférons le second car il converge plus vite vers la bonne valeur.

**Contre-exemple**

Dans le cas d'un modèle $\mathcal{N}(\theta^\star, 1)$ où $\theta^\star \in \mathbb{R}$, l'estimateur $\widehat{g} = 0$ est nul pour $\theta^\star = 0$ mais différent de zéro dès que $\theta^\star \neq 0$, il est donc biaisé puisqu'il ne vérifie pas la condition pour tout $\theta^\star \in \mathbb{R}$.

**Remarque**

Nous pouvons faire trois remarques :

1. Le caractère non-biaisé d'un estimateur doit être vérifié pour tout $\theta^\star \in \Theta$ ; c'est-à-dire que, quelque soit la quantité à estimer, le biais doit être nul (pas juste pour des valeurs).
2. Le critère est non-asymptotique.
3. Le critère peut être discutable : à un moment, le biais était jugé très important (c'est pour cela que beaucoup de logiciels proposent des versions non biaisées par défaut de certains estimateurs). Néanmoins, il vaut toujours mieux un estimateur qui converge plus vite vers la bonne valeur qu'un estimateur non biaisé. De même, un estimateur non biaisé qui possède une forte variance a parfois une plus forte probabilité de fournir une estimation très loin de la quantité à estimer qu'un paramètre biaisé avec une faible variance.

Par rapport à la troisième remarque, nous préférons nous contenter de la version asymptotique.

**Définition 34 (Asymptotiquement sans biais)**

Étant donnée une observation $X$ et un estimateur $\widehat{g} = h(X)$ de $g(\theta^\star)$ ayant un moment d'ordre 1 ; c'est-à-dire tel que $\mathbb{E}_{\theta^\star}[\|\widehat{g}\|] < +\infty$, pour tout $\theta^\star \in \Theta$, l'estimateur $\widehat{g}_n$ sera dit **asymptotiquement sans biais** si pour tout $\theta^\star \in \Theta$ :

$$\lim_{n\to+\infty} \mathbb{E}_{\theta^\star}[\widehat{g}_n] = g(\theta^\star).$$

**Point méthode**

Pour le biais, il suffit donc de procéder en deux étapes :

1. Nous calculons $\mathbb{E}_{\theta^\star}[\widehat{g}_n]$ :
   - S'il vaut $g(\theta^\star)$ pour tout $n \in \mathbb{N}^\star$, l'estimateur est sans biais donc asymptotiquement sans biais aussi.
2. Sinon, nous regardons la limite :
   - Si elle vaut $g(\theta^\star)$, l'estimateur est asymptotiquement sans biais aussi.
   - Sinon, il peut y avoir des valeurs de $\theta^\star \in \Theta$ pour lesquelles, l'estimateur n'est pas consistant.

### 3.6.3 Estimation optimale

Dans cette partie, nous allons parler du risque d'un estimateur qui représente à quel point un estimateur peut s'écarter ou non de la valeur qu'il estime et nous essayerons de minimiser ce risque.

#### Risque d'un estimateur

Dans la partie précédente, nous nous sommes intéressé · e · s au fait que les estimations soient, en espérance, égale à la vraie valeur (statistique de position). Toutefois, nous pouvons être en espérance proche mais avec une forte dispersion (statistique de dispersion). Dans cette partie, nous étudions ainsi la variabilité des estimations et nous commençons par introduire le risque.

**Définition 35 (Risque d'un estimateur)**

Étant donné un estimateur $\widehat{g}$ de $g(\theta^\star)$, nous associons son **risque (quadratique)** défini comme l'application suivante :

$$\begin{array}{rcl} R : \Theta & \to & [0, +\infty] \\ \theta^\star & \mapsto & \mathbb{E}_\theta \left[ \left( \widehat{g} - g(\theta) \right)^2 \right] \end{array}$$

**Remarque**

Si plusieurs estimateurs sont étudiés en même temps, il est préférable d'indexer la fonction $R$ par le nom de l'estimateur.

Un estimateur sera d'autant meilleur que son risque est faible sur $\Theta$.

**Proposition 7**

Étant donné un estimateur $\widehat{g}$ de $g(\theta^\star)$, nous avons la relation suivante :

$$R(\theta) = \left( \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta) \right)^2 + \mathbb{E}_\theta \left[ \left( \widehat{g} - \mathbb{E}_\theta \left[ \widehat{g} \right] \right)^2 \right] = b(\theta)^2 + \mathbb{V}_\theta \left[ \widehat{g} \right].$$

**Preuve**

La démonstration consiste simplement à introduire l'espérance :

$$\begin{array}{rcl} R(\theta) & = & \mathbb{E}_\theta \left[ \left( \widehat{g} - g(\theta) \right)^2 \right] \\ & = & \mathbb{E}_\theta \left[ \left( \widehat{g} - \mathbb{E}_\theta \left[ \widehat{g} \right] + \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta^\star) \right)^2 \right] \\ & = & \mathbb{E}_\theta \left[ \left( \widehat{g} - \mathbb{E}_\theta \left[ \widehat{g} \right] \right)^2 \right] + 2\mathbb{E}_\theta \left[ \left( \widehat{g} - \mathbb{E}_\theta \left[ \widehat{g} \right] \right) \left( \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta^\star) \right) \right] + \mathbb{E}_\theta \left[ \left( \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta^\star) \right)^2 \right] \\ & = & \mathbb{V}_\theta \left[ \widehat{g} \right] + 2\mathbb{E}_\theta \left[ \widehat{g} - \mathbb{E}_\theta \left[ \widehat{g} \right] \right] \left( \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta^\star) \right) + b(\theta)^2 \\ & = & \mathbb{V}_\theta \left[ \widehat{g} \right] + 2 \underbrace{\left( \mathbb{E}_\theta \left[ \widehat{g} \right] - \mathbb{E}_\theta \left[ \widehat{g} \right] \right)}_{=0} \left( \mathbb{E}_\theta \left[ \widehat{g} \right] - g(\theta^\star) \right) + b(\theta)^2 \\ & = & b(\theta)^2 + \mathbb{V}_\theta \left[ \widehat{g} \right] \end{array}$$

**Corollaire 8**

Étant donné un estimateur $\widehat{g}$ sans biais de $g(\theta^\star)$, le risque est égal à la variance de l'estimateur.

Dans le cadre des estimateurs non biaisés, nous nous intéressons donc à ceux qui minimisent uniformément la variance.

**Définition 36 (Estimateur UMVU)**

Étant donnée une collection $\mathcal{G}$ d'estimateurs non biaisés de $g(\theta^\star)$, c'est-à-dire un ensemble d'applications mesurable en l'observation $X$ à valeurs dans l'espace d'arrivée de $g(\theta^\star)$, nous appelons **estimateur UMVU** pour (**Uniform Minimum Variance Unbiaised**) tout estimateur $\widehat{g}$ qui minimise la variance de façon uniforme c'est-à-dire vérifiant :

$$\forall \theta \in \Theta, \forall \widehat{h} \in \mathcal{G}, \mathbb{V}_{\theta^\star} \left[ \widehat{g} \right] \leq \mathbb{V}_{\theta^\star} \left[ \widehat{h} \right].$$

### Information de Fisher

Dans cette partie et les suivantes, nous avons besoin de restreindre l'ensemble des lois étudiées.

**Définition 37 (Modèle paramétrique régulier)**

Étant donné un modèle paramétrique $(f_\theta)_{\theta\in\Theta}$, nous dirons qu'il est **régulier** s'il vérifie les trois conditions suivantes :

1. $\Theta$ est un ouvert de $\mathbb{R}^p$.
2. Le support des densités $f_\theta$, c'est-à-dire l'ensemble $\left\{x\in\mathbb{R}^d \middle| f_\theta(x)>0\right\}$, est indépendant de $\theta$ et la fonction $\theta\mapsto f_\theta$ est dérivable sur $\Theta$ sauf éventuellement pour un nombre fini de points.
3. Pour toute application $g$ mesurable en $X$ dont l'espérance du module est finie sous la loi $f_\theta$ pour tout $\theta$, c'est-à-dire que $\mathbb{E}_\theta\left[\|g(X)\|\right]<+\infty$, nous supposons que l'application $\theta\mapsto\int_{\mathcal{X}} g(x)f_\theta(x)\mathrm{d}x$ est dérivable et que sa dérivée vaut :

$$\frac{\partial}{\mathrm{d}\theta}\int_{\mathcal{X}} g(x)f_\theta(x)\mathrm{d}x=\int_{\mathcal{X}} g(x)\frac{\partial}{\mathrm{d}\theta}f_\theta(x)\mathrm{d}x$$

où $\frac{\partial}{\mathrm{d}\theta}$ représente le gradient (donc si $p=1$, c'est juste la dérivée).

**Remarque**

Pour les deux premiers points, nous pouvons faire des remarques :

1. Un exemple classique d'ouverts de $\mathbb{R}^p$ est le pavé $]a_1,b_1[\times]a_2,b_2[\times\cdots\times]a_p,b_p[$ avec $(a_1,b_1,a_2,\ldots,b_p)\in\overline{\mathbb{R}}^{2p}$.
2. La loi uniforme $\mathcal{U}([0,\theta])$ ne vérifie pas ces conditions. Ceci ne veut pas dire que les résultats que nous démontrerons par la suite ne s'appliquent pas forcément ; par contre, les démonstrations seront plus compliquées.

**Définition 38 (Score)**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et un $n$-échantillon $\mathbf{X}=(X_1,\ldots,X_n)$, nous appelons **score**, noté $U(\cdot)$, la fonction $U:\theta\mapsto\mathbb{R}$ définie par

$$U(\theta)=\frac{\partial}{\mathrm{d}\theta}\log V_{\mathbf{X}}(\theta)=\frac{\partial}{\mathrm{d}\theta}\left(\sum_{i=1}^n\log f_\theta(X_i)\right).$$

**Propriétés 9**

Dans un modèle paramétrique régulier, nous avons pour tout $\theta\in\Theta$ :

$$\mathbb{E}_\theta\left[U(\theta)\right]=0.$$

**Preuve**

Pour montrer cela, il suffit de faire le calcul. Dans le cas continu, nous avons :

$$\begin{aligned}\mathbb{E}_\theta\left[U(\theta)\right] &= \mathbb{E}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\left(\sum_{i=1}^n\log f_\theta(X_i)\right)\right]\\ &= \sum_{i=1}^n\mathbb{E}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X_i)\right]\end{aligned}$$

$$
\begin{aligned}
&= n\int_{\mathcal{X}} \frac{\partial}{\mathrm{d}\theta} \log f_\theta(x) f_\theta(x) \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{\mathcal{X}} \frac{\frac{\partial}{\mathrm{d}\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) \mathrm{d}x \\
&= \sum_{i=1}^{n} \int_{\mathcal{X}} \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x \\
&= \sum_{i=1}^{n} \frac{\partial}{\mathrm{d}\theta} \underbrace{\int_{\mathcal{X}} f_\theta(x) \mathrm{d}x}_{=1} \\
&= 0.
\end{aligned}
$$

**Remarque**

L'estimateur du maximum de vraisemblance peut être vu comme la valeur annulant le score.

**Définition 39 (Information de Fisher)**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$, nous appelons **information de Fisher**, notée $I_n(\cdot)$, la matrice symétrique de taille $p \times p$ définie pour tout $\theta \in \Theta$ par :

$$
I_n(\theta) = \mathbb{E}_{\theta^\star}\left[U(\theta)U(\theta)^T\right] = \left(\mathbb{E}\left[\frac{\partial \log V_{\mathbf{X}}(\theta)}{\mathrm{d}\theta_i} \frac{\partial \log V_{\mathbf{X}}(\theta)}{\mathrm{d}\theta_j}\right]\right)_{1\le i,j\le p}.
$$

### Bornes de Cramer-Rao

Dans cette partie, nous faisons des hypothèses supplémentaires sur l'ensemble de définitions et sur la régularité des fonctions :

**Hypothèse**

Nous supposons que nous sommes dans un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ et que $\Theta \subset \mathbb{R}$.

**Corollaire 10 (Cas unidimensionnel)**

Dans le cas où $\theta \in \mathbb{R}$, nous avons :

$$
I_n(\theta) = nI_1(\theta) = n\int \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta(x)} \mathbb{1}_{\{f_\theta(x)>0\}} \mathrm{d}x.
$$

**Preuve**

Le fait que $I_n(\theta) = nI_1(\theta)$ vient de l'indépendance des observations et la deuxième partie est simplement le calcul.

**Remarque**

Ce corollaire permet de ne se concentrer que sur les propriétés de $I_1(\theta)$ que nous trouvons souvent simplement sous la notation $I(\theta)$.

**Corollaire 11 (Information de Fisher et variance)**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ unidimensionnel, nous avons :

$$I_1(\theta) = \mathbb{V}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right].$$

**Preuve**

Pour cela, nous faisons le calcul de la variance :

$$\begin{aligned}\mathbb{V}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right] &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right)^2\right] - \mathbb{E}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right]^2 \\ &= I_1(\theta) - \underbrace{\mathbb{E}_\theta\left[U(\theta)\right]^2}_{=0}.\end{aligned}$$

**Proposition 12 (Information de Fisher et dérivée seconde)**

Étant donné un modèle paramétrique régulier $(f_\theta)_{\theta\in\Theta}$ unidimensionnel, si nous pouvons dériver deux fois sous le signe somme alors nous avons :

$$I_1(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right].$$

**Preuve**

Il suffit de faire le calcul :

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right] &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(x)\right]f_\theta(x)\mathrm{d}x \\ &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\left[\frac{\partial}{\mathrm{d}\theta}\frac{\frac{\partial}{\mathrm{d}\theta}f_\theta(x)}{f_\theta(x)}\right]f_\theta(x)\mathrm{d}x \\ &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\left[\frac{\frac{\partial^2}{\mathrm{d}\theta^2}f_\theta(x)}{f_\theta(x)} - \frac{\left[\frac{\partial}{\mathrm{d}\theta}f_\theta(x)\right]^2}{f_\theta^2(x)}\right]f_\theta(x)\mathrm{d}x \\ &= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\frac{\partial^2}{\mathrm{d}\theta^2}f_\theta(x)\mathrm{d}x - \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}\frac{\left[\frac{\partial}{\mathrm{d}\theta}f_\theta(x)\right]^2}{f_\theta(x)}\mathrm{d}x \\ &= \frac{\partial^2}{\mathrm{d}\theta^2}\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}}f_\theta(x)\mathrm{d}x - I_1(\theta) \\ &= -I_1(\theta)\end{aligned}$$

**Exercices 3.5**

Montrer de deux façons différentes que pour un $n$-échantillon de loi de Poisson, nous avons $I_n(\theta) = n/\theta$.

**Théorème 13 (Borne de Cramer-Rao)**

Étant donné un estimateur $\widehat{g} = h(X)$ sans biais de $g(\theta)$ avec $g$ une fonction dérivable sur $\Theta$, si le modèle est régulier d'information de Fisher strictement positive sur $\Theta$ et si $\mathbb{E}_\theta\left[h^2(X)\right]$ est

bornée alors nous avons :

$$\forall \theta \in \Theta,\ \mathbb{V}_\theta\left[\widehat{g}\right] \geq \frac{\left(g'(\theta)\right)^2}{I_1(\theta)}.$$

**Preuve**

Comme $\widehat{g}$ est supposé sans biais, nous savons que $\mathbb{E}_\theta\left[\widehat{g}\right] = g(\theta)$ donc nous avons :

$$\begin{aligned}
g'(\theta) &= \frac{\partial}{\mathrm{d}\theta} g(\theta) \\
&= \frac{\partial}{\mathrm{d}\theta} \mathbb{E}_\theta\left[\widehat{g}\right] \\
&= \frac{\partial}{\mathrm{d}\theta} \mathbb{E}_\theta\left[h(X)\right] \\
&= \frac{\partial}{\mathrm{d}\theta} \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} h(x) f_\theta(x) \mathrm{d}x \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} h(x) \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} h(x) \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x - \underbrace{\mathbb{E}_\theta\left[h(X)\right]}_{\text{bornée}} \underbrace{\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x}_{=0} \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \left[h(x) - \mathbb{E}_\theta\left[h(X)\right]\right] \frac{\partial}{\mathrm{d}\theta} f_\theta(x) \mathrm{d}x \\
&= \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \left[h(x) - \mathbb{E}_\theta\left[h(X)\right]\right] \sqrt{f_\theta(x)} \frac{\frac{\partial}{\mathrm{d}\theta} f_\theta(x)}{f_\theta(x)} \sqrt{f_\theta(x)} \mathrm{d}x \\
&\leq \sqrt{\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \left[h(x) - \mathbb{E}_\theta\left[h(X)\right]\right]^2 f_\theta(x) \mathrm{d}x \int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta^2(x)} f_\theta(x) \mathrm{d}x} \\
&\quad \text{par Cauchy-Schwartz} \\
&\leq \sqrt{\mathbb{V}_\theta\left[h(X)\right]} \sqrt{\int_{\{x\in\mathcal{X}|f_\theta(x)>0\}} \frac{\left[\frac{\partial}{\mathrm{d}\theta} f_\theta(x)\right]^2}{f_\theta(x)} \mathrm{d}x} \\
&\leq \sqrt{\mathbb{V}_\theta\left[\widehat{g}\right]} \sqrt{I_1(\theta)}.
\end{aligned}$$

En passant au carré et en divisant par $I_1(\theta)$ (qui est strictment positif), nous avons le résultat.

**Définition 40 (Borne de Cramer-Rao)**

Dans le cas où la quantité d'intérêt est $g(\theta)$ avec $g$ dérivable, la **borne de Cramer-Rao** est la quantité :

$$\frac{\left(g'(\theta)\right)^2}{I_1(\theta)}.$$

**Corollaire 14 (Borne de Cramer-Rao et estimateur UMVU)**

Si un estimateur sans biais $\widehat{g}$ vérifie pour tout $\theta \in \Theta$ que sa variance est égale à la borne de Cramer-Rao alors c'est un estimateur UMVU.

**Corollaire 15 (Borne de Cramer-Rao pour un $n$-échantillon)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ d'information de Fisher $I_n$ et sous les conditions du théorème précédent, nous avons :

$$\forall \theta \in \Theta,\ \mathbb{V}_\theta\left[h(X_1, \ldots, X_n)\right] \geq \frac{\left(g'(\theta)\right)^2}{nI_1(\theta)}.$$

**Exemple**

Étant donné $\mathbf{X} = (X_1, \ldots, X_n)$ un $n$-échantillon de loi de Poisson $\mathcal{P}(\theta^\star)$ alors nous montrons que $\widehat{\theta}_{n,1} = \overline{\mathbf{X}}_n$ est un UMVU.

**Point méthode**

Pour un estimateur $UMVU$, il faut procéder en plusieurs étapes :

1. Vérifier s'il est non biaisé (sinon, cela s'arrête là).
2. Calculer la dérivée première ou la dérivée seconde en $\theta$ de la fonction densité $f_\theta$ (suivant si nous pouvons dériver qu'une seule ou deux fois sous le signe somme).
3. Calculer l'information de Fisher pour une observation.
4. Calculer la variance de l'estimateur.
5. Vérifier si borne de Cramer-Rao est atteinte ou non.

## 3.7 Solutions des exercices

**Exercices 3.1**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ suivant une loi de Poisson $\mathcal{P}(\theta^\star)$, nous avons vu qu'un estimateur possible de $\theta^\star$ est la variance empirique $\widehat{\theta}_{n,2} = \overline{X_n^2} - \overline{\mathbf{X}}_n^2 = \ell\left(\overline{X_n^2}, \overline{\mathbf{X}}_n\right)$ où $\ell(x, y) = x - y^2$ et nous souhaitons appliquer la méthode Delta.

Pour cela, nous prenons le vecteur $U_n = \begin{pmatrix} \overline{X_n^2} \\ \overline{\mathbf{X}}_n \end{pmatrix}$ dont nous devons calculer la loi asymptotique. Or, si nous prenons $Y_i = \begin{pmatrix} X_i^2 \\ X_i \end{pmatrix}$, nous voyons que le théorème de la limite centrale pourra nous aider à trouver la loi puisque nous avons des observations $(Y_1, \ldots, Y_n)$ indépendantes et de même loi. Commençons par calculer le moment d'ordre 1 de $Y_i$, pour cela, nous rappelons les moments suivants :

$$\mathbb{E}\left[X_i\right] = \theta^\star,\ \mathbb{E}\left[X_i^2\right] = \theta^{\star 2} + \theta^\star,$$
$$\mathbb{E}\left[X_i^3\right] = \theta^{\star 3} + 3\theta^{\star 2} + \theta^\star \text{ et } \mathbb{E}\left[X_i^4\right] = \theta^{\star 4} + 6\theta^{\star 3} + 7\theta^{\star 2} + \theta^\star.$$

Ceci nous permet de calculer les moments d'ordre 1 :

$$\begin{aligned} \mathbb{E}\left[Y_i\right] &= \begin{pmatrix} \mathbb{E}\left[X_1^2\right] \\ \mathbb{E}\left[X_1\right] \end{pmatrix} \\ &= \begin{pmatrix} \theta^{\star 2} + \theta^\star \\ \theta^\star \end{pmatrix}. \end{aligned}$$

Pour la matrice de variance-covariance, nous avons :

$$\begin{aligned} \mathbb{V}\left[X_i^2\right] &= \mathbb{E}\left[X_i^4\right] - \mathbb{E}\left[X_i^2\right]^2 \\ &= \theta^{\star 4} + 6\theta^{\star 3} + 7\theta^{\star 2} + \theta^\star - \left(\theta^{\star 2} + \theta^\star\right)^2 \end{aligned}$$

$$\begin{aligned}
&= \theta^{\star 4} + 6\theta^{\star 3} + 7\theta^{\star 2} + \theta^\star - \theta^{\star 4} - 2\theta^{\star 3} - \theta^{\star 2} \\
&= 4\theta^{\star 3} + 6\theta^{\star 2} + \theta^\star, \\
\mathbb{V}[X_i] &= \theta^\star, \\
\mathrm{Cov}\left(X_i^2, X_i\right) &= \mathbb{E}\left[X_i^3\right] - \mathbb{E}\left[X^2\right]\mathbb{E}\left[X_i\right] \\
&= \theta^{\star 3} + 3\theta^{\star 2} + \theta^\star - \theta^{\star 3} - \theta^{\star 2} \\
&= 2\theta^{\star 2} + \theta^\star.
\end{aligned}$$

Au final, par le théorème de la limite centrale, nous avons :

$$\sqrt{n}\left(U_n - \begin{pmatrix}\theta^{\star 2} + \theta^\star \\ \theta^\star\end{pmatrix}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4\theta^{\star 3} + 6\theta^{\star 2} + \theta^\star & 2\theta^{\star 2} + \theta^\star \\ 2\theta^{\star 2} + \theta^\star & \theta^\star\end{pmatrix}\right).$$

Maintenant, calculons la dérivée :

$$D\ell\left(\theta^{\star 2} + \theta^\star \quad \theta^\star\right) = \begin{pmatrix}1 \\ -2\theta^\star\end{pmatrix}.$$

Il reste donc à calculer la variance :

$$\begin{aligned}
&\begin{pmatrix}1 & -2\theta^\star\end{pmatrix}\begin{pmatrix}4\theta^{\star 3} + 6\theta^{\star 2} + \theta^\star & 2\theta^{\star 2} + \theta^\star \\ 2\theta^{\star 2} + \theta^\star & \theta^\star\end{pmatrix}\begin{pmatrix}1 \\ -2\theta^\star\end{pmatrix} \\
= &\begin{pmatrix}4\theta^{\star 3} + 6\theta^{\star 2} + \theta^\star - 4\theta^{\star 3} - 2\theta^{\star 2} & 2\theta^{\star 2} + \theta^\star - 2\theta^{\star 2}\end{pmatrix}\begin{pmatrix}1 \\ -2\theta^\star\end{pmatrix} \\
= &\begin{pmatrix}4\theta^{\star 2} + \theta^\star & \theta^\star\end{pmatrix}\begin{pmatrix}1 \\ -2\theta^\star\end{pmatrix} \\
= &\ 4\theta^{\star 2} + \theta^\star - 2\theta^{\star 2} \\
= &\ 2\theta^{\star 2} + \theta^\star
\end{aligned}$$

Donc, au final, nous avons :

$$\sqrt{n}\left(\widehat{\theta}_{n,2} - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, 2\theta^{\star 2} + \theta^\star\right).$$

**Exercices 3.2**

Nous pouvons au choix utiliser l'un ou l'autre des corollaires. Dans les deux cas, nous commençons par calculer le logarithme de la densité :

$$\begin{aligned}
\log f_\theta(x) &= \log\left(e^{-\theta}\frac{\theta^x}{x!}\right) \\
&= -\theta + x\log\theta - \log x!.
\end{aligned}$$

$$\begin{aligned}
I_1(\theta) &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\mathrm{d}\theta}\log f_\theta(X)\right)^2\right] \\
&= \mathbb{E}_\theta\left[\left(-1 + \frac{X}{\theta}\right)^2\right] \\
&= \mathbb{V}_\theta\left[\frac{X}{\theta}\right] \text{ car } \frac{X}{\theta} \sim \mathcal{E}(1), \\
&= \frac{1}{\theta^2}\theta
\end{aligned}$$

$$= \frac{1}{\theta}.$$

Autre solution, nous calculons la dérivée seconde :

$$\begin{aligned} I_1(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right] \\ &= -\mathbb{E}_\theta\left[\frac{\partial}{\mathrm{d}\theta}\left(-1+\frac{X}{\theta}\right)\right] \\ &= -\mathbb{E}_\theta\left[-\frac{X}{\theta^2}\right] \\ &= \frac{\mathbb{E}_\theta[X]}{\theta^2} \\ &= \frac{\theta}{\theta^2} \\ &= \frac{1}{\theta}. \end{aligned}$$

Le résultat final vient du fait que $I_n(\theta) = nI_1(\theta)$.

# Chapitre 4

# Intervalle et région de confiance

*"Capitaine, je suis sûr à 80% que c'est le meurtrier."*
Adrien Monk, personnage éponyme de la série *Monk*.

## 4.1 Objectifs

Le but de ce chapitre est de comprendre qu'un risque est toujours possible et qu'il faut donc en tenir compte.

## 4.2 Introduction

Ce chapitre commence par un certain nombre de définitions.

**Définition 41 (Région de confiance)**
Étant donné $\alpha \in ]0,1[$, une **région de confiance de** $g(\theta^\star)$ **au niveau** $1-\alpha$ est un ensemble $\widehat{\mathcal{C}}$ construit mesurablement par rapport à l'observation $X$ et tel que pour tout $\theta^\star \in \Theta$, nous ayons :

$$\mathbb{P}_{\theta^\star}\left(g(\theta^\star) \in \widehat{\mathcal{C}}\right) \geq 1-\alpha.$$

Lorsque l'inégalité devient une égalité, nous disons que le **niveau de confiance est exactement égal à** $1-\alpha$.

Nous avons immédiatement la généralisation pour un $n$-échantillon :

**Définition 42 (Région de confiance asymptotique)**
Étant donné $\alpha \in ]0,1[$, une **région de confiance asymptotique de** $g(\theta^\star)$ **au niveau** $1-\alpha$ est une suite d'ensembles $\left(\widehat{\mathcal{C}}_n\right)_{n\in\mathbb{N}}$, chacun construit mesurablement par rapport au $n$-échantillon $(X_1,\ldots,X_n)$ et telle que pour tout $\theta^\star \in \Theta$, nous ayons :

$$\liminf_{n\to+\infty} \mathbb{P}_{\theta^\star}\left(g(\theta^\star) \in \widehat{\mathcal{C}}_n\right) \geq 1-\alpha.$$

**Remarque**
Usuellement, nous prenons $\alpha = 5\%$ ou $\alpha = 1\%$.

**Remarque**
Étant donné un niveau de confiance fixé, une région de confiance est d'autant meilleure que son aire est petite.

**Définition 43 (Intervalle de confiance)**
Si $g(\theta^\star) \in \mathbb{R}$, nous parlons plutôt d'**intervalle de confiance**.

**Attention au piège**

Une région de confiance est une variable aléatoire et nous calculons donc la probabilité que cette variable aléatoire englobe le paramètre d'intérêt qui lui est fixé. En particulier, il ne faut pas confondre la région de confiance qui est aléatoire et son estimée (sa réalisation).
Si je prends une région de confiance $\widehat{\mathcal{C}}_n = [\widehat{A}_n, \widehat{B}_n]$ telle que :

$$\mathbb{P}_{\theta^\star}\left(\theta^\star \in [\widehat{A}_n, \widehat{B}_n]\right) = 95\%$$

et que je calcule une réalisation de celle-ci, par exemple $[0.33, 0.75]$ alors on ne peut plus dire que $\mathbb{P}_{\theta^\star}\left(\theta^\star \in [0.33, 0.75]\right) = 95\%$ puisque tout est fixé :

- Soit $\theta^\star$ est dans l'intervalle et la probabilité vaut 1.
- Soit il ne l'est pas et la probabilité est nulle.

La notion de région de confiance est de dire que si nous répétons l'expérience un grand nombre de fois alors le paramètre d'intérêt se trouvera dans la région de confiance estimée en moyenne 95% du temps.

## 4.3 Premières constructions

La première façon de construire un intervalle de confiance est d'utiliser les quantiles :

**Définition 44 (Quantile d'ordre $p$)**
Pour tout $\beta \in ]0;1[$, nous appelons **quantile d'ordre** $\beta$ d'une variable aléatoire $X$ de fonction de répartition $F_X$ est définie par

$$q_\beta = \inf\left\{q \in \mathbb{R} \,|\, F_X(q) \geq \beta\right\}.$$

Il est parfois noté $F^{-1}(\beta)$.

En particulier, nous avons :

**Définitions 45 (Quantiles particuliers)**

- La **médiane** est le quantile d'ordre 50%.
- Les **quartiles** sont les quantiles d'ordre $25k\%$ avec $k \in \{1, 2, 3\}$.
- Les **déciles** sont les quantiles d'ordre $10k\%$ avec $k \in \{1, \ldots, 9\}$.
- Les **centiles** sont les quantiles d'ordre $k\%$ avec $k \in \{1, \ldots, 99\}$.

**Remarque**

Si la fonction est inversible (voir figure 4.1 (a)), le quantile est directement l'inversion de cette dernière. Si la fonction possède un plateau horizontal (voir figure 4.1 (b)), il faut prendre au début du plateau (la fonction quantile aura donc un saut). Si la fonction possède un saut (voir figure 4.1 (c)), il faut prendre la valeur correspondante à ce saut (la fonction quantile aura donc un plateau horizontal).

**Exemple**

Pour la loi $\mathcal{N}(0, 1)$, deux quantiles classiques (à connaître) sont : $q_{0,975} \approx 1,96$ et $q_{0,95} \approx 1,645$.

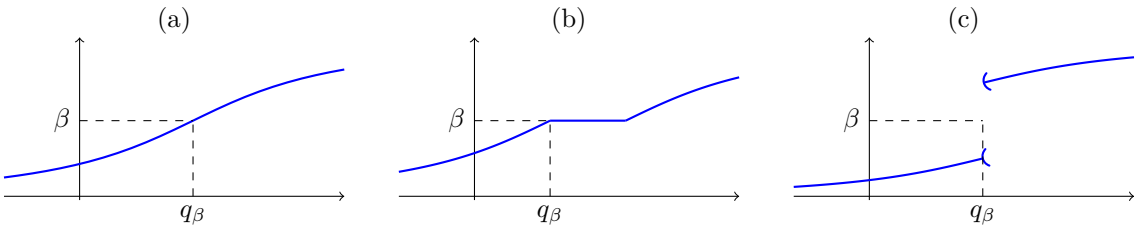



FIGURE 4.1 – Calcul de différents quantiles en fonction du type de fonction de répartition : inversible (a), avec un plateau (b) ou avec un saut (c).

**Exercices 4.1**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de même loi $\mathcal{N}(\theta^\star, 1)$. Proposer un intervalle de confiance de niveau exactement 95% de la forme $]-\infty, a]$, $[a, +\infty[$ et un **intervalle bilatère** c'est-à-dire de la forme $[a, b]$. Lequel préférez-vous ?

**Point méthode (Méthode du pivot)**

Pour calculer un intervalle de confiance, nous procédons souvent à l'aide de la **méthode dite du pivot** :

1. Choix d'un estimateur.
2. Calcul de sa loi en fonction de $\theta^\star$.
3. Transformation de l'estimateur pour obtenir une statistique dont la loi ne dépend plus de $\theta^\star$.
4. Détermination des quantiles de la loi pour estimer la région de confiance adéquat.

## 4.4 Intervalle de confiance de niveaux obtenus par des inégalités de probabilités

Dans cette partie, nous présentons des intervalles de confiance obtenus à partir d'inégalités de probabilités classiques.

### 4.4.1 Cas de variances uniformément bornée

**Proposition 16 (Inégalité de Bienaymé-Tchebychev)**

Soit une variable aléatoire $Y$ admettant un moment d'ordre 2 alors nous avons :

$$\forall \varepsilon > 0, \mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \varepsilon\right) \leq \frac{\mathbb{V}[Y]}{\varepsilon^2}.$$

Ce théorème est une conséquence de l'inégalité de Markov :

**Lemme 17 (Inégalité de Markov)**

Soit $Z$ une variable aléatoire réelle presque sûrement positive ou nulle alors nous avons :

$$\forall a > 0, \mathbb{P}\left(Z \geq a\right) \leq \frac{\mathbb{E}[Z]}{a}.$$

**Preuve**

Pour tout $a > 0$, nous avons presque sûrement $Z \geq a\mathbb{1}_{\{Z \geq a\}}$ et nous avons donc :

$$\mathbb{E}[Z] \geq \mathbb{E}\left[a\mathbb{1}_{\{Z \geq a\}}\right] = a\mathbb{P}(Z \geq a) \quad \Leftrightarrow \quad \mathbb{P}(Z \geq a) \leq \frac{\mathbb{E}[Z]}{a}.$$

La preuve de la proposition 16 est alors :

**Preuve**

Comme la fonction $x \mapsto x^2$ est bijective sur $\mathbb{R}^+$ à valeurs dans $\mathbb{R}^+$, nous avons par le lemme 17 pour tout $\varepsilon > 0$ :

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \quad = \quad \mathbb{P}\left((Y - \mathbb{E}[Y])^2 \geq \varepsilon^2\right)$$

et nous pouvons appliquer l'inégalité de Markov avec $Z = (Y - \mathbb{E}[Y])^2$ et $a = \varepsilon^2$.

**Exemple**

Étant donné un $n$-échantillon de loi $\mathcal{B}(\theta^\star)$, l'estimateur du maximum de vraisemblance est la moyenne $\overline{\mathbf{X}}_n$ et nous avons pour tout $\varepsilon > 0$ :

$$\mathbb{P}_{\theta^\star}\left(\left|\overline{\mathbf{X}}_n - \theta^\star\right| \geq \varepsilon\right) \leq \frac{\theta^\star(1-\theta^\star)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2} = \alpha.$$

Donc, un intervalle de confiance bilatéral de niveau $1-\alpha$ est :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{\mathbf{X}}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{\mathbf{X}}_n + \frac{1}{2\sqrt{n\alpha}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la longueur est d'environ 0.45.

### 4.4.2 Cas de lois à supports inclus dans un compact donné

**Lemme 18 (Inégalité de Hoeffding)**

Étant donnés, $(Y_1, \ldots, Y_n)$ des variables aléatoires réelles indépendantes et $(a_1, \ldots, a_n)$ et $(b_1, \ldots, b_n)$ des $n$-uplets de réels tels que pour tout $i \in \{1, \ldots, n\}$ :

- $\mathbb{E}[Y_i] = 0$,
- $a_i \leq Y_i \leq b_i$ presque sûrement.

Alors, pour tout $t > 0$, nous avons :

$$\mathbb{P}\left(\sum_{i=1}^n Y_i > t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

**Corollaire 19**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi commune à support inclus dans $[a, b]$ admettant un moment d'ordre 1, nous avons :

$$\forall \varepsilon > 0,\ \mathbb{P}_{\theta^\star}\left(\left|\overline{\mathbf{X}}_n - \mathbb{E}_{\theta^\star}[X_1]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}(\theta^\star)$ alors nous avons presque sûrement $a = 0 \leq X_i \leq b = 1$ et donc pour tout $\varepsilon > 0$ :

$$\mathbb{P}_{\theta^\star}\left(\left|\overline{\mathbf{X}}_n - \theta^\star\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

Le choix de $\varepsilon = \sqrt{1/(2n)\log(2/\alpha)}$ conduit à :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{\mathbf{X}}_n - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, \overline{\mathbf{X}}_n + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la longueur est d'environ 0.27.

**Remarque**

De manière générale, comme la variance d'une variable aléatoire encadrée par $a$ et $b$ est plus petite que $(b-a)^2/4$, comparer la précision des deux méthodes revient à comparer $1/\sqrt{2\alpha}$ et $\sqrt{\log(2/\alpha)}$. Pour les petites valeurs de $\alpha$, c'est l'intervalle fourni par l'inégalité de Hoeffding qui propose un meilleure résultat au prix d'une hypothèse plus forte.

## 4.5 Intervalle de confiance asymptotique

Le théorème de la limite centrale permet d'avoir à chaque fois un intervalle de confiance de niveau asymptotique $1-\alpha$ pour l'espérance d'une loi puisque nous avons :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mathbb{E}_{\theta^\star}[X_1]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}_{\theta^\star}[X_1]\right).$$

Nous pouvons alors proposer l'intervalle :

$$\left[\overline{\mathbf{X}}_n \pm q_{1-\alpha/2}\sqrt{\frac{\mathbb{V}_{\theta^\star}[X_1]}{n}}\right]$$

comme intervalle de confiance (asymptotique) à condition de connaître la variance. Or, cette variance dépend quasiment tout le temps de $\theta^\star$.

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}(\theta^\star)$ alors nous avons :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \theta^\star\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \theta^\star(1-\theta^\star)\right)$$

et il faudrait isoler $\theta^\star$ mais c'est compliqué.

### 4.5.1 Cas de variances uniformément bornées

Dans l'exemple précédent, si nous *admettons* connaître la variance $\theta^\star(1-\theta^\star)$ alors nous pouvons proposer :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{\mathbf{X}}_n \pm q_{1-\alpha/2}\sqrt{\frac{\theta^\star(1-\theta^\star)}{n}}\right] \subset \left[\overline{\mathbf{X}}_n \pm \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, nous trouvons une longueur de 0.196.

### 4.5.2 Estimation consistante de la variance

S'il existe une suite d'estimateurs $\left(\widehat{\sigma}_n^2\right)_{n\in\mathbb{N}^\star}$ de la variance consistante alors d'après le lemme de Slutsky, nous avons pour tout $\theta^\star \in \Theta$ :

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\left(\overline{\mathbf{X}}_n - \mathbb{E}_{\theta^\star}[X_1]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

D'où l'intervalle de confiance de niveau asymptotiquement exact $1-\alpha$ :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{\mathbf{X}}_n \pm q_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right].$$

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}(\theta^\star)$ alors nous avons :

$$IC_{1-\alpha}(\theta^\star) = \left[\overline{\mathbf{X}}_n \pm q_{1-\alpha/2}\sqrt{\frac{\overline{\mathbf{X}}_n\left(1-\overline{\mathbf{X}}_n\right)}{n}}\right].$$

### 4.5.3 Stabilisation de la variance

À l'aide de la méthode Delta, nous avons que pour toute fonction $\varphi$ continument dérivable sur $\Theta \subset \mathbb{R}$, nous avons :

$$\sqrt{n}\left(\varphi(\overline{\mathbf{X}}_n) - \varphi\left(\mathbb{E}_{\theta^\star}[X_1]\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \varphi'^2\mathbb{V}_{\theta^\star}[X_1]\right).$$

Donc, si prenons $\varphi$ telle que $\varphi'(\theta^\star) = 1/\sqrt{\mathbb{V}_{\theta^\star}[X_1]}$ alors nous aura la convergence en loi suivante :

$$\sqrt{n}\left(\varphi(\overline{\mathbf{X}}_n) - \varphi\left(\mathbb{E}_{\theta^\star}[X_1]\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

et nous pouvons en tirer l'intervalle de confiance suivant :

$$IC_{1-\alpha}(\theta^\star) = \varphi^{-1}\left(\left[\varphi(\overline{\mathbf{X}}_n) \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\right]\right)$$

c'est-à-dire l'image inverse de l'intervalle.

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{B}(\theta^\star)$, nous devons résoudre $\varphi'(\theta^\star) = 1/\sqrt{\theta^\star(1-\theta^\star)}$. Or, nous savons que la dérivée de la fonction arcsin est :

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}}$$

alors en prenant $\varphi(x) = 2\arcsin\sqrt{x}$, nous avons :

$$IC_{1-\alpha}(\theta^\star) = \left[\sin^2\left(\arcsin\sqrt{\overline{\mathbf{X}}_n} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\right)\right].$$

# Deuxième partie

# UE Test

# Chapitre 5

# Introduction aux tests

*"Dans la justice, il y a deux visions :*

- *En France, nous sommes présumé innocent jusqu'à ce que notre culpabilité soit établie ;*
- *Aux États-Unis, nous sommes présumé coupable jusqu'à ce que notre innocence soit établie."*

## 5.1 Objectifs

La notion de test est assez complexe car elle repose sur une asymétrie des deux hypothèses mises en concurrence.

## 5.2 Formalisme et démarche expérimentale

Dans le cadre du modèle paramétrique $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$, nous nous donnons deux sous-ensembles $\Theta_0$ et $\Theta_1$ **disjoints** et inclus dans $\Theta$ (nous n'imposons pas qu'ils forment une partition). À la vue d'une observation, nous souhaitons décider si $\theta^\star \in \Theta_0$ ou pas ; si ce n'est pas le cas, nous considérerons que $\theta^\star \in \Theta_1$.

**Définitions 46 (Hypothèses)**

Étant donnés deux sous-ensembles $\Theta_0$ et $\Theta_1$ **disjoints** et inclus dans $\Theta$, nous définissons respectivement :

$\mathcal{H}_0$ : l'**hypothèse nulle** considère que $\theta^\star \in \Theta_0$.

$\mathcal{H}_1$ : l'**hypothèse alternative** considère que $\theta^\star \in \Theta_1$.

Pour tout $j \in \{0, 1\}$, nous dirons que $\mathcal{H}_j$ est **simple** si $\Theta_j$ est réduit à un singleton et **composite** sinon.

**Exemple**

Dans le cas d'une pièce, nous nous intéressons souvent à savoir si la probabilité d'avoir *pile* est égale à celle d'avoir *face* ($\Theta_0 = \{1/2\}$) ou pas ($\Theta_1 = [0, 1]\backslash\{1/2\}$).

À l'aide de ces notations, nous pouvons introduire la définition d'un test :

**Définitions 47 (Test)**

Nous appelons **test de l'hypothèse $\mathcal{H}_0$ contre l'hypothèse $\mathcal{H}_1$** toute fonction $\phi(X)$ à valeur dans $\{0, 1\}$ avec $\phi$ mesurable et pouvant dépendre de $\Theta_0$ et $\Theta_1$ :

- Lorsque $\phi(X) = 0$, nous disons que nous **conservons** l'hypothèse $\mathcal{H}_0$.
- Lorsque $\phi(X) = 1$, nous disons que nous **rejetons** l'hypothèse $\mathcal{H}_0$ et nous **acceptons** l'hypothèse $\mathcal{H}_1$.

**Remarque**

Un test peut toujours s'écrire $\phi(X) = \mathbb{1}_{\{X \in \mathcal{R}\}}$ où $\mathcal{R} \in \mathcal{E}$ ou encore $\phi(X) = \mathbb{1}_{\{h(X) \in \mathcal{R}'\}}$ avec $h$ mesurable et $\mathcal{R}' \in \mathcal{B}(\mathbb{R})$ la tribu des boréliens de $\mathbb{R}$.

**Définitions 48 (Statistique de test)**

Dans le cas où $\phi(X) = \mathbb{1}_{\{h(X) \in \mathcal{R}'\}}$ avec $h$ mesurable et $\mathcal{R}' \in \mathcal{B}(\mathbb{R})$ la tribu des boréliens de $\mathbb{R}$, nous appelons $\mathcal{R}'$ la **région de rejet** et $h(X)$ la **statistique de test**.
Le complémentaire de la région de rejet est parfois appelée **région d'acceptation**.

### 5.2.1 Mesure de la qualité d'un test

Pour mesurer la qualité d'un test, nous nous intéressons d'une part à la probabilité d'accepter à tord et d'autre part de rejeter à tord.

**Définitions 49 (Risques de première et seconde espèce)**

Les **risques de première** et **de seconde espèce** du test $\phi(X)$ sont définis comme les fonctions $\underline{\alpha}$ sur $\Theta_0$ et $\underline{\beta}$ sur $\Theta_1$ par :

$$\underline{\alpha} : \begin{array}{ccc} \Theta_0 & \to & [0,1] \\ \theta^\star & \mapsto & \mathbb{P}_{\theta^\star}(\phi(X) = 1) \end{array} \quad \text{et} \quad \underline{\beta} : \begin{array}{ccc} \Theta_1 & \to & [0,1] \\ \theta^\star & \mapsto & \mathbb{P}_{\theta^\star}(\phi(X) = 0) \end{array}$$

et nous pouvons introduire la **puissance d'un test** comme la fonction $1 - \underline{\beta}$ et sa **taille** comme le réel $\alpha^\star$ défini par

$$\alpha^\star = \sup_{\theta^\star \in \Theta_0} \underline{\alpha}(\theta^\star) = \sup_{\theta^\star \in \Theta_0} \mathbb{P}_{\theta^\star}(\phi(X) = 1).$$

Enfin, nous disons que le test est de **niveau** $\alpha$ si sa taille $\alpha^\star$ est inférieure ou égale à $\alpha$.

Comme pour les intervalles de confiances, la notion de test s'étend également au cas asymptotique

**Définition 50 (Version asymptotique)**

Étant donnée une suite $(\phi_n)_{n \in \mathbb{N}}$ de tests et en notant $\alpha_n^\star$ la taille de $\phi_n$, nous disons que la suite de tests est de **niveau asymptotique** $\alpha$ si

$$\limsup_{n \to +\infty} \alpha_n^\star \leq \alpha.$$

**Point méthode (Représentation sous forme de tableau)**

Nous pouvons résumer les informations précédentes de la façon suivante :

| $\phi(X)$ \ $\theta^\star$ | $\Theta_0$ | $\Theta_1$ |
|---|---|---|
| 0 | Conservation correcte | Risque seconde espèce : $\underline{\beta}$ |
| 1 | Risque première espèce : $\underline{\alpha}$ | Puissance |

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ iid de loi $\mathcal{N}(\theta^\star, 1)$ avec $\theta \in \mathbb{R}$, nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star \geq 1, \\ \mathcal{H}_1 : & \theta^\star < 1, \end{cases} \quad \text{soit} \quad \begin{cases} \Theta_0 = [1, +\infty[, \\ \Theta_1 = ]-\infty, 1[. \end{cases}$$

L'intuition nous conduit à choisir $\phi(X) = \mathbb{1}_{\{\overline{\mathbf{X}}_n<1\}}$ : le test est fondé sur la statistique $\overline{\mathbf{X}}_n$ et la région de rejet $]-\infty, 1[$.

On voit que pour tout $\theta^\star \geq 1$, $\underline{\alpha}(\theta^\star) = \mathbb{P}_{\theta^\star}\left(\overline{\mathbf{X}}_n < 1\right) = F_{\mathcal{N}(0,1)}\left(\sqrt{n}\,(1-\theta^\star)\right)$ car $\overline{\mathbf{X}}_n \sim \mathcal{N}\left(\theta^\star, 1/n\right)$ et pour tout $\theta^\star < 1$, $\underline{\beta}(\theta^\star) = 1 - F_{\mathcal{N}(0,1)}\left(\sqrt{n}\,(1-\theta^\star)\right)$.

En revanche, la taille du test vaut $\alpha^\star = 1/2$ ce qui veut dire qu'on ne fait pas mieux qu'un lancer de pile ou face.

### 5.2.2 Dissymétrie des rôles des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$

Si nous voulons minimiser l'erreur de première espère, il suffirait de toujours conserver l'hypothèse $\mathcal{H}_0$ ; à l'opposé, si nous souhaitons minimiser l'erreur de seconde espèce, il suffit de rejeter à chaque fois l'hypothèse $\mathcal{H}_0$ au profit de l'hypothèse $\mathcal{H}_1$. Nous voyons donc que minimiser les deux erreurs en même temps est compliquée (voir impossible sauf dans certains cas). Il a donc été choisi de dissymétriser le problème en faisant jouer un rôle particulier à $\mathcal{H}_0$ et en contrôlant exclusivement le risque de première espèce. On prendra par exemple pour $\mathcal{H}_0$ :

- Une hypothèse communément admise : par exemple, il n'y a pas de traces de vie autre que sur Terre dans notre système solaire.
- Une hypothèse de prudence ou conservatrice : le nouveau médicament qui est en train d'être testé ne soigne pas mieux que l'ancien.
- Ou plus simplement la seule hypothèse que l'on peut facilement formuler.

Dans cette vision le statisticien impose le niveau $\alpha$ (par exemple 5%) et acceptera de rejeter à tord son hypothèse dans 5% des cas.

### 5.2.3 Démarche de construction et mise en œuvre pratique

**Exemple**

Étant donné un $n$-échantillon $X_1, \ldots, X_n$ de loi normale $\mathcal{N}(\theta^\star, 1)$ et regardons le test suivant :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star \geq 1, \\ \mathcal{H}_1 : & \theta^\star < 1. \end{cases}$$

La statistique naturelle est toujours $\overline{\mathbf{X}}_n$ et nous choisissons, à la vue de $\mathcal{H}_1$, une région de rejet de la forme $]-\infty; k_\alpha[$ c'est-à-dire une statistique de la forme $\phi(X) = \mathbb{1}_{\left\{\overline{\mathbf{X}}_n<k_\alpha\right\}}$. La question revient donc au choix de $k_\alpha$.

Pour ce faire, nous calculons la taille du test :

$$\begin{aligned} \alpha^\star &= \sup_{\theta^\star\in\Theta_0} \mathbb{P}_{\theta^\star}\left(\overline{\mathbf{X}}_n < k_\alpha\right) \\ &= \sup_{\theta^\star\geq 1} \mathbb{P}_{\theta^\star}\left(\sqrt{n}\left(\overline{\mathbf{X}}_n - \theta^\star\right) < \sqrt{n}\left(k_\alpha - \theta^\star\right)\right) \\ &= \sup_{\theta^\star\geq 1} F\left[\sqrt{n}\left(k_\alpha - \theta^\star\right)\right] \text{ où } F \text{ est la fonction de répartition d'une gaussienne centrée réduite,} \\ &= F\left[\sqrt{n}\left(k_\alpha - 1\right)\right] \text{ car } F \text{ est croissante.} \end{aligned}$$

Donc, pour que le niveau du test soit exactement $\alpha$, il suffit que :

$$\begin{aligned} \alpha = \alpha^\star &\Leftrightarrow \alpha = F\left[\sqrt{n}\left(k_\alpha - 1\right)\right] \\ &\Leftrightarrow q_\alpha = \sqrt{n}\left(k_\alpha - 1\right) \text{ où } q \text{ est le quantile d'une gaussienne centrée réduite,} \\ &\Leftrightarrow \frac{q_\alpha}{\sqrt{n}} + 1 = k_\alpha. \end{aligned}$$

Donc, le test est de la forme $\phi(X) = \mathbb{1}_{\left\{\overline{\mathbf{X}}_n < \frac{q_\alpha}{\sqrt{n}}+1\right\}}$ et a un niveau exactement de $\alpha$.

### 5.2.4 Qualités d'un test

Nous recensons un certain nombre de propriétés éventuelles.

**Définitions 51**

Un test est dit **sans biais** lorsque sa fonction puissance vérifie $1-\underline{\beta} > \alpha$ sur $\Theta_1$.

Une suite de tests est dit **consistante** s'ils sont tous de niveaux $\alpha$ et si, pour tout $\theta^\star \in \Theta_1$, nous avons :

$$\underline{\beta}_n(\theta^\star) \underset{n\to+\infty}{\longrightarrow} 0.$$

La **robustesse** exprime l'absence de sensibilité d'un test, éventuellement asymptotique, au modèle (où à la forme de l'observation). Entre d'autres termes, si la modélisation du phénomène est mauvaise, est-ce que nous obtenons les mêmes conclusions ?

Comme pour les estimateurs, nous pouvons chercher des tests ayant des propriétés plus intéressantes que les autres.

**Définition 52 (Test uniformément plus puissant)**

Pour deux tests $\phi(X)$ et $\phi'(X)$ de niveau $\alpha$, nous disons que $\phi(X)$ est **uniformément plus puissant (UPP)** que $\phi'(X)$ si leurs fonctions puissances $1-\underline{\beta}$ et $1-\underline{\beta}'$ vérifient $1-\underline{\beta} \geq 1-\underline{\beta}'$ ; c'est-à-dire que pour tout $\theta^\star \in \Theta_1$, nous avons :

$$\mathbb{P}_{\theta^\star}\left(\phi(X)=1\right) \geq \mathbb{P}_{\theta^\star}\left(\phi'(X)=1\right).$$

Nous disons que $\phi(X)$ est $UPP(\alpha)$

## 5.3 Un outil important : $p$-valeur

Le principe de la $p$-valeur (ou *p-value* en anglais) est expliquée sur la figure 5.1 : plutôt que de demander la valeur $\alpha$ visée par l'utilisateur, la plupart des logiciels renvoie la valeur estimée à partir de la moyenne ; il ne reste plus qu'à l'utilisateur de comparer cette valeur à la cible.

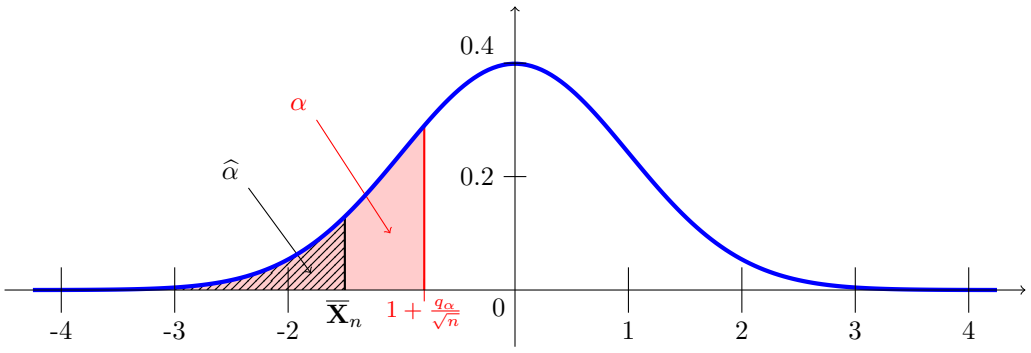



FIGURE 5.1 – Représentation schématique de la $p$-valeur (zone hachurée) par rapport à un niveau $\alpha$ visé (zone rouge).

**Définition 53 ($p$-valeur)**

Supposons avoir construit une famille de test $\phi_\alpha(X)$, chacun de niveau $\alpha$, pour $\alpha \in ]0,1[$. La **$p$-valeur** associée à cette famille et à l'observation $X$ est le réel défini par :

$$\widehat{\alpha}(X) = \sup\left\{\alpha \in ]0,1[ \,|\, \phi_\alpha(X)=0\right\}.$$

**Remarque**

En d'autres termes, $\widehat{\alpha}(X)$ est le plus grand niveau autorisant l'acceptation de $\mathcal{H}_0$ et forme donc, en un certain sens, un indice de crédibilité de $\mathcal{H}_0$ à la vue de $X$. La $p$-valeur mesure l'adéquation entre $\mathcal{H}_0$ et les observations : plus $\widehat{\alpha}(X)$ est faible et plus nos informations contredisent $\mathcal{H}_0$.

Le cas typique est celui où la famille $(\phi_\alpha(X))_{\alpha\in]0,1[}$ est p.s. croissante en $\alpha$, c'est-à-dire que $\phi_\alpha(X)$ décroît vers 0 lorsque $\alpha$ décroît vers 0 (ce qui est une condition naturelle). Ceci équivaut à dire que la famille des régions de rejet est croissante au sens de l'inclusion. Dans ce cas, on a également :

$$\widehat{\alpha}(X) = \inf\{\alpha \in ]0,1[ | \phi_\alpha(X) = 0\}.$$

En outre, le test $\phi_\alpha$ de niveau $\alpha$ conserve $\mathcal{H}_0$ à la vue de $X$ pour $\alpha < \widehat{\alpha}(X)$ et rejette $\alpha > \widehat{\alpha}(X)$.

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi $\mathcal{N}(\theta^\star, 1)$ et si nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star \geq 1, \\ \mathcal{H}_1 : & \theta^\star < 1. \end{cases}$$

Nous avons vu que la région de rejet est $\overline{\mathbf{X}}_n < k_\alpha = 1 + \frac{q_\alpha}{\sqrt{n}}$.

La $p$-valeur est caractérisée par le cas limite $k_{\widehat{\alpha}_n} = \overline{\mathbf{X}}_n$ soit $\widehat{\alpha}_n(\mathbf{X}) = F\left(\sqrt{n}\left(\overline{\mathbf{X}}_n - 1\right)\right)$.

Avec $\overline{X}_4 = 0.5$, nous obtenons $\widehat{\alpha}(\mathbf{X}) \approx 15,9\%$ et nous conservons donc l'hypothèse $\mathcal{H}_0$ avec un niveau de 10% ou même de 5%.

## 5.4 Test du rapport de vraisemblance

Il s'agit d'une méthode générale pour construire des tests $UPP(\alpha)$ sous certaines hypothèses.

**Définitions 54 (Statistique du rapport de vraisemblance)**

Étant donnés deux ensembles $\Theta_0$ et $\Theta_1$ représentant les deux hypothèses, nous introduisons la **statistique du rapport de vraisemblance** notée $h$ par :

$$h(X) = \frac{\sup_{\theta\in\Theta_1} V_X(\theta)}{\sup_{\theta\in\Theta_0} V_X(\theta)}$$

où nous rappelons que $V_X(\theta) = f_\theta(X)$ est la vraisemblance du modèle. La zone de rejet naturelle est donc de la forme $]k_\alpha, +\infty[$ où $k_\alpha$ est à fixer suivant le contexte et le **test du rapport de vraisemblance** s'écrit donc $\phi(X) = \mathbb{1}_{\{h(X)>k_\alpha\}}$.

**Remarque**

Si la valeur $\sup_{\theta\in\Theta_0} V_X(\theta)$ est nulle, nous supposons que le test vaut 1 puisque le modèle 0 est rejeté. Toutefois, si le numérateur **et** le dénominateur sont nuls, il faudra remettre en cause la pertinence des hypothèses.

**Exemple**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ iid de loi $\mathcal{N}(\theta^\star, 1)$ avec $\theta^\star \in \mathbb{R}$, nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star \geq 1, \\ \mathcal{H}_1 : & \theta^\star < 1. \end{cases}$$

La vraisemblance s'écrit alors :

$$V_X(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right]$$

est donc strictement croissante sur $]-\infty, \overline{\mathbf{X}}_n]$ et strictement décroissante sur $[\overline{\mathbf{X}}_n, +\infty[$. Ainsi, nous avons le rapport de vraisemblance suivant :

$$h(\mathbf{X}) = \begin{cases} \frac{V_{\mathbf{X}}(\overline{\mathbf{X}}_n)}{V_{\mathbf{X}}(1)} & \text{si } \overline{\mathbf{X}}_n < 1, \\ \frac{V_{\mathbf{X}}(1)}{V_{\mathbf{X}}(\overline{\mathbf{X}}_n)} & \text{si } \overline{\mathbf{X}}_n \geq 1. \end{cases}$$

Or, pour tout $\theta_0$ et $\theta_1$ de $\mathbb{R}$, nous avons :

$$\frac{V_{\mathbf{X}}(\theta_1)}{V_{\mathbf{X}}(\theta_0)} = \exp\left(\frac{n}{2}\left[\theta_0^2 - \theta_1^2 + 2\overline{\mathbf{X}}_n(\theta_1 - \theta_0)\right]\right)$$

de sorte que le rapport de vraisemblance s'écrit $h(X) = \exp\left[\frac{n}{2}\left(\overline{\mathbf{X}}_n - 1\right)^2\left(2\mathbb{1}_{\{\overline{\mathbf{X}}_n<1\}} - 1\right)\right]$. Or, il s'agit d'une fonction strictement décroissante en $\overline{\mathbf{X}}_n$ donc le test du rapport de vraisemblance peut se réécrire de la forme $\mathbb{1}_{\{\overline{\mathbf{X}}_n<k_\alpha\}}$.

**Théorème 20 (Lemme de Neyman-Pearson)**

Étant donnés un modèle paramétré par $\Theta$ et deux points $\theta_0$ et $\theta_1$ de $\Theta$ distincts, nous voulons tester :

$$\begin{cases} \mathcal{H}_0 : & \theta^\star = \theta_0, \\ \mathcal{H}_1 : & \theta^\star = \theta_1. \end{cases}$$

Étant donné $\alpha \in ]0,1[$, si nous supposons qu'il existe un seul $k_\alpha$ tel que le test de rapport de vraisemblance $\phi(X) = \mathbb{1}_{\{h(X;\theta_0,\theta_1)>k_\alpha\}}$ avec

$$h(X;\theta_0,\theta_1) = \frac{\sup\limits_{\theta\in\Theta_1} V_X(\theta)}{\sup\limits_{\theta\in\Theta_0} V_X(\theta)} = \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

vérifie $\mathbb{P}_{\theta_0}(\phi(X) = 1) = \alpha$ alors ce test est $UPP(\alpha)$.

**Preuve**

Nous considérons qu'il existe un autre test $\Psi$ de niveau $\alpha$ tel que $\mathbb{E}_{\theta_0}[\Psi(X)] = \mathbb{P}_{\theta_0}(\Psi(X) = 1) \leq \alpha$. Nous avons donc par construction $\mathbb{E}_{\theta_0}[\phi(X) - \Psi(X)] \geq 0$. Nous voulons donc montrer que le test de rapport de vraisemblance est au moins aussi puissant c'est-à-dire que

$$\mathbb{E}_{\theta_1}[\phi(X) - \Psi(X)] = \mathbb{P}_{\theta_1}(\phi(X) = 1) - \mathbb{P}_{\theta_1}(\Psi(X) = 1) \geq 0.$$

Pour ce faire, nous allons faire un changement de densité en remarquant que partout où la densité $\theta_0$ est non nulle, nous avons

$$f_{\theta_1}(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} f_{\theta_0}(x).$$

Nous avons donc :

$$\begin{aligned} &\mathbb{E}_{\theta_1}[\phi(X) - \Psi(X)] - k_\alpha \mathbb{E}_{\theta_0}[\phi(X) - \Psi(X)] \\ = \;& \mathbb{E}_{\theta_0}\left[(\phi(X) - \Psi(X))\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right] + \mathbb{E}_{\theta_1}\left[(\phi(X) - \Psi(X))\mathbb{1}_{\{f_{\theta_0}(X)=0\}}\right] - k_\alpha \mathbb{E}_{\theta_0}[\phi(X) - \Psi(X)] \\ = \;& \mathbb{E}_{\theta_0}\left[(\phi(X) - \Psi(X))\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} - k_\alpha\right)\right] + \mathbb{E}_{\theta_1}\left[(\phi(X) - \Psi(X))\mathbb{1}_{\{f_{\theta_0}(X)=0\}}\right] \text{ car } k_\alpha \text{ fini si } \alpha \in ]0,1[. \end{aligned}$$

Or, $\phi(X) = \mathbb{1}_{\{h(X;\theta_0,\theta_1)>k_\alpha\}}$ donc si elle vaut 0 nous avons :

$$\mathbb{E}_{\theta_0}\left[-\underbrace{\Psi(X)}_{>0}\underbrace{\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} - k_\alpha\right)}_{<0}\right] > 0,$$

et si elle vaut 1, l'espérance est également positive par construction. De plus, pour les valeurs $x$ telles que $f_{\theta_0}(x) = 0$, l'hypothèse nulle ne peut pas être choisie donc comme $k_\alpha$ est fini, la statistique vaut 1. Au final, comme $k_\alpha > 0$, nous avons donc :

$$\mathbb{E}_{\theta_1}\left[\phi(X) - \Psi(X)\right] \geq k_\alpha \mathbb{E}_{\theta_0}\left[\phi(X) - \Psi(X)\right] \geq 0.$$

## 5.5 Démarche générale

S'il ne faut retenir qu'une seule chose de ce cours, c'est la démarche générale. Toute la suite consistera à l'appliquer dans des cas particuliers :

**Point méthode**

Démarche générale des tests La démarche générale pour faire un test est la suivante :

1. Bien comprendre la problématique pour proposer une modélisation adéquate.
2. Définir l'hypothèse $\mathcal{H}_0$ qui sera la plus conservatrice.
3. Définir l'hypothèse $\mathcal{H}_1$ alternative.
4. Définir la statistique de test et la région de rejet. Pour cela, il sera nécessaire de :
   - Proposer un estimateur du paramètre d'intérêt.
   - Définir la loi sous l'hypothèse $\mathcal{H}_0$, éventuellement asymptotique, de cet estimateur.
   - Étant donné $\alpha \in [0; 1]$, proposer une règle de décision de niveau de confiance $\alpha$, éventuellement asymptotique.
5. Étudier la puissance du test en déterminant la probabilité de rejeter à raison l'hypothèse $\mathcal{H}_0$ :
   - Soit de façon théorique si des calculs sont possibles.
   - Soit de façon empirique à l'aide de simulations.

**Exercices 5.1**

Une entreprise d'appels téléphoniques possède 10 personnes dont le contrat consiste à gérer une centaine d'appels par jour. Les syndicats affirment qu'ils ne sont pas assez nombreux pour traiter toutes les demandes journalières et qu'il faudrait embaucher une personne de plus. L'entreprise vous embauche afin de savoir s'il est effectivement nécessaire de recruter quelqu'un en plus ou pas et, pour cela, elle vous fournit le nombre d'appels journaliers des $n$ derniers jours. Quel test proposez-vous ?

*Indications :* La loi de Poisson est souvent utilisée pour les comptages. De plus, si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes suivant des lois de Poisson de paramètres $\lambda_1, \ldots, \lambda_n$ alors la somme $\sum_{i=1}^n X_i$ est une variable aléatoire de Poisson de paramètre $\sum_{i=1}^n \lambda_i$.

## 5.6 Solutions des exercices

**Exercices 5.1**

1. Nous supposons que le nombre d'appels journaliers suit une loi de Poisson de paramètre $\theta^\star$.
2. L'hypothèse la plus conservatrice pour l'entreprise est que les 10 personnes suffisent. Comme chaque personne peut faire en moyenne 100 appels, cela représente 1000 appels par jour donc l'hypothèse $\mathcal{H}_0$ est :
   $$\mathcal{H}_0 \ : \quad \theta^\star = 1000$$
   ou, autrement dit, $\Theta_0 = \{1000\}$.
3. L'hypothèse proposée par les syndicats est qu'il manque une personne donc qu'il y a plutôt 1100 appels quotidiens. L'hypothèse $\mathcal{H}_1$ est donc :
   $$\mathcal{H}_1 \ : \quad \theta^\star = 1100$$

ou, autrement dit, $\Theta_1 = \{1100\}$.

4. Comme nous sommes dans un cas où les deux hypothèses sont simples, nous savons par le lemme de Neyman-Pearson que le test du rapport de vraisemblance est *uniformément plus puissant*. Nous commençons donc par calculer la vraisemblance étant donnée un paramètre $\theta \in \mathbb{R}_+^\star$ quelconque :

$$\begin{aligned} V_{\mathbf{X}}(\theta) &= \prod_{i=1}^{n} \left( e^{-\theta} \frac{\theta^{X_i}}{X_i!} \right) \\ &= e^{\sum_{i=1}^{n}(-\theta)} \frac{\theta^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n}(X_i!)} \\ &= e^{-n\theta} \frac{\theta^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n}(X_i!)}. \end{aligned}$$

Ainsi, la statistique de test est :

$$\begin{aligned} h(X; 1000, 1100) &= \frac{\sup_{\theta \in \Theta_1} V_X(\theta)}{\sup_{\theta \in \Theta_0} V_X(\theta)} \\ &= \frac{V_{\mathbf{X}}(1100)}{V_{\mathbf{X}}(1000)} \\ &= \frac{e^{-1100n} \frac{1100^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n}(X_i!)}}{e^{-1000n} \frac{1000^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n}(X_i!)}} \\ &= e^{-1100n+1000n} \left( \frac{1100}{1000} \right)^{\sum_{i=1}^{n} X_i} \\ &= e^{-100n} \left( \frac{11}{10} \right)^{\sum_{i=1}^{n} X_i}. \end{aligned}$$

La statistique de test est donc du type :

$$\phi(\mathbf{x}) = \mathbb{1}_{\{h(X;1000,1100) > k_\alpha\}} \tag{5.1}$$

où $k_\alpha$ est choisi tel que le niveau de confiance soit de $\alpha$. Or, la fonction $x \mapsto e^{-100n} \left(\frac{11}{10}\right)^x$ est croissante en $x$ donc l'équation (5.1) revient à trouver une valeur $k_\alpha^{\text{somme}}$ telle que :

$$\phi(\mathbf{x}) = \mathbb{1}_{\left\{\sum_{i=1}^{n} X_i > k_\alpha^{\text{somme}}\right\}} \tag{5.2}$$

ou une valeur $k_\alpha^{\text{moyenne}}$ telle que :

$$\phi(\mathbf{x}) = \mathbb{1}_{\left\{\overline{\mathbf{X}}_n > k_\alpha^{\text{moyenne}}\right\}}. \tag{5.3}$$

Deux solutions s'offre à nous :

- Comme, sous $\mathcal{H}_0$, les $X_i$ sont un $n$-échantillon de loi $\mathcal{P}(1000)$ alors $\sum_{i=1}^{n} X_i$ suit une loi de Poisson de paramètre $\sum_{i=1}^{n} 1000 = 1000n$ et nous pouvons prendre pour $k_\alpha^{\text{somme}}$ le quantile d'ordre $1-\alpha$ de la loi de Poisson $\mathcal{P}(1000n)$. Dans ce cas, le test (5.2) sera quasiment de niveau exact $\alpha$ (*quasiment* car la loi étant entière, il y a une petite fluctuation. Nous verrons cela en TP).
- Sinon, sous $\mathcal{H}_0$, $\overline{\mathbf{X}}_n$ suit asymptotiquement une loi normale de moyenne 1000 et de variance $1000/n$, nous pouvons recentrer et normaliser les termes et avoir le test (5.3) de niveau asymptotique $\alpha$.

# Chapitre 6

# Tests sur les variables quantitatives

## 6.1 Objectifs

Comprendre d'où viennent les tests relatifs aux variables quantitatives.

## 6.2 Test d'une espérance

Dans le cadre de variables quantitatives, les premières statistiques étudiées sont souvent la position et, en particulier, l'espérance. Ainsi, nous disposons d'un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de variables aléatoires suivant une loi d'espérance $\mu^\star$ et de variance $\sigma^2$ ; tous les deux inconnus. Dans notre cas, nous cherchons à savoir si la valeur de l'espérance vaut une valeur $\mu_0$ particulière ou si elle vaut autre chose :

$$\begin{cases} \mathcal{H}_0 : & \mu^\star = \mu_0, \\ \mathcal{H}_1 : & \mu^\star \neq \mu_0. \end{cases}$$

**Remarque**

Suivant le contexte, nous pouvons avoir une information supplémentaire sur la forme de $\mu^\star$ dans l'hypothèse alternative par exemple :

$$\mathcal{H}_1 : \mu^\star < \mu_0 \text{ ou } \mathcal{H}_1 : \mu^\star > \mu_0.$$

Dans ce cas, nous adapterons les zones des prédictions.

Pour estimer $\mu^\star$, nous pouvons prendre l'estimateur des moments $\overline{\mathbf{X}}_n$ (qui est, pour beaucoup de lois classiques, l'estimateur du maximum de vraisemblance également).

### 6.2.1 Cas gaussien avec la variance connue

Nous commençons par supposer que nous connaissons la variance $\sigma^2$. Dans ce cas, nous avons la loi suivante :

**Proposition 21 (Loi de la variable $\overline{\mathbf{X}}_n$ dans le cas gaussien avec variance $\sigma^2$ connue)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi gaussienne $\mathcal{N}\left(\mu^\star, \sigma^2\right)$ avec $\sigma^2$ connue alors la loi de $\overline{\mathbf{X}}_n$ suit une loi gaussienne $\mathcal{N}\left(\mu^\star, \sigma^2/n\right)$ :

$$\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu^\star, \sigma^2/n\right).$$

**Preuve**

Pour montrer cette proposition, nous avons besoin d'un résultat sur les vecteurs gaussiens :

**Lemme 22 (Linéarité des vecteurs gaussiens)**

Étant donnée $\mathbf{X} = (X_1, \dots, X_d)$ un vecteur gaussien d'espérance $m$ et de matrice de variance covariance $\Sigma$ alors pour toute matrice $A \in \mathcal{M}_{k\times d}(\mathbb{R})$ et pour tout vecteur $b \in \mathbb{R}^k$, le vecteur $A\mathbf{X} + b$ est gaussien de loi $\mathcal{N}(Am + b, A\Sigma A^T)$.

*Démonstration.* Ce lemme est une conséquence de la définition d'un vecteur gaussien. Pour montrer cette propriété, nous avons besoin de la forme de la fonction caractéristique et des propriétés de cette dernière. Pour une démonstration plus complète, nous encourageons les lectrices et lecteurs intéressé · e · s à lire le polycopié de Le Gall (2006) [1]ou la proposition 16 du polycopié de Brault (2021a) [2]. □

Le lemme 22 a pour conséquence que toute combinaison linéaire de variables gaussiennes suit une loi gaussienne (éventuellement triviale). Ainsi, nous savons que $\overline{\mathbf{X}}_n$ suit une loi normale et il ne reste qu'à connaître les paramètres. Par linéarité de l'espérance, nous avons :

$$\mathbb{E}\left[\overline{\mathbf{X}}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu^\star = \frac{n\mu^\star}{n} = \mu^\star.$$

Et pour la variance, comme les variables sont indépendantes, nous avons :

$$\begin{aligned}
\mathbb{V}\left[\overline{\mathbf{X}}_n\right] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \quad \text{car la variance est quadratique,} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[X_i\right] \quad \text{car les observations sont indépendantes,} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}$$

Ainsi, nous obtenons une loi gaussienne de moyenne $\mu^\star$ et de variance $\sigma^2/n$.

Ainsi, si nous connaissons la variance dans le cas de loi gaussienne, nous avons le point méthode suivant :

**Point méthode (Test d'espérance ; cas gaussien, variance connue)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi gaussienne $\mathcal{N}\left(\mu^\star, \sigma^2\right)$ avec $\sigma^2$ connue alors la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}$$

qui a pour loi une gaussienne centrée réduite sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star \neq \mu_0$ | $\mu^\star < \mu_0$ | $\mu^\star > \mu_0$ |
|---|---|---|---|
| $\Theta_1$ | $\mathbb{R} \setminus \{\mu_0\}$ | $]-\infty, \mu_0[$ | $]\mu_0; +\infty[$ |
| Règle | $\lvert T_n\rvert > q_{1-\alpha/2}$ | $T_n < q_\alpha$ | $T_n > q_{1-\alpha}$ |

où $q_\alpha$ est le quantile d'ordre $\alpha$ d'une loi normale centrée réduite ; c'est-à-dire que si $X$ suit une loi gaussienne centrée réduite $\mathcal{N}(0,1)$ alors $q_\alpha$ est tel que :

$$\mathbb{P}\left(X \leq q_\alpha\right) = \alpha.$$

1. Disponible à l'adresse `https://www.imo.universite-paris-saclay.fr/~jean-francois.le-gall/IPPA2.pdf`
2. Disponible à l'adresse `https://membres-ljk.imag.fr/Vincent.Brault/Cours/CoursStatistique_IF.pdf`

**Preuve**

À partir de la proposition 21 qui dit que $\overline{\mathbf{X}}_n$ suit une loi gaussienne d'espérance $\mu_0$ et de variance $\sigma^2/n$ et du lemme 22, nous montrons que, sous l'hypothèse $\mathcal{H}_0$ qui suppose que $\mu^\star = \mu_0$, $T_n$ suit une loi gaussienne dont il faut calculer l'espérance et la variance :

$$\mathbb{E}[T_n] = \mathbb{E}\left[\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}\right] = \sqrt{n}\frac{1}{\sigma}\left[\mathbb{E}\left[\overline{\mathbf{X}}_n\right] - \mu_0\right] = \sqrt{n}\frac{1}{\sigma}\left[\mu_0 - \mu_0\right] = 0.$$

Pour la variance, nous avons :

$$\begin{aligned}
\mathbb{V}[T_n] &= \mathbb{V}\left[\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}\right] \\
&= \frac{n}{\sigma^2}\left[\mathbb{V}\left[\overline{\mathbf{X}}_n\right] - \mathbb{V}[\mu_0]\right] \quad \text{car la variance est quadratique,} \\
&= \frac{n}{\sigma^2}\left[\sigma^2/n - 0\right] \quad \text{car la variance d'une constante est nulle,} \\
&= \frac{n\sigma^2}{n\sigma^2} = 1.
\end{aligned}$$

Et la loi de $T_n$ est une loi normale centrée réduite.

**Exercices 6.1**

Nous laissons en exercice le calcul de la puissance pour chaque règle de décision. En particulier, dans le cas où l'hypothèse $\mathcal{H}_1$ vaut $\mu^\star < \mu_0$ ou $\mu^\star > \mu_0$, montrer que la suite de tests est consistante.

### 6.2.2 Cas gaussien avec la variance inconnue

Dans le cas où la variance n'est pas connue, nous pouvons l'estimer. Dans ce cas, nous avons deux estimateurs possibles à savoir l'estimateur du maximum de vraisemblance ou basé sur les moments d'ordre 2

$$\widehat{\sigma}_n^2 = \overline{X^2}_n - \overline{\mathbf{X}}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)^2$$

ou sa version non biaisée :

$$S_n^2 = \frac{n}{n-1}\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)^2.$$

**Exercices 6.2**

Montrer que $\widehat{\sigma}_n^2$ est un estimateur biaisé de $\sigma^2$ alors que $S_n^2$ est sans biais.

Ainsi, nous allons étudier la statistique :

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{S_n}$$

dont nous devons trouver la loi :

**Proposition 23 (Loi de la variable $S_n$)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de la loi gaussienne $\mathcal{N}(\mu^\star, \sigma^2)$ avec $\sigma^2$ inconnue alors la variable $(n-1)S_n^2/\sigma^2$ suit une loi du $\chi^2(n-1)$ ; c'est-à-dire une loi du $\chi^2$ à $n-1$ degrés de libertés.

$$(n-1)S_n^2/\sigma^2 = \sum_{i=1}^n \frac{\left(X_i - \overline{\mathbf{X}}_n\right)^2}{\sigma^2} \sim \chi^2(n-1).$$

**Preuve**

La preuve se base sur un théorème que nous admettrons car il utilise des notions de projections (donc d'espérances conditionnelles) qui ne seront pas abordées (les lecteurices intéressé · e · s peuvent aller regarder le polycopié de Le Gall (2006) pour avoir la démonstration). Nous mettons toutefois l'énoncé ici.

**Lemme 24 (Théorème de Cochran)**

Étant donnés un vecteur gaussien $\mathbf{X}$ de $\mathbb{R}^d$ de loi $\mathcal{N}\left(m, \sigma^2 \mathbb{I}_d\right)$ et $F_1, \ldots, F_P$ des sous-espaces vectoriels de $\mathbb{R}^d$, orthogonaux deux à deux et de somme $\mathbb{R}^d$ dont nous notons $P_{F_p}$ les matrices de projection orthogonale sur $F_p$ et $n_p$ la dimension de $F_p$ alors nous avons :

- les vecteurs aléatoires $P_{F_1}\mathbf{X}, \ldots, P_{F_P}\mathbf{X}$ sont deux à deux indépendants et de lois respectives $\mathcal{N}\left(P_{F_1} m, \sigma^2 P_{F_1}\right), \ldots, \mathcal{N}\left(P_{F_P} m, \sigma^2 P_{F_P}\right)$.
- les variables aléatoires $\frac{\left\|P_{F_1}(\mathbf{X}-m)\right\|^2}{\sigma^2}, \ldots, \frac{\left\|P_{F_P}(\mathbf{X}-m)\right\|^2}{\sigma^2}$ sont deux à deux indépendantes et sont de lois respectives $\chi^2(n_1), \ldots, \chi^2(n_P)$.

Pour démontrer la proposition 23, nous commençons par définir pour tout $i \in \{1, \ldots, n\}$, la variable :

$$Z_i = \frac{X_i - \overline{\mathbf{X}}_n}{\sigma}.$$

Calculons l'espérance et la matrice de variance-covariance de la variable $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Sous l'hypothèse $\mathcal{H}_0$, nous avons de plus pour tout $i \in \{1, \ldots, n\}$ :

$$\begin{aligned}
\mathbb{E}[Z_i] &= \mathbb{E}\left[\frac{X_i - \overline{\mathbf{X}}_n}{\sigma}\right] \\
&= \frac{\mathbb{E}[X_i] - \mathbb{E}\left[\overline{\mathbf{X}}_n\right]}{\sigma} \\
&= \frac{\mu^\star - \mu^\star}{\sigma} \\
&= 0.
\end{aligned}$$

De plus, pour tout $i \in \{1, \ldots, n\}$, nous avons :

$$\begin{aligned}
\mathbb{V}[Z_i] &= \mathbb{V}\left[\frac{X_i - \overline{\mathbf{X}}_n}{\sigma}\right] \\
&= \frac{1}{\sigma^2}\mathbb{V}\left[X_i - \overline{\mathbf{X}}_n\right] \\
&= \frac{1}{\sigma^2}\left\{\mathbb{V}[X_i] + \mathbb{V}\left[\overline{\mathbf{X}}_n\right] - 2\mathrm{Cov}\left(X_i, \overline{\mathbf{X}}_n\right)\right\} \\
&= \frac{1}{\sigma^2}\left\{\sigma^2 + \frac{\sigma^2}{n} - 2\mathrm{Cov}\left(X_i, \frac{1}{n}\sum_{i'=1}^{n} X_{i'}\right)\right\} \\
&= \frac{n+1}{n} - \frac{2}{n\sigma^2}\sum_{i'=1}^{n}\mathrm{Cov}\left(X_i, X_{i'}\right) \\
&= \frac{n+1}{n} - \frac{2}{n\sigma^2}\left[\sum_{\substack{i'=1 \\ i' \neq i}}^{n} \underbrace{\mathrm{Cov}\left(X_i, X_{i'}\right)}_{=0} + \mathrm{Cov}\left(X_i, X_i\right)\right] \\
&= \frac{n+1}{n} - \frac{2}{n\sigma^2}\mathbb{V}[X_i] \\
&= \frac{n-1}{n}.
\end{aligned}$$

Donc la loi commune des $Z_i$ est centrée et matrice de variance covariance :

$$\Gamma_n = \mathbb{I}_n - \begin{pmatrix} 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{pmatrix}\begin{pmatrix} 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{pmatrix}.$$

La loi de la variable $\frac{Z_1+\ldots+Z_n}{n}$ est une loi normale centrée de covariance $\Gamma_n$. Or, cette loi est le projeté d'un vecteur aléatoire de $\mathbb{R}^K$ suivant une loi normale centrée réduite sur l'hyperplan orthogonal au vecteur colonne $(1/\sqrt{n}, \cdots, 1/\sqrt{n})^T$.
De plus, nous voyons que la variable $(n-1)S_n^2/\sigma^2$ est la norme au carré de cette variable donc, d'après le théorème de Cochran, la loi asymptotique est une loi du $\chi^2$ à $n-1$ degrés de libertés.

Ainsi, nous obtenons le corollaire suivant :

**Corollaire 25 (Loi de la statistique)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de la loi gaussienne $\mathcal{N}(\mu^\star, \sigma^2)$ avec $\sigma^2$ inconnue alors la loi de

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{S_n}$$

suit une loi de Student centrée à $n-1$ degrés de liberté :

$$T_n \sim \mathcal{T}(n-1).$$

**Preuve**

Avant de faire la démonstration, rappelons la définition d'une loi de Student :

**Définition 55 (Loi de Student)**

Soient $Z$ une variable de loi gaussienne centrée réduite et $U$ une variable aléatoire indépendante de $Z$ et distribuée suivant une loi du $\chi^2$ à $k$ degrés de liberté alors la variable :

$$\frac{Z}{\sqrt{U/k}}$$

suit une loi de Student centrée à $k$ degrés de liberté $\mathcal{T}(n-1)$.

Nous avons :

$$\begin{aligned}\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sqrt{S_n^2}} &= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sigma}}{\frac{\sqrt{S_n^2}}{\sigma}} \\ &= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sigma}}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{Y}{n-1}}}\end{aligned}$$

avec $Z$ qui suit une loi normale centrée réduite et $Y$ qui suit une loi du $\chi^2$ à $n-1$ degrés de liberté indépendante d'après le théorème 24 de Cochran donc, par définition de la loi de Student, nous avons le résultat.

Ainsi, nous avons :

**Point méthode (Test d'espérance ; cas gaussien, variance inconnue)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi gaussienne $\mathcal{N}(\mu^\star, \sigma^2)$ avec $\sigma^2$ inconnue alors la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{S_n}$$

qui a pour loi une loi de Student centrée à $n-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la

règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star \neq \mu_0$ | $\mu^\star < \mu_0$ | $\mu^\star > \mu_0$ |
|---|---|---|---|
| $\Theta_1$ | $\mathbb{R} \setminus \{\mu_0\}$ | $]-\infty, \mu_0[$ | $]\mu_0; +\infty[$ |
| Règle | $\lvert T_n \rvert > s_{1-\alpha/2}^{(n-1)}$ | $T_n < s_\alpha^{(n-1)}$ | $T_n > s_{1-\alpha}^{(n-1)}$ |

où $s_\alpha^{(n-1)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $n-1$ degrés de liberté $\mathcal{T}(n-1)$ alors $s_\alpha^{(n-1)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(n-1)}\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rnorm(100)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,mu=0)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,mu=1)
```

### Remarque

Nous pouvons aussi utiliser $\widehat{\sigma}_n$ à la place de $S_n$ dans le calcul de la statistique de $T_n$ mais, dans ce cas, le test est de niveau asymptotique seulement.

**Point méthode (Test d'espérance ; cas gaussien, variance inconnue avec $\widehat{\sigma}_n$)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi gaussienne $\mathcal{N}\left(\mu^\star, \sigma^2\right)$ avec $\sigma^2$ inconnue alors la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\widehat{\sigma}_n}$$

qui a pour loi **asymptotique** une loi de Student centrée à $n-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star \neq \mu_0$ | $\mu^\star < \mu_0$ | $\mu^\star > \mu_0$ |
|---|---|---|---|
| $\Theta_1$ | $\mathbb{R} \setminus \{\mu_0\}$ | $]-\infty, \mu_0[$ | $]\mu_0; +\infty[$ |
| Règle | $\lvert T_n \rvert > s_{1-\alpha/2}^{(n-1)}$ | $T_n < s_\alpha^{(n-1)}$ | $T_n > s_{1-\alpha}^{(n-1)}$ |

où $s_\alpha^{(n-1)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi gde Student centrée à $n-1$ degrés de liberté $\mathcal{T}(n-1)$ alors $s_\alpha^{(n-1)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(n-1)}\right) = \alpha.$$

### Preuve

Par la même démonstration que pour la proposition 23, nous montrons que la variable $n\widehat{\sigma}_n^2/\sigma^2$ suit une loi du $\chi^2(n-1)$ ; c'est-à-dire une loi du $\chi^2$ à $n-1$ degrés de libertés. Ainsi, nous avons :

$$\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sqrt{\widehat{\sigma}_n^2}} = \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sigma}}{\frac{\sqrt{\widehat{\sigma}_n^2}}{\sigma}}$$

$$
\begin{aligned}
&= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n-\mu^\star}{\sigma}}{\frac{1}{\sqrt{n}}\sqrt{\frac{n\widehat{\sigma}_n^2}{\sigma^2}}} \\
&= \frac{1}{\sqrt{\frac{n}{n-1}}} \times \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n-\mu^\star}{\sigma}}{\sqrt{\frac{n\widehat{\sigma}_n^2}{\sigma^2(n-1)}}} \\
&= \underbrace{\frac{1}{\sqrt{\frac{n-1}{n}}}}_{\substack{\longrightarrow 1 \\ n\to+\infty}} \times \frac{Z}{\sqrt{\frac{Y}{n-1}}}
\end{aligned}
$$

avec $Z$ qui suit une loi normale centrée réduite et $Y$ qui suit une loi du $\chi^2$ à $n-1$ degrés de liberté indépendante d'après le théorème 24 de Cochran et comme asymptotiquement $n/(n-1)$ tend vers 1 alors, asymptotiquement et sous l'hypothèse que $\mu^\star = \mu_0$, $T_n$ suit une loi de Student centrée à $n-1$ degrés de liberté.

### 6.2.3 Test de la normalité

Comme vous pouvez le voir dans les sections 6.2.1 et 6.2.2, nous avons supposé que les variables étaient gaussiennes. Dans la section 6.2.4 suivante, nous verrons comment faire lorsque les variables ne semblent pas gaussiennes. Toutefois, la question qui se pose alors est de savoir comment tester si une variable est gaussienne ou non. Pour cela, il existe différents tests (voir le chapitre 8), nous présentons les principes de deux d'entre eux. En particulier, nous faisons le test suivant :

$$
\begin{cases}
\mathcal{H}_0 & : & \mathbf{X} \text{ suit une loi normale,} \\
\mathcal{H}_1 & : & \mathbf{X} \text{ ne suit pas une loi normale.}
\end{cases}
\tag{6.1}
$$

**Remarque**

Remarquons que le test n'empêche pas les corrélations entre les observations.

**Test de Shapiro-Wilk (voir Shapiro et Wilk (1965)) :** Le test de Shapiro-Wilk repose sur la même idée que celui des graphiques *QQ-plot* c'est-à-dire au fait d'ordonner les valeurs de la plus petite à la plus grande et de les comparer aux quantiles d'une loi gaussienne centrée réduite. Concrètement, étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$, nous organisons les variables de la plus petite à la plus grande $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ et nous les comparons aux espérances des quantiles d'une loi normale ayant les mêmes paramètres. Autrement dit, si nous notons $\mathbf{m} = (m_1, \ldots, m_n)$ les espérances des statistiques d'ordre d'un $n$-échantillon suivant une loi normale et $\widehat{\Sigma}_n$ la matrice de variance covariance estimée des observation :

$$
\forall (i,j) \in \{1,\ldots,n\}^2, \quad \left(\widehat{\Sigma}_n\right)_{i,j} = \mathrm{Cov}_n\left(X_i, X_j\right),
$$

nous comparons les observations aux constantes $\mathbf{a} = (a_1, \ldots, a_n)$ définies par :

$$
\mathbf{a} = \frac{\mathbf{m}^T\widehat{\Sigma}_n^{-1}}{\left[\mathbf{m}^T\widehat{\Sigma}_n^{-1}\widehat{\Sigma}_n^{-1}\mathbf{m}\right]^{1/2}}.
$$

Ainsi, nous avons la méthode suivante :

**Point méthode (Test de normalité ; méthode de Shapiro et Wilk (1965))**

Étant données des variables aléatoires $\mathbf{X} = (X_1, \ldots, X_n)$ de la loi gaussienne multivariée $\mathcal{N}_n\left(\boldsymbol{\mu}^\star, \Sigma_n^2\right)$ alors la statistique

$$
W_n = \frac{\left[\sum_{i=1}^n a_i X_{(i)}\right]^2}{\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)^2}
$$

suit une loi qui ne porte pas de nom mais dont les valeurs peuvent être calculées à l'aide de simulations de Monte-Carlo. La règle de décision du test (6.1) est alors :

$$W_n > w_{1-\alpha}^n$$

où $w_\alpha^n$ est le quantile d'ordre $\alpha$ estimé par simulation de Monte-Carlo (voir par exemple Royston (1982)).

### Codage en R

La fonction R permettant de faire le test de normalité en utilisant la méthode de Shapiro-Wilk est `shapiro.test`. Par exemple :

```
## Acceptation avec 95% de chances
shapiro.test(rnorm(100))
## Rejet normalement
shapiro.test(rexp(100))
```

### Remarque

Comme nous l'avons dit précédemment, ce test est à mettre en lien avec les graphiques $QQ$-plot. Sur la figure 6.1, nous avons représenté les $QQ$-plot pour 4 types de distributions différentes. Les deux du haut sont des distributions gaussiennes (indépendantes à gauche et corrélées à droite) ; les $p$-valeurs ne sont pas trop petites donc on va avoir tendance à conserver, à raison, l'hypothèse de normalité. Les deux distributions sont respectivement une loi exponentielle $\mathcal{E}(1)$ et une loi Cauchy $\mathcal{C}\text{au}(0,1)$ ; dans ces deux cas, l'hypothèse $\mathcal{H}_0$ est largement rejetée, à raison, au profit de l'hypothèse $\mathcal{H}_1$.

### Codage en R

Le code utilisé pour créer la figure 6.1 est le suivant :

```

#####
# Packages
#####
library(ggpubr) # Pour qqplot
library(grid) # Pour affichage multiple

#####
# Préparation pour représenter les 4 graphiques
#####
grid.newpage()
pushViewport(viewport(layout = grid.layout(2,2)))

###
# Simulation gaussienne iid
###
n<-100
dat<-data.frame(x=rnorm(n))
## Affichage
q<-ggqqplot(dat, x = "x",
            ggtheme = theme_pubclean())+
  ggtitle("Variables gaussiennes iid",subtitle = paste0("p-value=",signif(shapiro.test(dat$x)$p.value),2))
print(q,vp=viewport(layout.pos.row = 1, layout.pos.col = 1))

######
# Simulation gaussienne corrélée
######
## Triangulaire inférieure
P<-matrix(runif(n*n)*2-1,n,n)
P[abs(P)<0.3]<-0
```

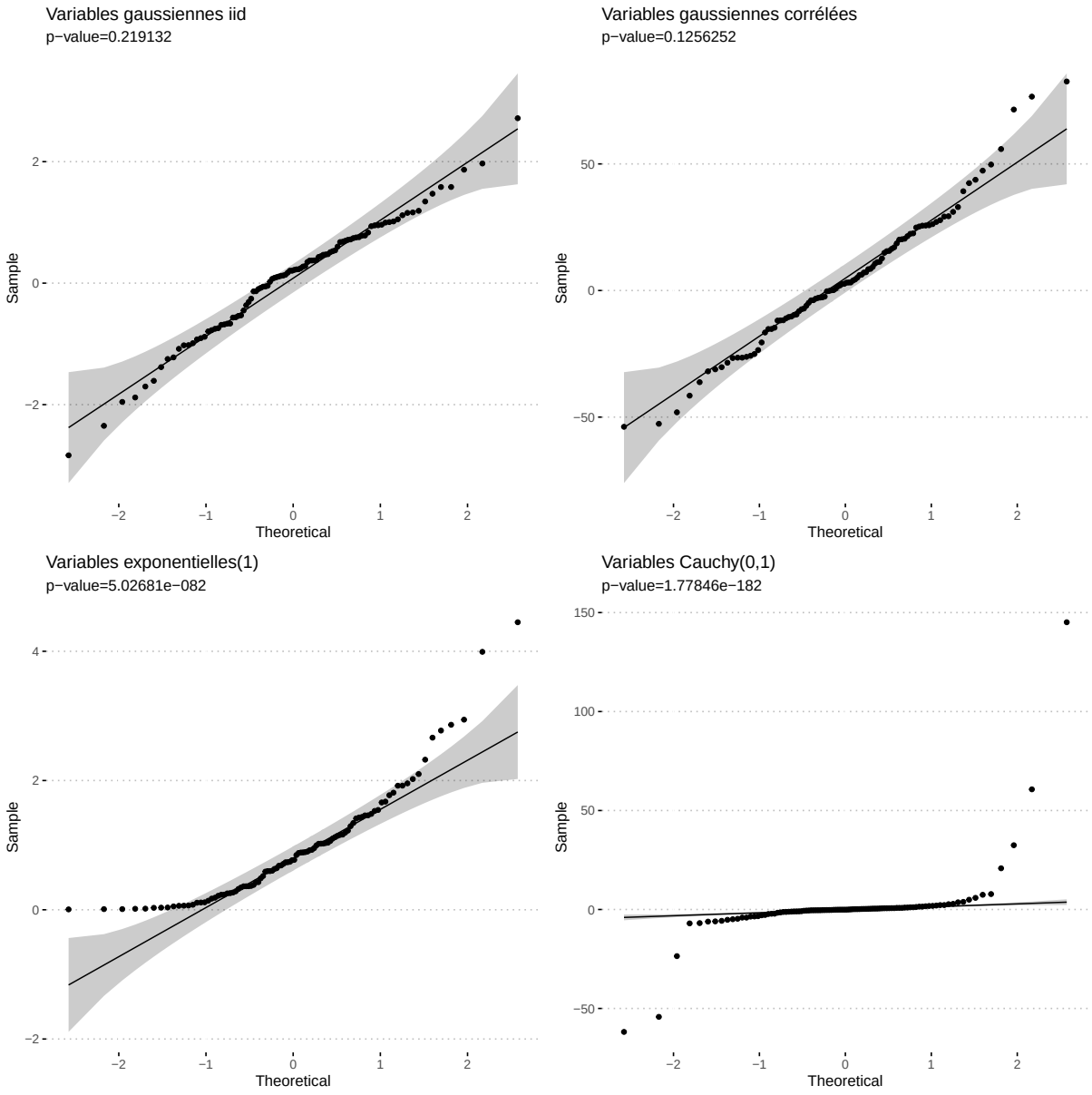



FIGURE 6.1 – *QQ*-plot de plusieurs 100-échantillons suivant différentes lois : loi gaussienne indépendante (en haut à gauche), loi gaussienne corrélée (en haut à droite), loi exponentielle $\mathcal{E}(1)$ (en bas à gauche) et loi Cauchy $\mathcal{C}\text{au}\,(0,1)$ (en bas à droite). La $p$-valeur obtenue par le test de Shapiro-Wilk a été ajouté à chaque fois dans le sous titre.

```
for (i in 1:(n-1)){
  for (j in (i+1):n){
    P[i,j]=0
  }
}
## Diagonale forte pour éviter les valeurs nulles
diag(P)<-runif(n)+0.5
Sigma<-P%*%t(P)
## Vérification de la matrice
sum(abs(Sigma-t(Sigma))) ## Symétrique
min(eigen(Sigma)$values) ## Définie positive
## Simulation
dat<-data.frame(x=Sigma%*%rnorm(n)+1)
## Affichage
```

```r
q<-ggqqplot(dat, x = "x",
            ggtheme = theme_pubclean())+
  ggtitle("Variables gaussiennes corrélées",subtitle = paste0("p-value=",signif(shapiro.test(dat$x)$p.value),2))
print(q,vp=viewport(layout.pos.row = 1, layout.pos.col = 2))

######
# Simulation loi exponentielle
######
## Simulation
dat<-data.frame(x=rexp(n))
## Affichage
q<-ggqqplot(dat, x = "x",
            ggtheme = theme_pubclean())+
  ggtitle("Variables exponentielles(1)",subtitle = paste0("p-value=",signif(shapiro.test(dat$x)$p.value),2))
print(q,vp=viewport(layout.pos.row = 2, layout.pos.col = 1))

######
# Simulation loi Cauchy
######
## Simulation
dat<-data.frame(x=rcauchy(n))
## Affichage
q<-ggqqplot(dat, x = "x",
            ggtheme = theme_pubclean())+
  ggtitle("Variables Cauchy(0,1)",subtitle = paste0("p-value=",signif(shapiro.test(dat$x)$p.value),2))
print(q,vp=viewport(layout.pos.row = 2, layout.pos.col = 2))
```

**Test de Lilliefors :** Le test de Liliefors est une adaptation du test de Kolmogorov-Smirnov qui est développé dans la section 8.1 sur les tests non paramétriques. Avant ça, nous avons besoin de redéfinir la fonction de répartition empirique (dont la définition a été donnée en section 19) :

**Définition 56 (Fonction de répartition empirique)**

Étant donné un $n$-échantillon $\mathbf{X}$ de variables réelles, nous appelons **fonction de répartition empirique** la fonction définie sur $\mathbb{R}$ par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Ainsi, nous avons la méthode suivante :

**Point méthode (Test de normalité ; méthode de Lilliefors (1967))**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de la loi gaussienne $\mathcal{N}\left(\mu^\star, \sigma^2\right)$ alors nous faisons la procédure suivante :

1. Calcul de $\overline{\mathbf{X}}_n$ et de $\widehat{\sigma}_n^2$ les estimateurs de l'espérance et de la variance de $\mathbf{X}$.
2. Calcul du maximum de la dispersion entre la fonction de répartition empirique $F_n$ de l'échantillon $\mathbf{X}$ et la fonction de répartition $F_{\mathcal{N}(\overline{\mathbf{X}}_n, \widehat{\sigma}_n^2)}$ d'une loi gaussienne ayant les paramètres estimés en 1 :

$$L_n = \sqrt{n}\, sup_{x\in\mathbb{R}} \left| F_{\mathcal{N}(\overline{\mathbf{X}}_n, \widehat{\sigma}_n^2)}(x) - F_n(x) \right|$$

alors la statistique $L_n$ suit une distribution de Lilliefors qui se calcule aussi par une méthode de Monte-Carlo. La règle de décision du test (6.1) est alors :

$$L_n > \ell_{1-\alpha}^n$$

où $\ell_\alpha^n$ est le quantile d'ordre $\alpha$ d'une distribution de Lilliefors estimé par simulation de Monte-Carlo.

### Codage en R

La fonction R permettant de faire le test de normalité en utilisant la méthode de Lilliefors est `lillie.test` du package `nortest`. Par exemple :

```
## Chargement du package
library(nortest)
## Acceptation avec 95% de chances
lillie.test(rnorm(100))
## Rejet normalement
lillie.test(rexp(100))
```

### Remarque

Comme nous l'avons dit précédemment, ce test est basé sur les fonctions de répartition. Sur la figure 6.2, nous avons représenté les fonctions de répartition empirique et les fonctions de répartition théoriques d'une loi gaussienne $\mathcal{N}\left(\overline{\mathbf{X}}_n, \widehat{\sigma}_n^2\right)$ pour 4 types de distributions différentes. Les deux du haut sont des distributions gaussiennes (indépendantes à gauche et corrélées à droite) ; les $p$-valeurs ne sont pas trop petites donc on va avoir tendance à conserver, à raison, l'hypothèse de normalité. Les deux distributions sont respectivement une loi exponentielle $\mathcal{E}(1)$ et une loi Cauchy $\mathcal{C}\text{au}\,(0,1)$ ; dans ces deux cas, l'hypothèse $\mathcal{H}_0$ est largement rejetée, à raison, au profit de l'hypothèse $\mathcal{H}_1$.

### Codage en R

Le code utilisé pour créer la figure 6.1 est le suivant :

```
#####
# Packages
#####
library(ggplot2)# Pour qqplot
library(grid) # Pour affichage multiple
library(nortest)# Pour le test

#####
# Préparation pour représenter les 4 graphiques
#####
grid.newpage()
pushViewport(viewport(layout = grid.layout(2,2)))

###
# Simulation gaussienne iid
###
n<-100
```

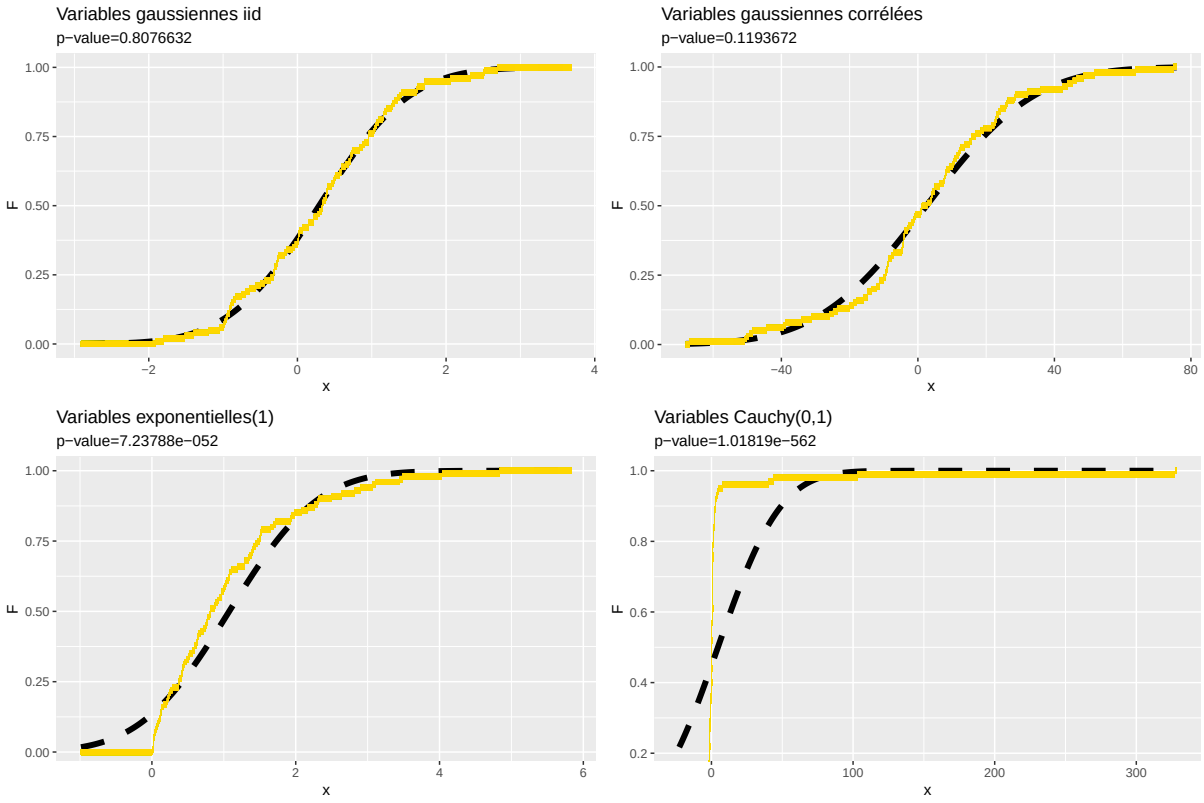



FIGURE 6.2 – Fonctions de répartition empirique (en or) et les fonctions de répartition théoriques (en noir) d'une loi gaussienne $\mathcal{N}\left(\overline{\mathbf{X}}_n, \widehat{\sigma}_n^2\right)$ de plusieurs 100-échantillons suivant différentes lois : loi gaussienne indépendante (en haut à gauche), loi gaussienne corrélée (en haut à droite), loi exponentielle $\mathcal{E}(1)$ (en bas à gauche) et loi Cauchy $\mathcal{C}\text{au}\,(0,1)$ (en bas à droite). La $p$-valeur obtenue par le test de Shapiro-Wilk a été ajouté à chaque fois dans le sous titre.

```
X<−rnorm(n)
### Fonction de répartition de référence
x_ref<−seq(min(X)−1,max(X)+1,by=0.1)
dat<−data.frame(x=x_ref,F=pnorm(x_ref,mean = mean(X),sd=sd(X)))
## Affichage
q<−ggplot(dat, aes(x = x,y=F))+geom_line(lwd=2,lty="dashed")+
  ggtitle("Variables gaussiennes iid",subtitle = paste0("p−value=",signif(lillie.test(X)$p.value),2))
## Ajout fonction de répartition empirique
X_ord<−c(x_ref[1],sort(X),x_ref[length(x_ref)])
for (it in 1:(length(X_ord)−1)){
  q<−q+geom_segment(x = X_ord[it],xend=X_ord[it+1],
                    y=(it−1)/n,yend=(it−1)/n,col="gold",alpha=0.5,lwd=2)
}
## Affichage
print(q,vp=viewport(layout.pos.row = 1, layout.pos.col = 1))

######
# Simulation gaussienne corrélée
######
## Triangulaire inférieure
P<−matrix(runif(n*n)*2−1,n,n)
P[abs(P)<0.3]<−0
for (i in 1:(n−1)){
  for (j in (i+1):n){
    P[i,j]=0
  }
}
```

```r
## Diagonale forte pour éviter les valeurs nulles
diag(P)<-runif(n)+0.5
Sigma<-P%*%t(P)
## Vérification de la matrice
sum(abs(Sigma-t(Sigma))) ## Symétrique
min(eigen(Sigma)$values) ## Définie positive
## Simulation
X<-Sigma%*%rnorm(n)+1
### Fonction de répartition de référence
x_ref<-seq(min(X)-1,max(X)+1,by=0.1)
dat<-data.frame(x=x_ref,F=pnorm(x_ref,mean = mean(X),sd=sd(X)))
## Affichage
q<-ggplot(dat, aes(x = x,y=F))+geom_line(lwd=2,lty="dashed")+
  ggtitle("Variables gaussiennes corrélées",subtitle = paste0("p-value=",signif(lillie.test(X)$p.value),2))
## Ajout fonction de répartition empirique
X_ord<-c(x_ref[1],sort(X),x_ref[length(x_ref)])
for (it in 1:(length(X_ord)-1)){
  q<-q+geom_segment(x = X_ord[it],xend=X_ord[it+1],
                    y=(it-1)/n,yend=(it-1)/n,col="gold",alpha=0.5,lwd=2)
}
## Affichage
print(q,vp=viewport(layout.pos.row = 1, layout.pos.col = 2))

######
# Simulation loi exponentielle
######
## Simulation
X<-rexp(n)
### Fonction de répartition de référence
x_ref<-seq(min(X)-1,max(X)+1,by=0.1)
dat<-data.frame(x=x_ref,F=pnorm(x_ref,mean = mean(X),sd=sd(X)))
## Affichage
q<-ggplot(dat, aes(x = x,y=F))+geom_line(lwd=2,lty="dashed")+
  ggtitle("Variables exponentielles(1)",subtitle = paste0("p-value=",signif(lillie.test(X)$p.value),2))
## Ajout fonction de répartition empirique
X_ord<-c(x_ref[1],sort(X),x_ref[length(x_ref)])
for (it in 1:(length(X_ord)-1)){
  q<-q+geom_segment(x = X_ord[it],xend=X_ord[it+1],
                    y=(it-1)/n,yend=(it-1)/n,col="gold",alpha=0.5,lwd=2)
}
## Affichage
print(q,vp=viewport(layout.pos.row = 2, layout.pos.col = 1))

######
# Simulation loi Cauchy
######
## Simulation
X<-rcauchy(n)
### Fonction de répartition de référence
x_ref<-seq(min(X)-1,max(X)+1,by=0.1)
dat<-data.frame(x=x_ref,F=pnorm(x_ref,mean = mean(X),sd=sd(X)))
## Affichage
q<-ggplot(dat, aes(x = x,y=F))+geom_line(lwd=2,lty="dashed")+
  ggtitle("Variables Cauchy(0,1)",subtitle = paste0("p-value=",signif(lillie.test(X)$p.value),2))
## Ajout fonction de répartition empirique
X_ord<-c(x_ref[1],sort(X),x_ref[length(x_ref)])
for (it in 1:(length(X_ord)-1)){
  q<-q+geom_segment(x = X_ord[it],xend=X_ord[it+1],
                    y=(it-1)/n,yend=(it-1)/n,col="gold",alpha=0.5,lwd=2)
}
## Affichage
print(q,vp=viewport(layout.pos.row = 2, layout.pos.col = 2))
```

### 6.2.4 Cas non gaussien

Enfin, si le test de normalité rejette l'hypothèse que la distribution soit gaussienne au profit d'une distribution gaussienne mais que nous admettons que la distribution ait un moment d'ordre 2 alors nous pouvons réutiliser la statistique de la section 6.2.2 mais avec une loi asymptotique :

**Point méthode (Test d'espérance ; cas non gaussien, variance inconnue avec $S_n$)**

Étant donné un $n$-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ de loi admettant un moment d'ordre 2 alors la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{S_n}$$

qui a pour loi **asymptotique** une loi de Student centrée à $n-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star \neq \mu_0$ | $\mu^\star < \mu_0$ | $\mu^\star > \mu_0$ |
|---|---|---|---|
| $\Theta_1$ | $\mathbb{R} \setminus \{\mu_0\}$ | $]-\infty, \mu_0[$ | $]\mu_0; +\infty[$ |
| Règle | $\lvert T_n\rvert > s_{1-\alpha/2}^{(n-1)}$ | $T_n < s_\alpha^{(n-1)}$ | $T_n > s_{1-\alpha}^{(n-1)}$ |

où $s_\alpha^{(n-1)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi gde Student centrée à $n-1$ degrés de liberté $\mathcal{T}(n-1)$ alors $s_\alpha^{(n-1)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(n-1)}\right) = \alpha.$$

**Preuve**

D'après le théorème de la limite centrale et comme la distribution admet un moment d'ordre 2, la loi de $\overline{\mathbf{X}}_n$ converge vers une loi gaussienne d'espérance $\mu^\star$ et de variance inconnue que nous noterons $\sigma^2$. Par la même démonstration que pour la proposition 23, nous montrons que la variable $n\widehat{\sigma}_n^2/\sigma^2$ suit une loi asymptotique du $\chi^2(n-1)$ ; c'est-à-dire une loi du $\chi^2$ à $n-1$ degrés de libertés. Ainsi, nous avons :

$$\begin{aligned}
\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sqrt{\widehat{\sigma}_n^2}} &= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sigma}}{\frac{\sqrt{\widehat{\sigma}_n^2}}{\sigma}} \\
&= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu^\star}{\sigma}}{\sqrt{\frac{(n-1)\widehat{\sigma}_n^2}{\sigma^2(n-1)}}} \\
&= \times \frac{Z}{\sqrt{\frac{Y}{n-1}}}
\end{aligned}$$

avec $Z$ qui suit une loi asymptotique normale centrée réduite et $Y$ qui suit une loi asymptotique du $\chi^2$ à $n-1$ degrés de liberté indépendante d'après le théorème 24 de Cochran alors, asymptotiquement et sous l'hypothèse que $\mu^\star = \mu_0$, $T_n$ suit une loi de Student centrée à $n-1$ degrés de liberté.

### 6.2.5 Quel(s) test(s) utilisé(s) ?

Pour faire le bilan de cette partie, nous avons créé un arbre de décision pour savoir quel(s) test(s) utilisé(s) disponible sur la figure 6.3.

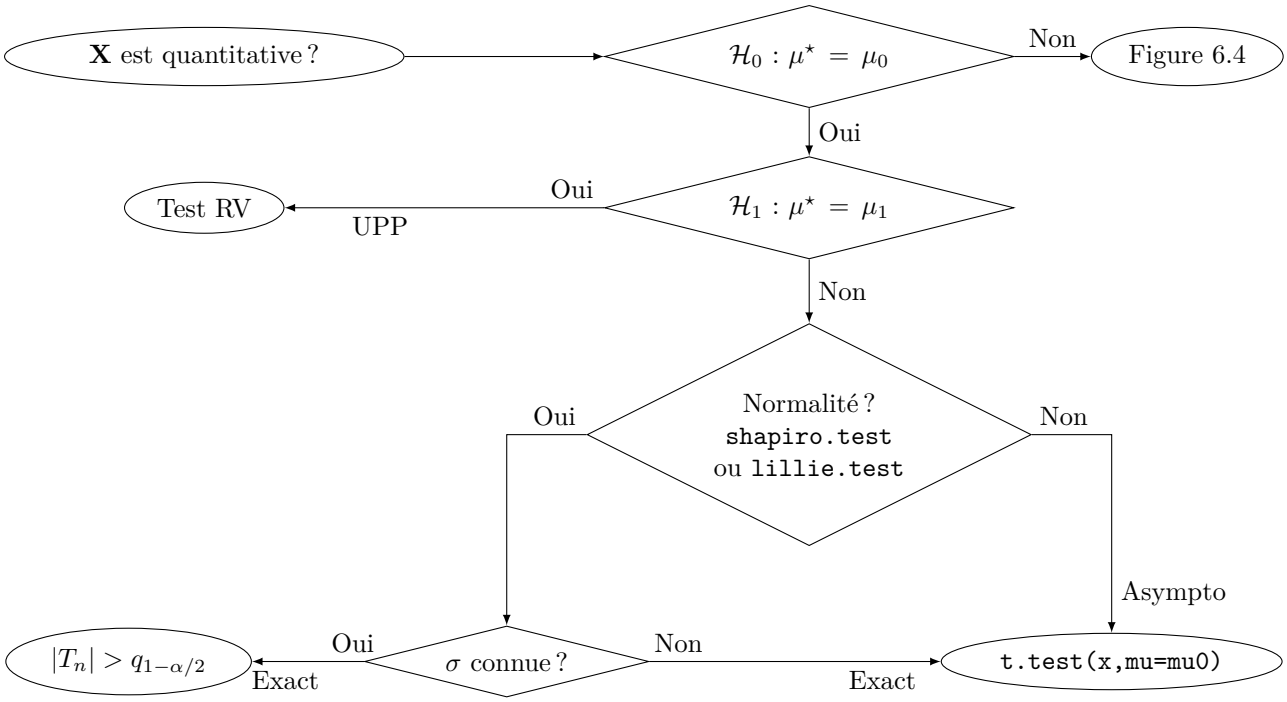



FIGURE 6.3 – Représentation de l'arbre de décision pour savoir quel test utilisé pour tester si l'espérance est égale à une valeur fixée en amont.

## 6.3 Test de comparaison des égalités d'espérances de deux échantillons

Dans cette partie, nous supposons que nous avons au moins deux échantillons $\mathbf{X} = (X_1, \ldots, X_{n_\mathbf{X}})$ et $\mathbf{Y} = (Y_1, \ldots, Y_{n_\mathbf{Y}})$ dont les lois associées possèdent une espérance $\mu^\star_\mathbf{X}$ et $\mu^\star_\mathbf{Y}$ et nous regardons si ces espérances sont égales ou pas. Dans ce cas, nous étudions donc le test paramétrique suivant :

$$\begin{cases} \mathcal{H}_0 : & \mu^\star_\mathbf{X} = \mu^\star_\mathbf{Y}, \\ \mathcal{H}_1 : & \mu^\star_\mathbf{X} \neq \mu^\star_\mathbf{Y}. \end{cases} \tag{6.2}$$

**Hypothèse (Test paramétrique et espérance)**

Pour enfoncer le clou, nous insistons sur le fait que nous ne pouvons comparer les espérances qu'à condition que les deux lois possèdent une espérance.

### 6.3.1 Cas d'échantillons appariés

Nous commençons par définir ce que sont les échantillons appariés :

**Définition 57 (Échantillons appariés)**

Nous disons que deux échantillons $\mathbf{X}$ et $\mathbf{Y}$ sont **appariés** si pour tout $i \in \{1, \ldots, n\}$, $X_i$ et $Y_i$ appartiennent au même individu. Dans ce cas, nous regarderons la loi de la paire de variables $(X_i, Y_i)_{1 \leq i \leq n}$.

Nous avons le corollaire suivant :

**Corollaire 26**

Si nous supposons que les échantillons $\mathbf{X}$ et $\mathbf{Y}$ sont appariés alors $n_\mathbf{X} = n_\mathbf{Y} = n$.

Pour ce faire, nous regardons pour tout $i \in \{1, \dots, n\}$, nous regardons la différence $D_i = X_i - Y_i$ de chaque paire et le test (6.2) devient alors le test suivant :

$$\begin{cases} \mathcal{H}_0 & : \quad \mu^\star_{\mathbf{D}} = 0, \\ \mathcal{H}_1 & : \quad \mu^\star_{\mathbf{D}} \neq 0. \end{cases}$$

Ainsi, nous pouvons reprendre les tests présentés dans la section 6.2 c'est-à-dire en adaptant la statistique $T_n$ présentée dans le corollaire 25 :

**Définition 58 (Test d'égalité d'espérance ; statistique pour échantillons appariés)**

Pour tester deux échantillons appariés $(\mathbf{X}, \mathbf{Y})$, nous utilisons la statistique de test suivante :

$$T_n = \sqrt{n} \frac{\overline{\mathbf{D}}_n}{S_n^{(\mathbf{D})}}$$

avec $\overline{\mathbf{D}}_n$ la moyenne empirique du $n$-échantillon $\mathbf{D} = (D_1, \dots, D_n)$ et ${S_n^{(\mathbf{D})}}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(D_i - \overline{\mathbf{D}}_n\right)^2$ l'estimateur non biaisée de la variance.

**Cas gaussien :** Dans le cas où les deux échantillons sont gaussiens, par le lemme 22, l'échantillon $\mathbf{D}$ est donc gaussien et nous avons la loi suivante :

**Point méthode (Test d'égalité des espérances ; cas gaussien et échantillons appariés)**

Étant donnés deux $n$-échantillons appariés $(\mathbf{X}, \mathbf{Y}) = [(X_1, Y_1), \dots, (X_n, Y_n)]$ de loi gaussienne $\mathcal{N}\left[\begin{pmatrix} \mu^\star_{\mathbf{X}} \\ \mu^\star_{\mathbf{Y}} \end{pmatrix}, \Sigma_2\right]$ avec $\Sigma_2$ la matrice de variance covariance alors la statistique de test est

$$T_n = \sqrt{n} \frac{\overline{\mathbf{D}}_n}{S_n^{(\mathbf{D})}}$$

qui a pour loi une loi de Student centrée à $n-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star_{\mathbf{X}} \neq \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} < \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} > \mu^\star_{\mathbf{Y}}$ |
|---|---|---|---|
| $\Theta_1$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x \neq y\right\}$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x < y\right\}$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x > y\right\}$ |
| Règle | $\lvert T_n \rvert > s^{(n-1)}_{1-\alpha/2}$ | $T_n < s^{(n-1)}_{\alpha}$ | $T_n > s^{(n-1)}_{1-\alpha}$ |

où $s^{(n-1)}_\alpha$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $n-1$ degrés de liberté $\mathcal{T}(n-1)$ alors $s^{(n-1)}_\alpha$ est tel que :

$$\mathbb{P}\left(S \leq s^{(n-1)}_\alpha\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rnorm(100)
Y <- rnorm(100)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,paired=T)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,Y+1,paired=T)
## Simulation d'un jeu de données avec variances différentes
X <- rnorm(100)*2
Y <- rnorm(100)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
```

```
t.test(X,Y,paired=T)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,Y+1,paired=T)
```

**Cas non gaussien :** Dans le cas où les deux échantillons ne sont pas gaussiens, nous supposons toutefois que les lois admettent un moment d'ordre 2 et nous utilisons à nouveau le théorème de la limite centrale :

**Point méthode (Test d'égalité des espérances ; cas non gaussien et échantillons appariés)**

Étant donnés deux $n$-échantillons appariés $(\mathbf{X}, \mathbf{Y}) = [(X_1, Y_1), \ldots, (X_n, Y_n)]$ de loi quelconque admettant un moment d'ordre 2 alors la statistique de test est

$$T_n = \sqrt{n}\frac{\overline{\mathbf{D}}_n}{{S_n^{(\mathbf{D})}}^2}$$

qui a pour loi **asymptotique** une loi de Student centrée à $n-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star_{\mathbf{X}} \neq \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} < \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} > \mu^\star_{\mathbf{Y}}$ |
|---|---|---|---|
| $\Theta_1$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x \neq y\right\}$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x < y\right\}$ | $\left\{(x,y) \in \mathbb{R}^2 \middle\vert x > y\right\}$ |
| Règle | $\lvert T_n\rvert > s^{(n-1)}_{1-\alpha/2}$ | $T_n < s^{(n-1)}_{\alpha}$ | $T_n > s^{(n-1)}_{1-\alpha}$ |

où $s^{(n-1)}_\alpha$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $n-1$ degrés de liberté $\mathcal{T}(n-1)$ alors $s^{(n-1)}_\alpha$ est tel que :

$$\mathbb{P}\left(S \leq s^{(n-1)}_\alpha\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rexp(100)
Y <- rexp(100)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,paired=T)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,rexp(100,1/2),paired=T)
```

### 6.3.2 Test de comparaison des égalités des variances de deux échantillons

Dans le cas de deux échantillons indépendants, nous commençons par supposer que les lois ont un moment d'ordre 2 et nous notons ${\sigma^\star_{\mathbf{X}}}^2$ et ${\sigma^\star_{\mathbf{Y}}}^2$ les variances associées à chaque échantillon. La première question à se poser est de savoir si les variances sont égales ou non. Pour ce faire, nous regardons donc le test :

$$\left\{\begin{array}{rcl} \mathcal{H}_0 & : & {\sigma^\star_{\mathbf{X}}}^2 = {\sigma^\star_{\mathbf{Y}}}^2, \\ \mathcal{H}_1 & : & {\sigma^\star_{\mathbf{X}}}^2 \neq {\sigma^\star_{\mathbf{Y}}}^2. \end{array}\right. \tag{6.3}$$

Pour ce faire, nous allons utiliser les deux estimateurs non biaisés ${S^{(\mathbf{X})}_{n_\mathbf{X}}}^2 = \frac{1}{n_\mathbf{X}-1}\sum_{i=1}^{n_\mathbf{X}}\left(X_i - \overline{\mathbf{X}}_{n_\mathbf{X}}\right)^2$ et ${S^{(\mathbf{Y})}_{n_\mathbf{Y}}}^2 = \frac{1}{n_\mathbf{Y}-1}\sum_{i=1}^{n_\mathbf{Y}}\left(Y_i - \overline{\mathbf{Y}}_{n_\mathbf{Y}}\right)^2$. Comme les deux valeurs sont (strictement) positives (sauf si toutes les observations sont égales), nous étudions le rapport des deux estimations :

$$F_n = \frac{{S^{(\mathbf{X})}_{n_\mathbf{X}}}^2}{{S^{(\mathbf{Y})}_{n_\mathbf{Y}}}^2}.$$

**Cas gaussien :**

Si les observations sont gaussiennes, nous savons par le théorème 24 de Cochran que les deux statistiques $(n_{\mathbf{X}}-1)\,{S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2$ et $(n_{\mathbf{Y}}-1)\,{S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2$ suivent des lois du $\chi^2$ à $n_{\mathbf{X}}-1$ et $n_{\mathbf{Y}}-1$ degrés de libertés respectivement et nous avons :

**Définitions 59 (Test d'égalité des variances ; statistique pour échantillons indépendants)**

Pour tester deux échantillons indépendants $\mathbf{X}=(X_1,\dots,X_{n_{\mathbf{X}}})$ et $\mathbf{Y}=(Y_1,\dots,Y_{n_{\mathbf{Y}}})$, nous utilisons la statistique de test suivante :

$$F_n=\frac{{S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2}{{S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2}.$$

avec ${S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2=\frac{1}{n_{\mathbf{X}}-1}\sum_{i=1}^{n_{\mathbf{X}}}\left(X_i-\overline{\mathbf{X}}_{n_{\mathbf{X}}}\right)^2$ et ${S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2=\frac{1}{n_{\mathbf{Y}}-1}\sum_{i=1}^{n_{\mathbf{Y}}}\left(Y_i-\overline{\mathbf{Y}}_{n_{\mathbf{Y}}}\right)^2$ des estimateurs non biaisés respectifs de la variance pour chaque échantillon.

Et nous obtenons

**Point méthode (Test d'égalité des variances ; cas gaussien)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X}=(X_1,\dots,X_{n_{\mathbf{X}}})$ et $\mathbf{Y}=(Y_1,\dots,Y_{n_{\mathbf{Y}}})$ de lois gaussiennes $\mathcal{N}\left(\mu_{\mathbf{X}}^\star,{\sigma_{\mathbf{X}}^\star}^2\right)$ et $\mathcal{N}\left(\mu_{\mathbf{Y}}^\star,{\sigma_{\mathbf{Y}}^\star}^2\right)$ alors la statistique de test est

$$F_n=\frac{{S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2}{{S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2}$$

qui a pour loi une loi de Fisher de paramètres $n_{\mathbf{X}}-1$ et $n_{\mathbf{Y}}-1$ sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha\in]0;1[$ est donc :

| $\mathcal{H}_1$ | ${\sigma_{\mathbf{X}}^\star}^2\neq{\sigma_{\mathbf{Y}}^\star}^2$ | ${\sigma_{\mathbf{X}}^\star}^2<{\sigma_{\mathbf{Y}}^\star}^2$ | ${\sigma_{\mathbf{X}}^\star}^2>{\sigma_{\mathbf{Y}}^\star}^2$ |
|---|---|---|---|
| $\Theta_1$ | $\left\{(x,y)\in\mathbb{R}_+^2 \middle\vert x\neq y\right\}$ | $\left\{(x,y)\in\mathbb{R}_+^2 \middle\vert x<y\right\}$ | $\left\{(x,y)\in\mathbb{R}_+^2 \middle\vert x>y\right\}$ |
| Règle | $F_n<f_{\alpha/2}^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ ou $F_n>f_{1-\alpha/2}^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ | $F_n<f_{\alpha}^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ | $F_n>f_{1-\alpha}^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ |

où $f_\alpha^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ est le quantile d'ordre $\alpha$ d'une loi de Fisher de paramètres $n_{\mathbf{X}}-1$ et $n_{\mathbf{Y}}-1$ ; c'est-à-dire que si $F$ suit une loi de Fisher de paramètres $n_{\mathbf{X}}-1$ et $n_{\mathbf{Y}}-1$ $\mathcal{F}(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)$ alors $f_\alpha^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}$ est tel que :

$$\mathbb{P}\left(F\leq f_\alpha^{(n_{\mathbf{X}}-1,n_{\mathbf{Y}}-1)}\right)=\alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rnorm(100)
Y <- rnorm(200)+1
## Test sous H_0 : sigma_X=sigma_Y -> acceptation avec 95% de chances
var.test(X,Y)
## Test sous H_0 : sigma_X=sigma_Y -> rejet normalement
var.test(X,Y*2)
```

**Cas non gaussien :**

Dans ce cas, nous utilisons une nouvelle fois le théorème de la limite centrale :

**Point méthode (Test d'égalité des variances ; cas non gaussien)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X} = (X_1, \dots, X_{n_\mathbf{X}})$ et $\mathbf{Y} = (Y_1, \dots, Y_{n_\mathbf{Y}})$ de lois admettant un moment d'ordre deux alors la statistique de test est

$$F_n = \frac{{S_{n_\mathbf{X}}^{(\mathbf{X})}}^2}{{S_{n_\mathbf{Y}}^{(\mathbf{Y})}}^2}$$

qui a pour loi **asymptotique** une loi de Fisher de paramètres $n_\mathbf{X} - 1$ et $n_\mathbf{Y} - 1$ sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | ${\sigma_\mathbf{X}^\star}^2 \neq {\sigma_\mathbf{Y}^\star}^2$ | ${\sigma_\mathbf{X}^\star}^2 < {\sigma_\mathbf{Y}^\star}^2$ | ${\sigma_\mathbf{X}^\star}^2 > {\sigma_\mathbf{Y}^\star}^2$ |
|---|---|---|---|
| $\Theta_1$ | $\{(x,y) \in \mathbb{R}_+^2 \mid x \neq y\}$ | $\{(x,y) \in \mathbb{R}_+^2 \mid x < y\}$ | $\{(x,y) \in \mathbb{R}_+^2 \mid x > y\}$ |
| Règle | $F_n < f_{\alpha/2}^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ ou $F_n > f_{1-\alpha/2}^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ | $F_n < f_{\alpha}^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ | $F_n > f_{1-\alpha}^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ |

où $f_\alpha^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ est le quantile d'ordre $\alpha$ d'une loi de Fisher de paramètres $n_\mathbf{X} - 1$ et $n_\mathbf{Y} - 1$ ; c'est-à-dire que si $F$ suit une loi de Fisher de paramètres $n_\mathbf{X} - 1$ et $n_\mathbf{Y} - 1$ $\mathcal{F}(n_\mathbf{X} - 1, n_\mathbf{Y} - 1)$ alors $f_\alpha^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}$ est tel que :

$$\mathbb{P}\left(F \leq f_\alpha^{(n_\mathbf{X}-1, n_\mathbf{Y}-1)}\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rpois(100,2)
Y <- rexp(200,1/sqrt(2))
## Test sous H_0 : sigma_X=sigma_Y -> acceptation avec 95% de chances
var.test(X,Y)
## Test sous H_0 : sigma_X=sigma_Y -> rejet normalement
var.test(X,rexp(200,1))
```

### 6.3.3 Cas d'échantillons indépendants

Dans le cas de deux échantillons indépendants, nous commençons par supposer que les lois ont un moment d'ordre 2 et nous notons ${\sigma_\mathbf{X}^\star}^2$ et ${\sigma_\mathbf{Y}^\star}^2$ les variances associées à chaque échantillon. La première question à se poser est de savoir si les variances sont égales ou non. Pour ce faire, nous renvoyons à la section 6.3.2.

#### Cas des variances égales

Dans le cas où les variances ${\sigma_\mathbf{X}^\star}^2$ et ${\sigma_\mathbf{Y}^\star}^2$ sont égales alors nous regarderons la statistique suivante :

**Définitions 60 (Test d'égalité d'espérance ; statistique pour échantillons indépendants)**

Pour tester deux échantillons indépendants $\mathbf{X} = (X_1, \dots, X_{n_\mathbf{X}})$ et $\mathbf{Y} = (Y_1, \dots, Y_{n_\mathbf{Y}})$, nous utilisons la statistique de test suivante :

$$Z_{n_\mathbf{X}, n_\mathbf{Y}} = \frac{\overline{\mathbf{X}}_{n_\mathbf{X}} - \overline{\mathbf{Y}}_{n_\mathbf{Y}}}{S_{n_\mathbf{X}, n_\mathbf{Y}}}$$

avec $\overline{\mathbf{X}}_{n_\mathbf{X}}$ et $\overline{\mathbf{Y}}_{n_\mathbf{Y}}$ les moyennes empiriques respectives et $S^2_{n_\mathbf{X}, n_\mathbf{Y}}$ est un estimateur non biaisé

de la variance commune

$$S_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \sqrt{\frac{(n_{\mathbf{X}}-1)\,{S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2 + (n_{\mathbf{Y}}-1)\,{S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2}{n_{\mathbf{X}}+n_{\mathbf{Y}}-2}}$$

avec ${S_{n_{\mathbf{X}}}^{(\mathbf{X})}}^2 = \frac{1}{n_{\mathbf{X}}-1}\sum_{i=1}^{n_{\mathbf{X}}}\left(X_i - \overline{\mathbf{X}}_{n_{\mathbf{X}}}\right)^2$ et ${S_{n_{\mathbf{Y}}}^{(\mathbf{Y})}}^2 = \frac{1}{n_{\mathbf{Y}}-1}\sum_{i=1}^{n_{\mathbf{Y}}}\left(Y_i - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}\right)^2$ des estimateurs non biaisés respectifs de la variance pour chaque échantillon.

**Cas gaussien :** Dans le cas où les deux échantillons sont gaussiens, nous avons la loi suivante :

**Point méthode (Test d'égalité des espérances ; cas gaussien et échantillons indépendants)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X} = (X_1, \ldots, X_{n_{\mathbf{X}}})$ et $\mathbf{Y} = (Y_1, \ldots, Y_{n_{\mathbf{Y}}})$ de lois gaussiennes $\mathcal{N}\left(\mu_{\mathbf{X}}^\star, {\sigma_{\mathbf{X}}^\star}^2\right)$ et $\mathcal{N}\left(\mu_{\mathbf{Y}}^\star, {\sigma_{\mathbf{Y}}^\star}^2\right)$ avec ${\sigma_{\mathbf{X}}^\star}^2 = {\sigma_{\mathbf{Y}}^\star}^2$ alors la statistique de test est

$$Z_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \frac{\overline{\mathbf{X}}_{n_{\mathbf{X}}} - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}}{S_{n_{\mathbf{X}},n_{\mathbf{Y}}}}$$

qui a pour loi une loi de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu_{\mathbf{X}}^\star \neq \mu_{\mathbf{Y}}^\star$ | $\mu_{\mathbf{X}}^\star < \mu_{\mathbf{Y}}^\star$ | $\mu_{\mathbf{X}}^\star > \mu_{\mathbf{Y}}^\star$ |
|---|---|---|---|
| $\Theta_1$ | $\left\{(x,y)\in\mathbb{R}^2 \middle\vert x \neq y\right\}$ | $\left\{(x,y)\in\mathbb{R}^2 \middle\vert x < y\right\}$ | $\left\{(x,y)\in\mathbb{R}^2 \middle\vert x > y\right\}$ |
| Règle | $\lvert Z_n\rvert > s_{1-\alpha/2}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ | $Z_n < s_{\alpha}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ | $Z_n > s_{1-\alpha}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ |

où $s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté $\mathcal{T}(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)$ alors $s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rnorm(100)
Y <- rnorm(200)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,var.equal=T)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,Y+1,var.equal=T)
```

**Cas non gaussien :** Dans le cas où les deux échantillons ne sont pas gaussiens, nous supposons toutefois que les lois admettent un moment d'ordre 2 et nous utilisons à nouveau le théorème de la limite centrale :

**Point méthode (Test d'égalité des espérances ; cas non gaussien et échantillons indépendants)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X} = (X_1, \ldots, X_{n_{\mathbf{X}}})$ et $\mathbf{Y} = (Y_1, \ldots, Y_{n_{\mathbf{Y}}})$ de loi quelconque admettant un moment d'ordre 2 alors la statistique de test est

$$Z_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \frac{\overline{\mathbf{X}}_{n_{\mathbf{X}}} - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}}{S_{n_{\mathbf{X}},n_{\mathbf{Y}}}}$$

qui a pour loi une loi **asymptotique** de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star_{\mathbf{X}} \neq \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} < \mu^\star_{\mathbf{Y}}$ | $\mu^\star_{\mathbf{X}} > \mu^\star_{\mathbf{Y}}$ |
|---|---|---|---|
| $\Theta_1$ | $\{(x,y)\in\mathbb{R}^2 \mid x \neq y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x < y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x > y\}$ |
| Règle | $\lvert Z_n\rvert > s_{1-\alpha/2}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ | $Z_n < s_{\alpha}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ | $Z_n > s_{1-\alpha}^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ |

où $s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $n_{\mathbf{X}}+n_{\mathbf{Y}}-2$ degrés de liberté $\mathcal{T}(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)$ alors $s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rchisq(100,25)
Y <- rchisq(200,25)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,var.equal=T)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,Y+5,var.equal=T)
```

#### Cas des variances inégales

Dans le cas où les variances ${\sigma^\star_{\mathbf{X}}}^2$ et ${\sigma^\star_{\mathbf{Y}}}^2$ sont différentes alors nous regarderons la statistique suivante :

**Définitions 61 (Test d'égalité d'espérance ; statistique pour échantillons indépendants)**

Pour tester deux échantillons indépendants $\mathbf{X}=(X_1,\dots,X_{n_{\mathbf{X}}})$ et $\mathbf{Y}=(Y_1,\dots,Y_{n_{\mathbf{Y}}})$, nous utilisons la statistique de test suivante :

$$Z_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \frac{\overline{\mathbf{X}}_{n_{\mathbf{X}}} - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}}{\sqrt{\frac{S^{(\mathbf{X})}_{n_{\mathbf{X}}}}{n_{\mathbf{X}}} + \frac{S^{(\mathbf{Y})}_{n_{\mathbf{Y}}}}{n_{\mathbf{Y}}}}}$$

avec $\overline{\mathbf{X}}_{n_{\mathbf{X}}}$ et $\overline{\mathbf{Y}}_{n_{\mathbf{Y}}}$ les moyennes empiriques respectives et ${S^{(\mathbf{X})}_{n_{\mathbf{X}}}}^2 = \frac{1}{n_{\mathbf{X}}-1}\sum_{i=1}^{n_{\mathbf{X}}}\left(X_i - \overline{\mathbf{X}}_{n_{\mathbf{X}}}\right)^2$ et ${S^{(\mathbf{Y})}_{n_{\mathbf{Y}}}}^2 = \frac{1}{n_{\mathbf{Y}}-1}\sum_{i=1}^{n_{\mathbf{Y}}}\left(Y_i - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}\right)^2$ des estimateurs non biaisés respectifs de la variance pour chaque échantillon.

**Cas gaussien :** Dans le cas où les deux échantillons sont gaussiens, nous avons la loi suivante :

**Point méthode (Test de Welch d'égalité des espérances)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X}=(X_1,\dots,X_{n_{\mathbf{X}}})$ et $\mathbf{Y}=(Y_1,\dots,Y_{n_{\mathbf{Y}}})$ de lois gaussiennes $\mathcal{N}\left(\mu^\star_{\mathbf{X}},{\sigma^\star_{\mathbf{X}}}^2\right)$ et $\mathcal{N}\left(\mu^\star_{\mathbf{Y}},{\sigma^\star_{\mathbf{Y}}}^2\right)$ avec ${\sigma^\star_{\mathbf{X}}}^2 \neq {\sigma^\star_{\mathbf{Y}}}^2$ alors la statistique de test est

$$Z_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \frac{\overline{\mathbf{X}}_{n_{\mathbf{X}}} - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}}{\sqrt{\frac{S^{(\mathbf{X})}_{n_{\mathbf{X}}}}{n_{\mathbf{X}}} + \frac{S^{(\mathbf{Y})}_{n_{\mathbf{Y}}}}{n_{\mathbf{Y}}}}}$$

qui a pour loi une loi de Student centrée à $\nu$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ avec $\nu$ la

valeur entière la plus proche de :

$$\frac{\left[\frac{S^{(\mathbf{X})}_{n_\mathbf{X}}}{n_\mathbf{X}}+\frac{S^{(\mathbf{Y})}_{n_\mathbf{Y}}}{n_\mathbf{Y}}\right]^2}{\frac{S^{(\mathbf{X})\,2}_{n_\mathbf{X}}}{(n_\mathbf{X}-1)n_\mathbf{X}^2}+\frac{S^{(\mathbf{Y})\,2}_{n_\mathbf{Y}}}{(n_\mathbf{Y}-1)n_\mathbf{Y}^2}}$$

et la règle de rejet de niveau approché $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star_\mathbf{X} \neq \mu^\star_\mathbf{Y}$ | $\mu^\star_\mathbf{X} < \mu^\star_\mathbf{Y}$ | $\mu^\star_\mathbf{X} > \mu^\star_\mathbf{Y}$ |
|---|---|---|---|
| $\Theta_1$ | $\{(x,y)\in\mathbb{R}^2 \mid x \neq y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x < y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x > y\}$ |
| Règle | $\lvert Z_n\rvert > s^{(\nu)}_{1-\alpha/2}$ | $Z_n < s^{(\nu)}_{\alpha}$ | $Z_n > s^{(\nu)}_{1-\alpha}$ |

où $s^{(\nu)}_\alpha$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $\nu$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $\nu$ degrés de liberté $\mathcal{T}(\nu)$ alors $s^{(n_\mathbf{X}+n_\mathbf{Y}-2)}_\alpha$ est tel que :

$$\mathbb{P}\left(S \leq s^{(\nu)}_\alpha\right) = \alpha.$$

Ce test est connu sous le nom de **test de Welch**.

### Codage en R

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rnorm(100)
Y <- rnorm(200)*2
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,var.equal=F)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,Y+1,var.equal=F)
```

**Cas non gaussien :** Dans le cas où les deux échantillons ne sont pas gaussiens, nous supposons toutefois que les lois admettent un moment d'ordre 2 et nous utilisons à nouveau le théorème de la limite centrale :

**Point méthode (Test d'égalité des espérances ; cas non gaussien et échantillons indépendants)**

Étant donnés deux $n$-échantillons indépendants $\mathbf{X} = (X_1,\dots,X_{n_\mathbf{X}})$ et $\mathbf{Y} = (Y_1,\dots,Y_{n_\mathbf{Y}})$ de loi quelconque admettant des moments d'ordre 2 différents alors la statistique de test est

$$Z_{n_\mathbf{X},n_\mathbf{Y}} = \frac{\overline{\mathbf{X}}_{n_\mathbf{X}} - \overline{\mathbf{Y}}_{n_\mathbf{Y}}}{\sqrt{\frac{S^{(\mathbf{X})}_{n_\mathbf{X}}}{n_\mathbf{X}}+\frac{S^{(\mathbf{Y})}_{n_\mathbf{Y}}}{n_\mathbf{Y}}}}$$

qui a pour loi une loi **asymptotique** de Student centrée à $\nu$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ avec $\nu$ la valeur entière la plus proche de :

$$\frac{\left[\frac{S^{(\mathbf{X})}_{n_\mathbf{X}}}{n_\mathbf{X}}+\frac{S^{(\mathbf{Y})}_{n_\mathbf{Y}}}{n_\mathbf{Y}}\right]^2}{\frac{S^{(\mathbf{X})\,2}_{n_\mathbf{X}}}{(n_\mathbf{X}-1)n_\mathbf{X}^2}+\frac{S^{(\mathbf{Y})\,2}_{n_\mathbf{Y}}}{(n_\mathbf{Y}-1)n_\mathbf{Y}^2}}$$

et la règle de rejet de niveau **asymptotique** approché $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\mu^\star_\mathbf{X} \neq \mu^\star_\mathbf{Y}$ | $\mu^\star_\mathbf{X} < \mu^\star_\mathbf{Y}$ | $\mu^\star_\mathbf{X} > \mu^\star_\mathbf{Y}$ |
|---|---|---|---|
| $\Theta_1$ | $\{(x,y)\in\mathbb{R}^2 \mid x \neq y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x < y\}$ | $\{(x,y)\in\mathbb{R}^2 \mid x > y\}$ |
| Règle | $\lvert Z_n\rvert > s^{(\nu)}_{1-\alpha/2}$ | $Z_n < s^{(\nu)}_{\alpha}$ | $Z_n > s^{(\nu)}_{1-\alpha}$ |

où $s_\alpha^{(\nu)}$ est le quantile d'ordre $\alpha$ d'une loi de Student centrée à $\nu$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi de Student centrée à $\nu$ degrés de liberté $\mathcal{T}(\nu)$ alors $s_\alpha^{(n_{\mathbf{X}}+n_{\mathbf{Y}}-2)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(\nu)}\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `t.test`. Par exemple :

```
## Simulation d'un jeu de données
X <- rexp(100)
Y <- rchisq(200,1)
## Test sous H_0 : mu=0 -> acceptation avec 95% de chances
t.test(X,Y,var.equal=F)
## Test sous H_0 : mu=1 -> rejet normalement
t.test(X,rexp(200,1/2),var.equal=F)
```

### 6.3.4 Quel(s) test(s) utilisé(s) ?

Pour faire le bilan de cette partie, nous avons créé un arbre de décision pour savoir quel(s) test(s) utilisé(s) disponible sur la figure 6.4.

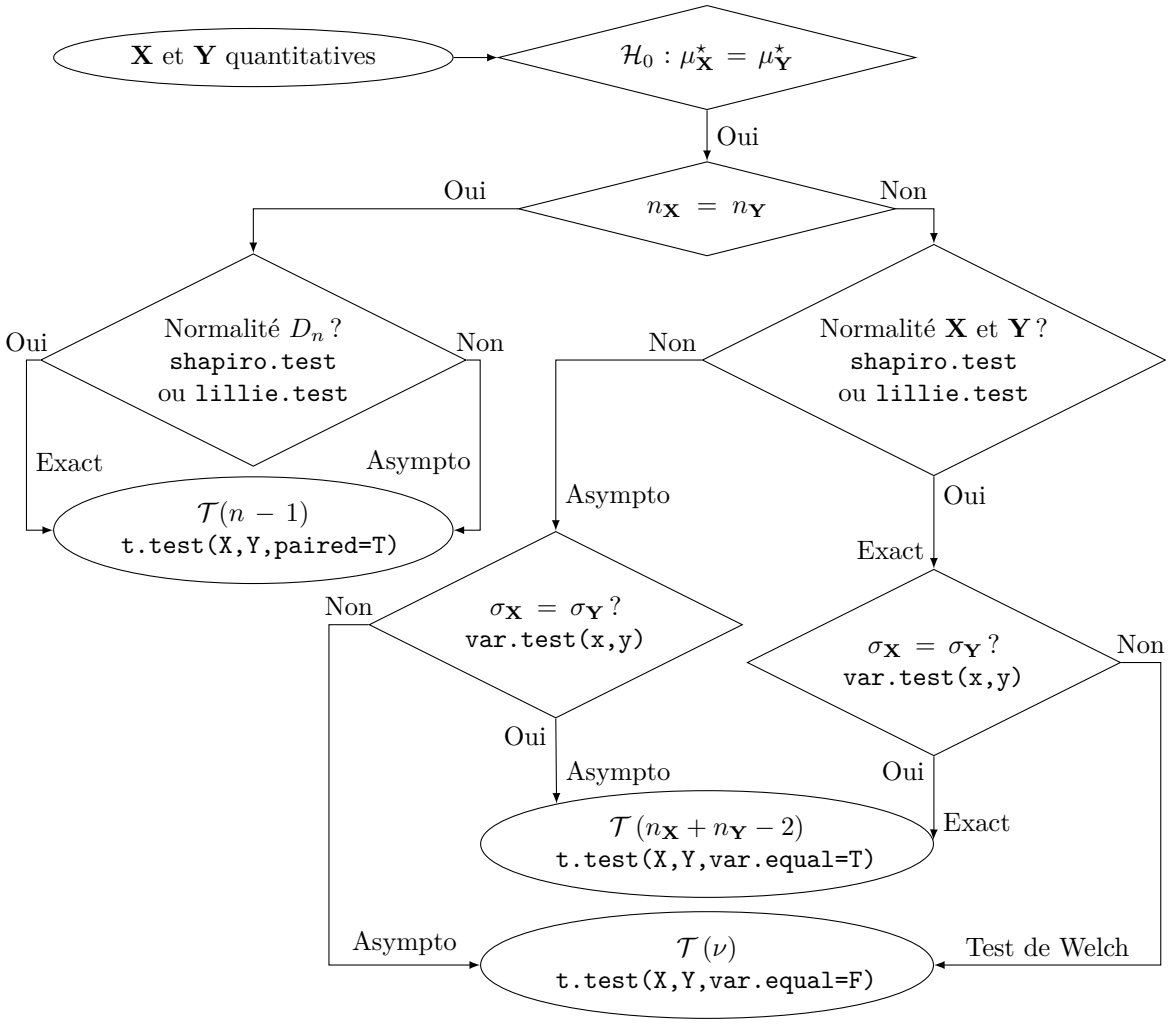



FIGURE 6.4 – Représentation de l'arbre de décision pour savoir quel test utilisé dans le cas de deux échantillons où nous cherchons à savoir si leurs espérances sont égales.

### 6.3.5 Généralisation à plus de deux échantillons dans le cas gaussien

Ces tests peuvent être généralisés à plus de deux échantillons. Nous mettons ici les grandes idées.

#### Comparaison de variances, échantillons indépendants

Nous supposons posséder $K$ échantillons $(X_{1,1}, \ldots, X_{n_1,1})$, $(X_{1,2}, \ldots, X_{n_2,2})$ et ainsi de suite jusqu'à $(X_{1,K}, \ldots, X_{n_K,K})$ admettant des moments d'ordre deux. Nous testons donc les variances $\sigma^\star_1, \ldots, \sigma^\star_K$ :

$$\begin{cases} \mathcal{H}_0 & : \quad \sigma^\star_1 = \sigma^\star_2 = \cdots = \sigma^\star_K, \\ \mathcal{H}_1 & : \quad \exists k \neq k', \sigma^\star_k \neq \sigma^\star_{k'}. \end{cases}$$

On introduit les estimateurs non biaisés de variance et nous avons :

**Définitions 62 (Test d'égalité des variances ; statistique pour plusieurs échantillons indépendants**

Pour tester plusieurs échantillons indépendants $(X_{1,1}, \ldots, X_{n_1,1})$, $(X_{1,2}, \ldots, X_{n_2,2})$ et ainsi de suite jusqu'à $(X_{1,K}, \ldots, X_{n_K,K})$ admettant des moments d'ordre deux, nous utilisons la statistique de test suivante :

$$F_n = \frac{(n-k)\ln\left[\frac{1}{n-K}\sum_{k=1}^K (n_k-1)\,S_k^2\right] - \sum_{k=1}^K (n_k-1)\ln\left[S_k^2\right]}{1 + \frac{1}{3(K-1)}\left[\sum_{k=1}^K \left(\frac{1}{n_k-1}\right) - \frac{1}{n-K}\right]}$$

où $n = \sum_{k=1}^{K} n_k$ et pour tout $k \in \{1, \ldots, K\}$, $S_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(X_{i,k} - \overline{X}_{\cdot,k}\right)^2$ avec $\overline{X}_{\cdot,k}$ est la moyenne de l'échantillon $k$.

Et nous obtenons

**Point méthode (Test de Barlett)**

Étant donnés plusieurs échantillons indépendants $(X_{1,1}, \ldots, X_{n_1,1})$, $(X_{1,2}, \ldots, X_{n_2,2})$ et ainsi de suite jusqu'à $(X_{1,K}, \ldots, X_{n_K,K})$ admettant des moments d'ordre deux alors la statistique de test est

$$F_n = \frac{(n-k)\ln\left[\frac{1}{n-K}\sum_{k=1}^{K}(n_k-1)S_k^2\right] - \sum_{k=1}^{K}(n_k-1)\ln\left[S_k^2\right]}{1+\frac{1}{3(K-1)}\left[\sum_{k=1}^{K}\left(\frac{1}{n_k-1}\right)-\frac{1}{n-K}\right]}$$

qui a pour loi **asymptotique** une loi du $\chi^2(K-1)$ à $K-1$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| | |
|---|---|
| $\mathcal{H}_1$ | $\exists k \neq k',\ \sigma_k^\star \neq \sigma_{k'}^\star$ |
| $\Theta_1$ | $\mathbb{R}_+^K \setminus \left\{(x_1, x_2, \ldots, x_K) \in \mathbb{R}_+^K \middle\vert x_1 = x_2 = \cdots = x_K\right\}$ |
| Règle | $F_n > s_{1-\alpha}^{(K-1)}$ |

où $s^{(K-1)}$ est le quantile d'ordre $\alpha$ d'une loi du $\chi^2(K-1)$ à $K-1$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi du $\chi^2(K-1)$ à $K-1$ degrés de liberté alors $s_\alpha^{(K-1)}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{(K-1)}\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `bartlett.test`. Par exemple :

```
## Simulation d'un jeu de données
X<-data.frame(X=c(rnorm(100),rnorm(200)+1,rnorm(300)-1),
              groupe=factor(c(rep(1,100),rep(2,200),rep(3,300))))

## Test sous H_0 : sigma_X=sigma_Y -> acceptation avec 95% de chances
bartlett.test(X$X,X$groupe)
## Test sous H_0 : sigma_X=sigma_Y -> rejet normalement
X<-data.frame(X=c(rnorm(100),2*rnorm(200)+1,rnorm(300)-1),
              groupe=factor(c(rep(1,100),rep(2,200),rep(3,300))))
bartlett.test(X$X,X$groupe)
```

### Comparaison d'espérances, échantillons appariés

Si les échantillons sont appariés alors nous avons une matrice $\mathbf{X} = (X_{i,k})$ où chaque colonne $X_{\cdot,k}$ possède sa propre loi d'espérance $\mu_k^\star$. Nous testons donc :

$$\begin{cases} \mathcal{H}_0 & : \quad \mu_1^\star = \mu_2^\star = \cdots = \mu_K^\star, \\ \mathcal{H}_1 & : \quad \exists k \neq k',\ \mu_k^\star \neq \mu_{k'}^\star. \end{cases}$$

Ce test revient à tester l'effet du facteur expérience dans un modèle d'ANOVA 2 sans interaction (voir le cours de modélisation linéaire). Nous avons :

**Définitions 63 (Test d'égalité des espérances ; statistique pour plusieurs échantillons appariés)**

Pour tester plusieurs échantillons appariés $\mathbf{X} = (X_{i,k})$, nous utilisons la statistique de test suivante :

$$F_n = \frac{n\sum_{k=1}^{K}\left(\overline{X}_{\cdot,k} - \overline{X}_{\cdot,\cdot}\right)^2}{\frac{1}{n-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{i,k} - \overline{X}_{\cdot,k} - \overline{X}_{i,\cdot} + \overline{X}_{\cdot,\cdot}\right)^2}$$

où $\overline{X}_{\cdot,k}$ et $\overline{X}_{i,\cdot}$ sont respectivement les moyennes par colonne et par ligne et $\overline{X}_{\cdot,\cdot}$ est la moyenne globale.

Et nous obtenons

**Point méthode (Test d'égalité des espérances ; cas gaussien)**

Étant donnés plusieurs échantillons appariés $\mathbf{X} = (X_{i,k})$ de lois gaussiennes alors la statistique de test est

$$F_n = \frac{n\sum_{k=1}^{K}\left(\overline{X}_{\cdot,k} - \overline{X}_{\cdot,\cdot}\right)^2}{\frac{1}{n-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{i,k} - \overline{X}_{\cdot,k} - \overline{X}_{i,\cdot} + \overline{X}_{\cdot,\cdot}\right)^2}$$

qui a pour loi une loi de Fisher de paramètres $K-1$ et $(n-1)(K-1)$ sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| | |
|---|---|
| $\mathcal{H}_1$ | $\exists k \neq k',\ \mu_k^\star \neq \mu_{k'}^\star$ |
| $\Theta_1$ | $\mathbb{R}^K \setminus \left\{(x_1, x_2, \ldots, x_K) \in \mathbb{R}^K \middle\vert x_1 = x_2 = \cdots = x_K\right\}$ |
| Règle | $F_n < f_{\alpha/2}^{(K-1,(n-1)(K-1))}$ ou $F_n > f_{1-\alpha/2}^{(K-1,(n-1)(K-1))}$ |

où $f_\alpha^{(K-1,(n-1)(K-1))}$ est le quantile d'ordre $\alpha$ d'une loi de Fisher de paramètres $K-1$ et $(n-1)(K-1)$ ; c'est-à-dire que si $F$ suit une loi de Fisher de paramètres $n_{\mathbf{X}}-1$ et $n_{\mathbf{Y}}-1$ $\mathcal{F}(n_{\mathbf{X}}-1, n_{\mathbf{Y}}-1)$ alors $f_\alpha^{(K-1,(n-1)(K-1))}$ est tel que :

$$\mathbb{P}\left(F \leq f_\alpha^{(K-1,(n-1)(K-1))}\right) = \alpha.$$

### Codage en R

Sous condition que les espérances soient égales (voir le test de Barett), la fonction R permettant de faire le test bilatéral est `aov`. Par exemple :

```
## Simulation d'un jeu de données
X<-data.frame(X=c(rnorm(100),rnorm(200),rnorm(300)),
              groupe=factor(c(rep(1,100),rep(2,200),rep(3,300))))
## Test sous H_0 : sigma_X=sigma_Y -> acceptation avec 95% de chances
summary(aov(X$X~X$groupe))
## Test sous H_0 : sigma_X=sigma_Y -> rejet normalement
X<-data.frame(X=c(rnorm(100),rnorm(200)+1,rnorm(300)-1),
              groupe=factor(c(rep(1,100),rep(2,200),rep(3,300))))
summary(aov(X$X~X$groupe))
```

## 6.4 Test de corrélation

Enfin, le dernier test des variables quantitatives est celui des corrélations. Dans ce cadre, nous supposons avoir des échantillons appariés $(X_i, Y_i)_{1\leq i\leq n}$ qui proviennent d'une loi binormale d'espérance $(\mu_{\mathbf{X}}^\star, \mu_{\mathbf{Y}}^\star)$ et de matrice de variance covariance :

$$\begin{pmatrix} {\sigma_{\mathbf{X}}^\star}^2 & \sigma_{\mathbf{X}}^\star\sigma_{\mathbf{Y}}^\star\rho \\ \sigma_{\mathbf{X}}^\star\sigma_{\mathbf{Y}}^\star\rho & {\sigma_{\mathbf{Y}}^\star}^2 \end{pmatrix}$$

avec $\rho \in \mathbb{R}$ appelé **coefficient de corrélation** (voir la section 1.5.2). L'estimateur naturel est le coefficient de corrélation empirique :

**Définition 64 (Coefficient de corrélation empirique)**

Le coefficient de corrélation empirique $\widehat{\rho}_n$ est défini par :

$$\widehat{\rho}_n = \frac{\mathrm{Cov}_n(\mathbf{X},\mathbf{Y})}{\widehat{\sigma}_n^{(\mathbf{X})}\widehat{\sigma}_n^{(\mathbf{Y})}} = \frac{\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)\left(Y_i - \overline{\mathbf{Y}}_{n_{\mathbf{Y}}}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \overline{\mathbf{Y}}_n\right)^2}}.$$

Ainsi, la question que nous nous posons est de savoir si $\rho = 0$ ou pas :

$$\begin{cases} \mathcal{H}_0 & : \quad \rho = 0, \\ \mathcal{H}_1 & : \quad \rho \neq 0. \end{cases}$$

Nous avons :

**Définitions 65 (Test de corrélation)**

Pour tester la corrélation de deux échantillons appariés $(X_i, Y_i)_{1\leq i\leq n}$, nous utilisons la statistique de test suivante :

$$T_n = \sqrt{n-2} \times \frac{\widehat{\rho}_n}{\sqrt{1-\widehat{\rho}_n^2}}.$$

Et nous obtenons

**Point méthode (Test de corrélation)**

Étant donnés deux échantillons appariés $(X_i, Y_i)_{1\leq i\leq n}$ de loi binormale alors la statistique de test est

$$T_n = \sqrt{n-2} \times \frac{\widehat{\rho}_n}{\sqrt{1-\widehat{\rho}_n^2}}$$

qui a pour loi une loi de Student $\mathcal{T}(n-2)$ à $n-2$ degrés de libertés sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\rho \neq 0$ |
|---|---|
| $\Theta_1$ | $\mathbb{R}^\star$ |
| Règle | $\lvert T_n\rvert > t_{1-\alpha/2}^{(n-2)}$ |

où $t_\alpha^{(n-2)}$ est le quantile d'ordre $\alpha$ d'une loi de Student $\mathcal{T}(n-2)$ à $n-2$ degrés de libertés ; c'est-à-dire que si $T$ suit une loi de Student $\mathcal{T}(n-2)$ à $n-2$ degrés de libertés alors $t_\alpha^{(n-2)}$ est tel que :

$$\mathbb{P}\left(T \leq t_\alpha^{(n-2)}\right) = \alpha.$$

### Codage en R

La fonction R permettant de faire le test bilatéral est `cor.test`. Par exemple :

```
## Simulation d'un jeu de données
X<-rnorm(100)
Y<-rnorm(100)
## Test sous H_0 : corrélation -> acceptation avec 95% de chances
cor.test(X,Y,method="pearson")
## Test sous H_0 : corrélation -> rejet normalement
cor.test(X,X+0.5*Y,method="pearson")
```

## 6.5 Solutions des exercices

**Exercices 6.1**

Pour la puissance, nous supposons que $\mu^\star$ est différent de $\mu_0$ donc, toujours d'après la proposition 21 $\overline{\mathbf{X}}_n$ est de loi gaussienne d'espérance $\mu^\star \neq \mu_0$ et de variance $\sigma^2/n$. Ainsi, nous avons que $T_n$ suit une loi gaussienne dont il faut calculer l'espérance et la variance :

$$\mathbb{E}[T_n] = \mathbb{E}\left[\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}\right] = \sqrt{n}\frac{1}{\sigma}\left[\mathbb{E}\left[\overline{\mathbf{X}}_n\right] - \mu_0\right] = \sqrt{n}\frac{1}{\sigma}\left[\mu^\star - \mu_0\right].$$

Et par la même démonstration que pour le point méthode, la variance est égale à 1. Ainsi, sous l'hypothèse que $\mu^\star \neq \mu_0$, nous avons :

$$T_n \sim \mathcal{N}\left(\frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right], 1\right)$$

donc nous avons par exemple dans le cas où $\mathcal{H}_1$ est $\mu^\star \neq \mu_0$ :

$$\begin{aligned}
\underline{\beta}(\mu^\star) &= \mathbb{P}\left(|T_n| \leq q_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(-q_{1-\alpha/2} < T_n \leq q_{1-\alpha/2}\right) \\
&= 1 - \mathbb{P}\left(T_n \leq -q_{1-\alpha/2}\right) + \mathbb{P}\left(T_n \leq q_{1-\alpha/2}\right) - \left[\mathbb{P}\left(T_n \leq q_{1-\alpha/2}\right) - \mathbb{P}\left(T_n \leq -q_{1-\alpha/2}\right)\right] \\
&= \mathbb{P}\left(T_n \leq q_{1-\alpha/2}\right) - \mathbb{P}\left(T_n < -q_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(T_n \leq q_{1-\alpha/2}\right) - \mathbb{P}\left(T_n \leq -q_{1-\alpha/2}\right) \text{ car } T_n \text{ est continue,} \\
&= \mathbb{P}\left(T_n - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \leq q_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) \\
&\quad -\mathbb{P}\left(T_n - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \leq -q_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) \\
&= F_{\mathcal{N}(0,1)}\left(q_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) - F_{\mathcal{N}(0,1)}\left(-q_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right)
\end{aligned}$$

où $F_{\mathcal{N}(0,1)}$ est la fonction de répartition d'une loi gaussienne centrée réduite. En particulier, nous voyons que plus $\mu^\star$ est loin de $\mu_0$ plus l'erreur de type deux va tendre vers 0.

Par la même démonstration, dans le cas où $\mu^\star > \mu_0$, nous montrons que le risque de seconde espèce $\underline{\beta}$ vaut :

$$\begin{aligned}
\underline{\beta}(\mu^\star) &= \mathbb{P}\left(\Phi(X) = 0\right) \\
&= \mathbb{P}\left(T_n \leq q_{1-\alpha}\right) \\
&= \mathbb{P}\left(T_n - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \leq q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) \\
&= F_{\mathcal{N}(0,1)}\left(q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right)
\end{aligned}$$

et nous voyons que pour tout $\mu^\star > \mu_0$, nous avons :

$$\begin{aligned}
\mu^\star > \mu_0 &\Leftrightarrow \mu^\star - \mu_0 > 0 \\
&\Leftrightarrow \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] > 0 \\
&\Leftrightarrow -\frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] < 0 \\
&\Leftrightarrow q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] < q_{1-\alpha} \\
&\Leftrightarrow F_{\mathcal{N}(0,1)}\left(q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) < F_{\mathcal{N}(0,1)}\left(q_{1-\alpha}\right) \text{ car } F_{\mathcal{N}(0,1)} \text{ est croissante,} \\
&\Leftrightarrow \underline{\beta}(\mu^\star) < 1 - \alpha
\end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \quad -\underline{\beta}\left(\mu^\star\right) > \alpha - 1 \\ &\Leftrightarrow \quad 1 - \underline{\beta}\left(\mu^\star\right) > \alpha \end{aligned}$$

et la suite de tests est bien sans biais. De plus, nous avons pour tout $\mu^\star > \mu_0$ :

$$\begin{aligned} \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \underset{n\to+\infty}{\longrightarrow} +\infty \quad &\Rightarrow \quad -\frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \underset{n\to+\infty}{\longrightarrow} -\infty \\ &\Rightarrow \quad q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right] \underset{n\to+\infty}{\longrightarrow} -\infty \\ &\Rightarrow \quad F_{\mathcal{N}(0,1)}\left(q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}\left[\mu^\star - \mu_0\right]\right) \underset{n\to+\infty}{\longrightarrow} 0 \end{aligned}$$

et la suite de tests est bien consistante (puisque sans biais et avec un risque de seconde espèce qui tend vers 0 lorsque $n$ tend vers l'infini).

**Exercices 6.2**

Nous commençons par remarquer que pour tout $i \in \{1,\dots,n\}$, $X_i - \overline{\mathbf{X}}_n$ est une loi gaussienne par le lemme 22 dont nous devons calculer la moyenne et la variance :

$$\mathbb{E}\left[X_i - \overline{\mathbf{X}}_n\right] \quad = \quad \mathbb{E}\left[X_i\right] - \mathbb{E}\left[\overline{\mathbf{X}}_n\right] = \mu^\star - \mu^\star = 0.$$

Et pour la variance, il faut faire attention au fait que $\overline{\mathbf{X}}_n$ n'est pas indépendant de $X_i$ puisque la somme inclut cette variable. Il faut donc la traiter à part dans le calcul :

$$\begin{aligned} \mathbb{V}\left[X_i - \overline{\mathbf{X}}_n\right] &= \mathbb{V}\left[X_i\right] + \mathbb{V}\left[\overline{\mathbf{X}}_n\right] - 2\mathrm{Cov}\left(X_i, \overline{\mathbf{X}}_n\right) \\ &= \sigma^2 + \frac{\sigma^2}{n} - 2\mathrm{Cov}\left(X_i, \frac{1}{n}\sum_{i'=1}^{n} X_{i'}\right) \\ &= \frac{(n+1)\sigma^2}{n} - \frac{2}{n}\sum_{i'=1}^{n}\mathrm{Cov}\left(X_i, X_{i'}\right) \\ &= \frac{(n+1)\sigma^2}{n} - \frac{2}{n}\left[\sum_{\substack{i'=1 \\ i'\neq i}}^{n}\underbrace{\mathrm{Cov}\left(X_i, X_{i'}\right)}_{=0} + \mathrm{Cov}\left(X_i, X_i\right)\right] \\ &= \frac{(n+1)\sigma^2}{n} - \frac{2}{n}\mathbb{V}\left[X_i\right] \\ &= \frac{(n-1)\sigma^2}{n}. \end{aligned}$$

Or, comme les variables $Z_i = X_i - \overline{\mathbf{X}}_n$ sont centrées alors pour tout $i \in \{1,\dots,n\}$, nous avons :

$$\mathbb{E}\left[\left(X_i - \overline{\mathbf{X}}_n\right)^2\right] = \mathbb{E}\left[Z_i^2\right] = \mathbb{V}\left[Z_i\right] = \frac{(n-1)\sigma^2}{n}.$$

Pour vérifier le biais, nous devons calculer l'espérance de $S_n^2$ :

$$\begin{aligned} \mathbb{E}\left[\widehat{\sigma}_n^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{\mathbf{X}}_n\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \overline{\mathbf{X}}_n\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Z_i^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(n-1)\sigma^2}{n} \\
&= \frac{n}{n}\frac{(n-1)\sigma^2}{n} \\
&= \frac{(n-1)\sigma^2}{n}
\end{aligned}
$$

et l'estimateur est biaisé alors que nous avons :

$$\mathbb{E}\left[S_n^2\right] = \mathbb{E}\left[\frac{n}{n-1}\widehat{\sigma}_n^2\right] = \frac{n}{n-1}\mathbb{E}\left[\widehat{\sigma}_n^2\right] = \frac{n}{n-1}\times\frac{(n-1)\sigma^2}{n} = \sigma^2$$

qui est non biaisé.

# Chapitre 7

# Tests sur les variables qualitatives

## 7.1 Objectifs

Dans ce chapitre, nous nous concentrons sur les variables qualitatives. Dans ce cadre, nous ne pouvons regarder que les proportions de chaque modalité.

## 7.2 Test d'une probabilité pour une variable de Bernoulli

La variable de Bernoulli permet de caractériser la présence ou l'absence d'une particularité dans la population. Par exemple, nous pouvons nous intéresser à la présence du caractère $A$ par opposition à un autre caractère. Une utilisation documentée de cette question est celle de Pierre-Simon de Laplace sur la proportion de naissances masculines à Paris (présentée en section 2.2). Dans ce cadre, nous supposons avoir un $n$-échantillon de loi de Bernoulli $\mathcal{B}(p^\star)$ et nous nous demandons si elle est égale à une certaine valeur $p_0 \in ]0;1[$ :

$$\begin{cases} \mathcal{H}_0 & : \quad p^\star = p_0, \\ \mathcal{H}_1 & : \quad p^\star \neq p_0. \end{cases}$$

L'estimateur usuel de $p^\star$, que ce soit celui des moments ou du maximum de vraisemblance, est $\widehat{p}_n = \overline{\mathbf{X}}_n$. Deux solutions s'offrent alors à nous pour résoudre ce problème.

### 7.2.1 Méthode exacte

La première solution est d'utiliser la loi de $n\overline{\mathbf{X}}_n$ qui est une loi binomiale de paramètre $n$ et $p^\star$. Ainsi, nous avons :

**Point méthode (Test d'une probabilité ; niveau exact)**

Étant donné un échantillons $\mathbf{X} = (X_i)_{1\leq i\leq n}$ de loi de Bernoulli $\mathcal{B}(p^\star)$ alors la statistique de test est

$$n\overline{\mathbf{X}}_n$$

qui a pour loi une loi binomiale $\mathcal{B}in(n, p_0)$ sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau exact $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $p^\star \neq p_0$ | $p^\star < p_0$ | $p^\star > p_0$ |
|---|---|---|---|
| $\Theta_1$ | $]0, p_0[\cup]p_0, 1[$ | $]0, p_0[$ | $]p_0, 1[$ |
| Règle | $n\overline{\mathbf{X}}_n < b_{\alpha/2}^{(n,p_0)}$ ou $n\overline{\mathbf{X}}_n > b_{1-\alpha/2}^{(n,p_0)}$ | $n\overline{\mathbf{X}}_n < b_{\alpha}^{(n,p_0)}$ | $n\overline{\mathbf{X}}_n > b_{1-\alpha}^{(n,p_0)}$ |

où $b_\alpha^{(n,p_0)}$ est le quantile d'ordre $\alpha$ d'une loi binomiale de paramètres $n$ et $p_0$ ; c'est-à-dire que si $B$ suit une loi binomiale de paramètres $n$ et $p_0$ alors $b_\alpha^{(n,p_0)}$ est tel que :

$$\mathbb{P}\left(B \leq b_\alpha^{(n,p_0)}\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `binom.test`. Par exemple :

```
## Simulation d'un jeu de données
p_0<-0.5
n<-100
X <- rbinom(n,1,p_0)
## Test sous H_0 : p_0=0.5 -> acceptation avec 95% de chances
binom.test(c(sum(X),n-sum(X)),p = p_0)
## Test sous H_0 : p_0=0.75 -> rejet normalement
binom.test(c(sum(X),n-sum(X)),p = 0.75)
```

### 7.2.2 Méthode approchée

La deuxième solution est d'utiliser le théorème de la limite centrale en sachant que nous connaissons la variance théorique sous la loi de $\mathcal{H}_0$. Ainsi, nous avons :

**Définition 66 (Test d'une probabilité ; statistique approchée)**

Pour tester un échantillons $\mathbf{X} = (X_i)_{1\leq i\leq n}$ de loi de Bernoulli $\mathcal{B}(p^\star)$ et vérifier si $p^\star = p_0$, nous utilisons la statistique de test suivante :

$$S_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - p_0}{\sqrt{p_0(1-p_0)}}$$

avec $\overline{\mathbf{X}}_n$ la moyenne empirique du $n$-échantillon.

Alors nous avons :

**Point méthode (Test d'une probabilité ; niveau approché)**

Étant donné un échantillons $\mathbf{X} = (X_i)_{1\leq i\leq n}$ de loi de Bernoulli $\mathcal{B}(p^\star)$ alors la statistique de test est

$$S_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - p_0}{\sqrt{p_0(1-p_0)}}$$

qui a pour loi **asymptotique** une loi normale $\mathcal{N}(0,1)$ centrée réduite sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $p^\star \neq p_0$ | $p^\star < p_0$ | $p^\star > p_0$ |
|---|---|---|---|
| $\Theta_1$ | $]0,p_0[\cup]p_0,1[$ | $]0,p_0[$ | $]p_0,1[$ |
| Règle | $\lvert S_n\rvert > q_{1-\alpha/2}$ | $S_n < q_\alpha$ | $S_n > q_{1-\alpha}$ |

où $q_\alpha$ est le quantile d'ordre $\alpha$ d'une loi normale centrée réduite ; c'est-à-dire que si $X$ suit une loi gaussienne centrée réduite $\mathcal{N}(0,1)$ alors $q_\alpha$ est tel que :

$$\mathbb{P}(X \leq q_\alpha) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `prop.test`. Par exemple :

```
## Simulation d'un jeu de données
p_0<-0.5
n<-100
X <- rbinom(n,1,p_0)
## Test sous H_0 : p_0=0.5 -> acceptation avec 95% de chances
prop.test(sum(X),n,p = p_0)
## Test sous H_0 : p_0=0.75 -> rejet normalement
prop.test(sum(X),n,p = 0.75)
```

## 7.3 Test de comparaison de deux probabilités pour deux échantillons appariés

Dans cette section, nous supposons avoir un $n$-échantillon apparié $(X_i, Y_i)_{1\leq i\leq n}$ prenant leurs valeurs dans l'ensemble $\{0,1\}^2$ suivant des lois de Bernoulli $\mathcal{B}(p^\star_{\mathbf{X}})$ et $\mathcal{B}(p^\star_{\mathbf{Y}})$ et symbolisant, par exemple, la présence d'un caractère dans deux expériences. Par exemple, si $X_i = 0$ et $Y_i = 0$, cela signifie que l'individu $i$ n'a pas eu la caractéristique dans aucune des expériences. La question que nous pouvons nous poser est l'égalité des probabilités :

$$\begin{cases} \mathcal{H}_0 & : \quad p^\star_{\mathbf{X}} = p^\star_{\mathbf{Y}}, \\ \mathcal{H}_1 & : \quad p^\star_{\mathbf{X}} \neq p^\star_{\mathbf{Y}}. \end{cases} \tag{7.1}$$

Nous avons donc 4 possibilités pour les paires $(X_i, Y_i)$ et nous comptons le nombre d'individus de chaque paire :

| | | **Y** | |
|---|---|---|---|
| | | 0 | 1 |
| **X** | 0 | $n_{0,0}$ | $n_{0,1}$ |
| | 1 | $n_{1,0}$ | $n_{1,1}$ |

En particulier, nous voyons que pour estimer $p^\star_{\mathbf{X}}$ et $p^\star_{\mathbf{Y}}$, nous pouvons compter le nombre moyen de fois où $X$ (resp. $Y$) vaut 1 :

$$\widehat{p}_{\mathbf{X},n} = \frac{n_{1,1} + n_{1,0}}{n} \text{ et } \widehat{p}_{\mathbf{Y},n} = \frac{n_{1,1} + n_{0,1}}{n}.$$

En particulier, nous voyons que la différence entre les deux estimateurs se fait uniquement sur les valeurs de $n_{1,0}$ et $n_{0,1}$. Ainsi, le test (7.1) est équivalent au test :

$$\begin{cases} \mathcal{H}_0 & : \quad p^{\star(\mathbf{X},\mathbf{Y})}_{0,1} = p^{\star(\mathbf{X},\mathbf{Y})}_{1,0}, \\ \mathcal{H}_1 & : \quad p^{\star(\mathbf{X},\mathbf{Y})}_{0,1} \neq p^{\star(\mathbf{X},\mathbf{Y})}_{1,0} \end{cases}$$

où $p^{\star(\mathbf{X},\mathbf{Y})}_{0,1}$ et $p^{\star(\mathbf{X},\mathbf{Y})}_{1,0}$ sont les probabilités croisées. En théorie, si ces probabilités sont égales, cela signifie que $n_{1,0} \approx n_{0,1}$ et donc, si nous notons $n_{1/0} = n_{1,0} + n_{0,1}$, elles valent chacune la moitié. Autrement dit, nous pouvons résumer par le tableau suivant :

| | $(\mathbf{X},\mathbf{Y})$ | | Total |
|---|---|---|---|
| | $(1,0)$ | $(0,1)$ | |
| Observé | $n_{1,0}$ | $n_{0,1}$ | $n_{1/0}$ |
| Théorique | $\frac{n_{1/0}}{2}$ | $\frac{n_{1/0}}{2}$ | $n_{1/0}$ |

L'idée est donc de regarder la différence entre les effectifs observés et les théoriques. Nous obtenons ainsi :

**Point méthode (Test de Mc Nemar)**

Étant donné deux échantillons appariés $(\mathbf{X},\mathbf{Y})$ ayant chacune une loi de Bernoulli $\mathcal{B}(p^\star_{\mathbf{X}})$ et $\mathcal{B}(p^\star_{\mathbf{Y}})$ alors la statistique de test est

$$mcN_n = \frac{\left(n_{1,0} - \frac{n_{1/0}}{2}\right)^2}{\frac{n_{1/0}}{2}} = \frac{(n_{1,0} - n_{0,1})^2}{n_{1,0} + n_{0,1}}$$

qui a pour loi **asymptotique** une loi du Chi2 $\chi^2(1)$ à 1 degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| | |
|---|---|
| $\mathcal{H}_1$ | $p^\star_{\mathbf{X}} \neq p^\star_{\mathbf{Y}}$ |
| $\Theta_1$ | $]0;1[^2 \setminus \{(x,y) \in ]0;1[^2 \mid x = y\}$ |
| Règle | $mcN_n > s^{(1)}_{1-\alpha}$ |

où $s^{(1)}_\alpha$ est le quantile d'ordre $\alpha$ d'une loi du Chi2 $\chi^2(1)$ à 1 degrés de liberté ; c'est-à-dire que si $S$ suit une loi du Chi2 $\chi^2(1)$ à 1 degrés de liberté alors $s^{(1)}_\alpha$ est tel que :

$$\mathbb{P}\left(S \leq s^{(1)}_\alpha\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `mcnemar.test`. Par exemple :

```
## Simulation d'un jeu de données
p_0<-0.5
n<-100
tab<-table(rbinom(n,1,p_0),rbinom(n,1,p_0))
## Test sous H_0 : p_X=p_Y -> acceptation avec 95% de chances
mcnemar.test(tab)
## Test sous H_0 : p_X= p_Y -> rejet normalement
mcnemar.test(table(rbinom(n,1,p_0),rbinom(n,1,0.9)))
```

## 7.4 Test de comparaison de deux probabilités pour deux échantillons indépendants

Dans le cas de deux échantillons $\mathbf{X}$ et $\mathbf{Y}$ indépendants de tailles $n_{\mathbf{X}}$ et $n_{\mathbf{Y}}$, le fait de comparer les deux probabilités $p^\star_{\mathbf{X}}$ et $p^\star_{\mathbf{Y}}$ comme dans le test (7.1) revient à tester la différence des deux probabilités :

$$\begin{cases} \mathcal{H}_0 & : \quad p^\star_{\mathbf{X}} - p^\star_{\mathbf{Y}} = 0, \\ \mathcal{H}_1 & : \quad p^\star_{\mathbf{X}} - p^\star_{\mathbf{Y}} \neq 0. \end{cases}$$

Or, par analogie, les deux estimateurs sont :

$$\widehat{p}_{\mathbf{X},n} = \frac{1}{n_{\mathbf{X}}} \sum_{i=1}^{n_{\mathbf{X}}} X_i = \frac{\mathbf{X}_+}{n_{\mathbf{X}}} \text{ et } \widehat{p}_{\mathbf{Y},n} = \frac{\mathbf{Y}_+}{n_{\mathbf{Y}}},$$

nous voyons que si les probabilités sont égales, c'est à dire qu'elles valent $p^\star$, alors un estimateur est :

$$\widehat{p}_{n,\mathbf{X},\mathbf{Y}} = \frac{\mathbf{X}_+ + \mathbf{Y}_+}{n_{\mathbf{X}} + n_{\mathbf{Y}}}.$$

A partir de ces notations, nous avons donc :

**Point méthode (Test d'égalité des proportions ; échantillons indépendants)**

Étant donné deux échantillons indépendants $\mathbf{X}$ et $\mathbf{Y}$ de tailles $n_{\mathbf{X}}$ et $n_{\mathbf{Y}}$ ayant chacune une loi de Bernoulli $\mathcal{B}(p^\star_{\mathbf{X}})$ et $\mathcal{B}(p^\star_{\mathbf{Y}})$ alors la statistique de test est

$$S_{n_{\mathbf{X}},n_{\mathbf{Y}}} = \frac{\widehat{p}_{\mathbf{X},n} - \widehat{p}_{\mathbf{Y},n}}{\sqrt{\widehat{p}_{n,\mathbf{X},\mathbf{Y}}\left(1 - \widehat{p}_{n,\mathbf{X},\mathbf{Y}}\right)\left(\frac{1}{n_{\mathbf{X}}} + \frac{1}{n_{\mathbf{Y}}}\right)}}$$

qui a pour loi **asymptotique** une loi normale $\mathcal{N}(0,1)$ centrée réduite sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $p^\star_{\mathbf{X}} \neq p^\star_{\mathbf{Y}}$ | $p^\star_{\mathbf{X}} < p^\star_{\mathbf{Y}}$ | $p^\star_{\mathbf{X}} > p^\star_{\mathbf{Y}}$ |
|---|---|---|---|
| $\Theta_1$ | $]0,p_0[\cup]p_0,1[$ | $]0,p_0[$ | $]p_0,1[$ |
| Règle | $\lvert S_{n_{\mathbf{X}},n_{\mathbf{Y}}}\rvert > q_{1-\alpha/2}$ | $S_{n_{\mathbf{X}},n_{\mathbf{Y}}} < q_\alpha$ | $S_{n_{\mathbf{X}},n_{\mathbf{Y}}} > q_{1-\alpha}$ |

où $q_\alpha$ est le quantile d'ordre $\alpha$ d'une loi normale centrée réduite ; c'est-à-dire que si $X$ suit une loi gaussienne centrée réduite $\mathcal{N}(0,1)$ alors $q_\alpha$ est tel que :

$$\mathbb{P}(X \leq q_\alpha) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test bilatéral est `prop.test`. Par exemple :

```
## Simulation d'un jeu de données
nX<-100
nY<-200
X<-rbinom(nX,1,0.5)
Y<-rbinom(nY,1,0.5)
## Test sous H_0 : p_X=p_Y -> acceptation avec 95% de chances
prop.test(matrix(c(sum(X),sum(Y),nX,nY),2,2))
## Test sous H_0 : p_X= p_Y -> rejet normalement
prop.test(matrix(c(sum(X),sum(rbinom(nY,1,0.75)),nX,nY),2,2))
```

## 7.5 Généralisation à plus de deux échantillons appariés

Dans le cas de $K$ échantillons appariés $\mathbf{X} = (X_{i,k})$ dont nous cherchons à savoir si toutes les proportions sont égales (c'est-à-dire que, qu'importe l'expérience, nous retrouvons la même proportion), nous avons le test :

$$\begin{cases} \mathcal{H}_0 & : \quad p_1^\star = p_2^\star = \cdots = p_K^\star, \\ \mathcal{H}_1 & : \quad \exists k \neq k',\ p_k^\star \neq p_{k'}^\star. \end{cases}$$

Dans ce cas, le test utilisé est une généralisation du test de Mc Nemar appelé test de Cochran-Mantel-Haenszel.

**Codage en R**

La fonction R permettant de faire le test bilatéral est `mantelhaen.test`. Par exemple :

```
## Simulation d'un jeu de données
X<-rbinom(100,1,0.5)
Y<-rbinom(100,1,0.5)
Z<-rbinom(100,1,0.5)
## Test sous H_0 : p_X=p_Y -> acceptation avec 95% de chances
mantelhaen.test(table(X,Y,Z))
## Test sous H_0 : p_X= p_Y -> rejet normalement
mantelhaen.test(table(X,rbinom(100,1,0.1),rbinom(100,1,0.9)))
```

## 7.6 Test de comparaison de plus de deux probabilités pour des échantillons indépendants

Enfin, nous nous intéressons au cas où un individu peut prendre une caractéristique $A_i$ avec $i \in \{1, \ldots, I\}$ dans une population $M_j$ où $j \in \{1, \ldots, J\}$ et nous nous intéressons au fait que la probabilité $p_{i,j}^\star$ d'avoir cette caractéristique dans la population $M_j$ soit invariable qu'importe la population. Autrement dit, pour tout couple $(j, j')$ de $\{1, \ldots, J\}^2$, $p_{i,j}^\star$ est égale à $p_{i,j'}^\star$. De plus, nous ajoutons l'hypothèse que cette propriété est vraie qu'importe la caractéristique. Au final, nous avons donc :

$$\begin{cases} \mathcal{H}_0 & : \quad \forall i \in \{1, \ldots, I\}, \forall\, (j, j') \in \{1, \ldots, J\}^2,\ p_{i,j}^\star = p_{i,j'}^\star, \\ \mathcal{H}_1 & : \quad \exists i \in \{1, \ldots, I\} \text{ et } \exists\, (j, j') \in \{1, \ldots, J\}^2,\ p_{i,j}^\star \neq p_{i,j'}^\star. \end{cases}$$

Comme précédemment, nous regardons les effectifs théoriques si nous avions l'indépendance et les comparons aux effectifs observés. Pour cela, nous avons besoin des notations suivantes :

| | | Population | | | | Total |
|---|---|---|---|---|---|---|
| | | $M_1$ | $M_2$ | $\cdots$ | $M_J$ | |
| Caractéristique | $A_1$ | $n_{1,1}$ | $n_{1,2}$ | $\cdots$ | $n_{1,J}$ | $n_{1,\cdot} = \sum_{j=1}^{J} n_{1,j}$ |
| | $A_2$ | $n_{2,1}$ | $n_{2,2}$ | $\cdots$ | $n_{2,J}$ | $n_{2,\cdot} = \sum_{j=1}^{J} n_{2,j}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $A_I$ | $n_{I,1}$ | $n_{I,2}$ | $\cdots$ | $n_{I,J}$ | $n_{I,\cdot} = \sum_{j=1}^{J} n_{I,j}$ |
| Total | | $n_{\cdot,1} = \sum_{i=1}^{I} n_{i,1}$ | $n_{\cdot,2} = \sum_{i=1}^{I} n_{i,2}$ | $\cdots$ | $n_{\cdot,J} = \sum_{i=1}^{I} n_{i,J}$ | $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j}$ |

et nous avons donc :

**Point méthode (Test de Chi2 d'indépendance de Pearson)**

Étant donné $K$ échantillons indépendants ayant chacune une loi de Bernoulli $\mathcal{B}(p_k^\star)$ alors la statistique de test est

$$\chi_n^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{i,j} - c_{i,j})^2}{c_{i,j}} \text{ avec } c_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{n}$$

qui a pour loi **asymptotique** une loi du Chi2 $\chi^2[(I-1)(J-1)]$ à $(I-1)(J-1)$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$ et la règle de rejet de niveau **asymptotique** $\alpha \in ]0;1[$ est donc :

| $\mathcal{H}_1$ | $\exists i \in \{1,\dots,I\}$ et $\exists (j,j') \in \{1,\dots,J\}^2$, $p_{i,j}^\star \neq p_{i,j'}^\star$ |
|---|---|
| Règle | $\chi_n^2 > s_{1-\alpha}^{[(I-1)(J-1)]}$ |

où $s_\alpha^{[(I-1)(J-1)]}$ est le quantile d'ordre $\alpha$ d'une loi du Chi2 $\chi^2(1)$ à $(I-1)(J-1)$ degrés de liberté ; c'est-à-dire que si $S$ suit une loi du Chi2 $\chi^2[(I-1)(J-1)]$ à $(I-1)(J-1)$ degrés de liberté alors $s_\alpha^{[(I-1)(J-1)]}$ est tel que :

$$\mathbb{P}\left(S \leq s_\alpha^{[(I-1)(J-1)]}\right) = \alpha.$$

**Codage en R**

La fonction R permettant de faire le test unilatéral est `chisq.test`.

# Chapitre 8

# Tests non-paramétriques

## 8.1 Test de Kolmogorov-Smirnov

Le deuxième test d'adéquation est celui de Kolmogorov-Smirnov et se base cette fois sur la fonction de répartition. Le principe est d'estimer cette dernière.

**Définition 67 (Fonction de répartition empirique)**

Étant donné un $n$-échantillon $\mathbf{X}$ de variables réelles, nous appelons **fonction de répartition empirique** la fonction définie sur $\mathbb{R}$ par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Le principe du test est de calculer le plus grand écart entre la fonction de répartition recherchée et la fonction de répartition empirique.

**Proposition 27 (Test de Kolmogorov-Smirnov)**

Étant donnés un $n$-échantillon $\mathbf{X}$ de variable réelle et une loi de référence de fonction de répartition $F^{(0)}$, alors nous définissons la statistique de test

$$D_n = \sqrt{n}\, sup_{x\in\mathbb{R}} \left| F^{(0)}(x) - F_n(x) \right|$$

et, sous l'hypothèse $\mathcal{H}_0$, nous avons :

$$D_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \sup_{x\in\mathbb{R}} \left| B_{F^{(0)}(x)} \right|$$

où $(B_t)_{t\in[0;1]}$ est un pont brownien ; c'est-à-dire un mouvement brownien conditionné par le fait que $B_0 = B_1 = 0$.

En particulier, dans le cas d'une fonction de répartition continue, nous avons pour toute constante $c$ strictement positive :

$$\mathbb{P}\left(D_n > c\right) \underset{n\to+\infty}{\longrightarrow} \alpha(c) = 2\sum_{r=1}^{+\infty} (-1)^{r-1} e^{-2c^2 r^2}.$$

**Remarque**

Nous remarquons qu'il existe une formule explicite pour calculer la $p$-valeur même si nous ne devrons nous contenter d'approximations numériques dans la pratique. En particulier, $\alpha(1,36) \approx 0,05$. Sur la figure 9.1, nous avons représenté la fonction $c \mapsto \alpha(c)$.

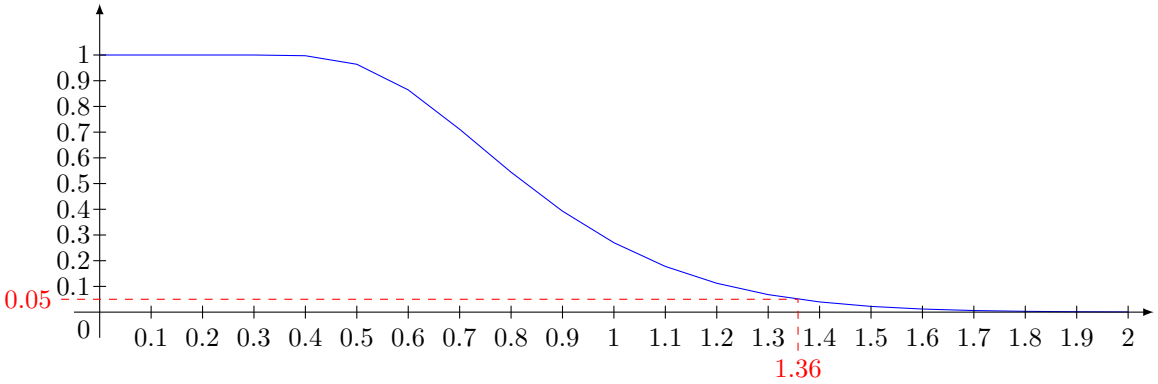



FIGURE 8.1 – Représentation de la fonction $c \mapsto \alpha(c)$ qui représente la probabilité $\mathbb{P}\left(\sup_{t\in[0;1]} |B_t| \geq c\right)$ (donc la $p$-valeur). La ligne hachée rouge permet de retrouver la valeur correspondante à une probabilité de 0.05.

**Corollaire 28 (Décision)**

À partir de la proposition 36, nous pouvons rejeter l'hypothèse nulle si la statistique $D_n$ dépasse le quantile d'ordre $1 - \alpha$ de la loi présentée dans la proposition.

# Chapitre 9

# Autres tests

## 9.1 Botanique des tests

Dans cette partie, nous mettons quelques tests classiques.

### 9.1.1 Tests paramétriques classiques

Dans cette section, nous mettons des tests classiques.

#### Test de proportion

Dans cette partie, nous testons la proportion d'une quantité (par exemple, le nombre d'étudiants du M1 de Maths Générale aimant la statistique) par rapport à une quantité fixée. Comme cette question peut être ramenée à une loi de Bernoulli, nous supposons avoir un $n$-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi de Bernoulli $\mathcal{B}(p^\star)$, une valeur $p_0$ de référence et nous regardons l'un des tests suivants :

$$\begin{cases} \mathcal{H}_0 : & p^\star = p_0, \\ \mathcal{H}_1 : & p^\star \neq p_0 \text{ ou } p^\star < p_0 \text{ ou } p^\star > p_0 \text{ (choix à fixer suivant le contexte).} \end{cases}$$

Notons que nous n'abordons pas le cas $\mathcal{H}_1 : p^\star = p_1$ car le théorème 20 montre que le test de vraisemblance est uniformément plus puissant.

Pour la suite, nous présentons le cas où l'hypothèse alternative est $p^\star \neq p_0$, la démarche étant similaire pour les autres. pour estimer le paramètre $p^\star$, nous utilisons la moyenne empirique qui est l'estimateur du maximum de vraisemblance et celui des moments basé sur l'espérance. Par le théorème de la limite centrale, nous avons, sous l'hypothèse $\mathcal{H}_0$, :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - p_0\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, p_0(1-p_0)\right).$$

Ainsi, la statistique la plus logique est :

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - p_0}{\sqrt{p_0(1-p_0)}}$$

qui suit asymptotiquement une loi normale centrée réduite.

**Corollaire 29 (Décision)**

À partir des propriétés que nous venons de voir, nous pouvons rejeter l'hypothèse nulle si la statistique $T_n$ dépasse le quantile d'ordre $1-\alpha/2$ ou est plus petite que le quantile d'ordre $\alpha/2$ de la loi gaussienne centrée réduite.

#### Test des moyennes

Dans cette partie, nous nous plaçons dans le cas d'un $n$-échantillon avec un moment d'ordre 2 et nous cherchons à savoir si son moment d'ordre 1 $\mu^\star$ est égal à une valeur connue ou pas. Ainsi, nous regardons l'un des tests suivants :

$$\begin{cases} \mathcal{H}_0 : & \mu^\star = \mu_0, \\ \mathcal{H}_1 : & \mu^\star \neq \mu_0 \text{ ou } \mu^\star < \mu_0 \text{ ou } \mu^\star > \mu_0 \text{ (choix à fixer suivant le contexte).} \end{cases}$$

Comme pour le test des proportions, une estimation de la moyenne se fait grâce à la moyenne empirique $\overline{\mathbf{X}}_n$ et, sous l'hypothèse, $\mathcal{H}_0$, nous avons :

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu_0\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{V}\left[X_1\right]\right).$$

Si nous connaissons la variance, nous pouvons reprendre la méthode précédente. Nous nous plaçons donc dans le cas où la variance est inconnue.

**Hypothèse**

Dans la suite de cette partie, nous supposons que la variance est inconnue.

Comme nous ne connaissons pas la moyenne, nous l'estimons par l'estimateur empirique :

$$S_n^2 = \overline{X_n^2} - \overline{\mathbf{X}}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}_n\right)^2.$$

Nous commençons par calculer la loi de cet estimateur.

**Proposition 30 (Loi de $S_n^2$)**

Étant donné un $n$-échantillon $\mathbf{X}$ admettant un moment d'ordre 2 alors $nS_n^2/\sigma$ (où $\sigma$ est l'écart-type commun) suit une loi du $\chi^2(n-1)$ ; c'est-à-dire une loi du $\chi^2$ à $n-1$ degrés de libertés.

**Preuve**

La preuve se base sur un théorème que nous admettrons car il utilise des notions de projections (donc d'espérances conditionnelles) qui ne seront abordées que durant le deuxième semestre. Nous mettons toutefois l'énoncé ici.

**Lemme 31 (Théorème de Cochran)**

Étant donnés un vecteur gaussien $\mathbf{X}$ de $\mathbb{R}^d$ de loi $\mathcal{N}\left(m, \sigma^2\mathbb{I}_d\right)$ et $F_1, \ldots, F_P$ des sous-espaces vectoriels de $\mathbb{R}^d$, orthogonaux deux à deux et de somme $\mathbb{R}^d$ dont nous notons $P_{F_p}$ les matrices de projection orthogonale sur $F_p$ et $n_p$ la dimension de $F_p$ alors nous avons :

- les vecteurs aléatoires $P_{F_1}\mathbf{X}, \ldots, P_{F_P}\mathbf{X}$ sont deux à deux indépendants et de lois respectives $\mathcal{N}\left(P_{F_1}m, \sigma^2 P_{F_1}\right), \ldots, \mathcal{N}\left(P_{F_P}m, \sigma^2 P_{F_P}\right)$.
- les variables aléatoires $\frac{\left\|P_{F_1}(\mathbf{X}-m)\right\|^2}{\sigma^2}, \ldots, \frac{\left\|P_{F_P}(\mathbf{X}-m)\right\|^2}{\sigma^2}$ sont deux à deux indépendantes et sont de lois respectives $\chi^2(n_1), \ldots, \chi^2(n_P)$.

Pour démontrer la proposition 30, nous commençons par définir pour tout $i \in \{1, \ldots, n\}$, la variable :

$$Z_i = \frac{X_i - \overline{\mathbf{X}}_n}{\sigma}.$$

Calculons l'espérance et la matrice de variance-covariance de la variable $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Sous l'hypothèse $\mathcal{H}_0$, nous avons de plus pour tout $i \in \{1, \ldots, n\}$ :

$$\begin{aligned}
\mathbb{E}\left[Z_i\right] &= \mathbb{E}\left[\frac{X_i - \overline{\mathbf{X}}_n}{\sigma}\right] \\
&= \frac{\mathbb{E}\left[X_i\right] - \mathbb{E}\left[\overline{\mathbf{X}}_n\right]}{\sigma} \\
&= \frac{\mu^\star - \mu^\star}{\sigma} \\
&= 0.
\end{aligned}$$

De plus, pour tout $i \in \{1, \dots, n\}$, nous avons :

$$\begin{aligned}
\mathbb{V}\left[\frac{X_i - \overline{\mathbf{X}}_n}{\sigma}\right] &= \frac{1}{\sigma^2}\mathbb{V}\left[X_i \frac{1}{n}\sum_{j=1}^{n} 1 - \frac{1}{n}\sum_{j=1}^{n} X_j\right] \\
&= \frac{1}{n^2\sigma^2}\mathbb{V}\left[\sum_{j=1}^{n}(X_i - X_j)\right] \\
&= \frac{1}{n^2\sigma^2}\mathbb{V}\left[\sum_{\substack{j=1 \\ j\neq i}}^{n}(X_i - X_j)\right] \\
&= \frac{1}{n^2\sigma^2}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n}\mathbb{V}\left[X_i - X_j\right] + \sum_{\substack{j_1=1 \\ j_1\neq i}}^{n}\sum_{\substack{j_2=1 \\ j_2\neq j_1 \text{ et } j_2\neq i}}^{n}\operatorname{Cov}\left(X_i - X_{j_1}, X_i - X_{j_2}\right)\right) \\
&= \frac{1}{n^2\sigma^2}\left[\sum_{\substack{j=1 \\ j\neq i}}^{n}\left(\mathbb{V}\left[X_i\right] + \mathbb{V}\left[X_j\right] - 2\underbrace{\operatorname{Cov}\left(X_i, X_j\right)}_{=0}\right)\right. \\
&\qquad \left. + \sum_{\substack{j_1=1 \\ j_1\neq i}}^{n}\sum_{\substack{j_2=1 \\ j_2\neq j_1 \text{ et } j_2\neq i}}^{n}\left(\mathbb{V}\left[X_i\right] - \underbrace{\operatorname{Cov}\left(X_{j_1}, X_i\right)}_{=0} - \underbrace{\operatorname{Cov}\left(X_i, X_{j_2}\right)}_{=0} + \underbrace{\operatorname{Cov}\left(X_{j_1}, X_{j_2}\right)}_{=0}\right)\right] \\
&= \frac{1}{n^2\sigma^2}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} 2\sigma^2 + \sum_{\substack{j_1=1 \\ j_1\neq i}}^{n}\sum_{\substack{j_2=1 \\ j_2\neq j_1 \text{ et } j_2\neq i}}^{n}\sigma^2\right) \\
&= \frac{1}{n^2\sigma^2}\left[2(n-1)\sigma^2 + (n-1)(n-2)\sigma^2\right] \\
&= \frac{n(n-1)\sigma^2}{n^2\sigma^2} \\
&= 1 - \frac{1}{n}.
\end{aligned}$$

Enfin, pour tout $1 \leq i_1 \neq i_2 \leq n$, nous avons :

$$\begin{aligned}
&\operatorname{Cov}\left(\frac{X_{i_1} - \overline{\mathbf{X}}_n}{\sigma}, \frac{X_{i_2} - \overline{\mathbf{X}}_n}{\sigma}\right) \\
&= \frac{1}{\sigma^2}\operatorname{Cov}\left(X_{i_1} - \overline{\mathbf{X}}_n, X_{i_2} - \overline{\mathbf{X}}_n\right) \\
&= \frac{1}{\sigma^2}\left[\underbrace{\operatorname{Cov}\left(X_{i_1}, X_{i_2}\right)}_{=0} - \operatorname{Cov}\left(X_{i_1}, \overline{\mathbf{X}}_n\right) - \operatorname{Cov}\left(\overline{\mathbf{X}}_n, X_{i_2}\right) + \operatorname{Cov}\left(\overline{\mathbf{X}}_n, \overline{\mathbf{X}}_n\right)\right] \\
&= \frac{1}{\sigma^2}\left[-\frac{1}{n}\sum_{j=1}^{n}\operatorname{Cov}\left(X_{i_1}, X_j\right) - \frac{1}{n}\sum_{j=1}^{n}\operatorname{Cov}\left(X_j, X_{i_2}\right) + \mathbb{V}\left[\overline{\mathbf{X}}_n\right]\right] \\
&= \frac{1}{\sigma^2}\left(-\frac{1}{n}\mathbb{V}\left[X_{i_1}\right] - \frac{1}{n}\mathbb{V}\left[X_{i_2}\right] + \frac{1}{n^2}\mathbb{V}\left[\sum_{j=1}^{n} X_j\right]\right) \\
&= \frac{1}{n\sigma^2}\left[-2\sigma^2 + \frac{1}{n}\left(\sum_{j=1}^{n}\mathbb{V}\left[X_j\right] + \sum_{j_1=1}^{n}\sum_{\substack{j_2=1 \\ j_2\neq j_1}}^{n}\underbrace{\operatorname{Cov}\left(X_{j_1}, X_{j_2}\right)}_{=0}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n\sigma^2}\left(-2\sigma^2 + \frac{n\sigma^2}{n}\right)\\
&= \frac{-\sigma^2}{n\sigma^2}\\
&= \frac{-1}{n}
\end{aligned}
$$

Donc la loi commune des $Z_i$ est centrée et matrice de variance covariance :

$$\Gamma_n = \mathbb{I}_n - \begin{pmatrix} 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{pmatrix} \begin{pmatrix} 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{pmatrix}.$$

Par le théorème multivarié de la limite centrale, la loi de la variable $\frac{Z_1+\ldots+Z_n}{n}$ converge vers loi normale centrée de covariance $\Gamma_n$. Or, cette loi est le projeté d'un vecteur aléatoire de $\mathbb{R}^K$ suivant une loi normale centrée réduite sur l'hyperplan orthogonal au vecteur colonne $(1/\sqrt{n}, \cdots, 1/\sqrt{n})^T$.
De plus, nous voyons que la variable $nS_n^2/\sigma$ est la norme au carré de cette variable donc, d'après le théorème de Cochran, la loi asymptotique est une loi du $\chi^2$ à $n-1$ degrés de libertés.

**Corollaire 32 (Loi de la statistique)**

Étant donné un $n$-échantillon $\mathbf{X}$ admettant un moment d'ordre 2 alors

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sqrt{S_n^2}}$$

suit asymptotiquement une loi de Student centrée à $n-1$ degrés de liberté.

**Preuve**

Nous avons :

$$
\begin{aligned}
\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sqrt{S_n^2}} &= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}}{\frac{\sqrt{S_n^2}}{\sigma}}\\
&= \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}}{\frac{1}{\sqrt{n}}\sqrt{\frac{nS_n^2}{\sigma^2}}}\\
&= \frac{1}{\sqrt{\frac{n}{n-1}}} \times \frac{\sqrt{n}\frac{\overline{\mathbf{X}}_n - \mu_0}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2(n-1)}}}\\
&= \underbrace{\frac{1}{\sqrt{\frac{n-1}{n}}}}_{\underset{n\to+\infty}{\longrightarrow} 1} \times \frac{Z}{\sqrt{\frac{Y}{n-1}}}
\end{aligned}
$$

avec $Z$ qui suit asymptotiquement une loi normale centrée réduite et $Y$ qui suit une loi du $\chi^2$ à $n-1$ degrés de liberté donc, par définition de la loi de Student, nous avons le résultat.

**Corollaire 33 (Décision)**

À partir des propriétés que nous venons de voir, nous pouvons rejeter l'hypothèse nulle si la statistique $T_n$ dépasse le quantile d'ordre $1-\alpha/2$ ou est plus petite que le quantile d'ordre $\alpha/2$ de la loi Student centrée de $n-1$ degrés de libertés.

### 9.1.2 Tests d'adéquation d'une loi de probabilité à des données

Dans cette partie, nous nous intéressons au cas où le $n$-échantillon $\mathbf{X}$ suit une loi de référence de densité $f_0$. Le test effectué est donc le suivant :

$$\begin{cases} \mathcal{H}_0 : & \text{La loi de } \mathbf{X} \text{ a pour densité } f_0, \\ \mathcal{H}_1 : & \text{La loi de } \mathbf{X} \text{ n'a pas pour densité } f_0. \end{cases}$$

Nous présentons ici deux tests qui permettent de répondre à cette question.

#### Test du $\chi^2$ d'adéquation

Dans cette partie, nous présentons essentiellement le cas où les valeurs de $\mathbf{X}$ sont dans un ensemble fini $I = \{a_1, \dots, a_K\}$.

**Définitions 68 (Loi multinomiale)**

Étant donnée un $n$-échantillon $\mathbf{X}$ prenant ses valeurs dans un ensemble fini $I = \{a_1, \dots, a_K\}$, nous notons $p_k$ la probabilité que la variable $X_1$ soit égale à $a_k$ pour tout $k$ compris entre 1 et $K$ :

$$p_k = \mathbb{P}(X_1 = a_k).$$

Pour tout $k \in \{1, \dots, K\}$, nous notons $N_k$ la variable aléatoire qui compte le nombre de variables $X_i$ qui ont pris la modalité $a_k$ :

$$N_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = a_k\}}.$$

Nous disons alors que la variable $\mathbf{N} = (N_1, \dots, N_K)$ suit une **loi multinomiale de paramètres** $n$ **et** $(p_1, \dots, p_K)$, notée $\mathcal{M}(n; p_1, \dots, p_K)$, dont la densité est :

$$\mathbb{P}(N_1 = n_1, \dots, N_K = n_k) = \frac{n!}{\prod_{k=1}^{K} n_k!} \left( \prod_{k=1}^{K} p_k \right) \mathbb{1}_{\{\sum_{k=1}^{K} n_k = n\}}.$$

**Remarque**

L'ensemble des valeurs possibles de la variable $\mathbf{N}$ est donc l'ensemble des valeurs discrètes d'un simplexe :

$$\left\{ (n_1, \dots, n_K) \in \{0, \dots, n\}^K \,\middle|\, \sum_{k=1}^{K} n_k = n \right\}.$$

Comme une loi discrète est entièrement caractérisée par les probabilités de valoir chaque modalité, le test revient à savoir si :

$$\begin{cases} \mathcal{H}_0 : & \forall k \in \{1, \dots, K\},\ p_k^\star = p_k^{(0)}, \\ \mathcal{H}_1 : & \exists k \in \{1, \dots, K\},\ p_k^\star \neq p_k^{(0)}. \end{cases}$$

**Proposition 34 (Test du $\chi^2$ d'adéquation)**

Étant donnés un $n$-échantillon de loi définie sur un ensemble discret de probabilité $p_k^\star$ va valoir la modalité $a_k$ et une loi de référence de probabilité $p_k^{(0)}$ de valoir la modalité $a_k$, alors nous définissons la statistique de test

$$T_n = \sum_{k=1}^{K} \frac{\left( N_k - n p_k^{(0)} \right)^2}{n p_k^{(0)}}$$

et, sous l'hypothèse $\mathcal{H}_0$, nous avons :

$$T_n \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(K-1)$$

c'est-à-dire une loi du $\chi^2$ à $K-1$ degrés de libertés.

**Preuve**

Pour démontrer la proposition 34, nous commençons par définir pour tout $i \in \{1,\dots,n\}$, la variable :

$$Z_i = \left( \frac{\mathbb{1}_{\{X_i=a_1\}} - p_1^{(0)}}{\sqrt{p_1^{(0)}}}, \dots, \frac{\mathbb{1}_{\{X_i=a_K\}} - p_K^{(0)}}{\sqrt{p_K^{(0)}}} \right).$$

Calculons l'espérance et la matrice de variance-covariance de la variable $\mathbf{Z} = (Z_1,\dots,Z_n)$. Sous l'hypothèse $\mathcal{H}_0$, nous avons de plus pour tout $i \in \{1,\dots,n\}$ :

$$\begin{aligned}
\mathbb{E}[Z_i] &= \mathbb{E}\left[ \left( \frac{\mathbb{1}_{\{X_i=a_1\}} - p_1^{(0)}}{\sqrt{p_1^{(0)}}}, \dots, \frac{\mathbb{1}_{\{X_i=a_K\}} - p_K^{(0)}}{\sqrt{p_K^{(0)}}} \right) \right] \\
&= \left( \mathbb{E}\left[ \frac{\mathbb{1}_{\{X_i=a_1\}} - p_1^{(0)}}{\sqrt{p_1^{(0)}}} \right], \dots, \mathbb{E}\left[ \frac{\mathbb{1}_{\{X_i=a_K\}} - p_K^{(0)}}{\sqrt{p_K^{(0)}}} \right] \right) \\
&= \left( \frac{\mathbb{E}[\mathbb{1}_{\{X_i=a_1\}}] - p_1^{(0)}}{\sqrt{p_1^{(0)}}}, \dots, \frac{\mathbb{E}[\mathbb{1}_{\{X_i=a_K\}}] - p_K^{(0)}}{\sqrt{p_K^{(0)}}} \right) \\
&= (0,\dots,0).
\end{aligned}$$

De plus, pour tout $k \in \{1,\dots,K\}$, nous avons :

$$\begin{aligned}
\mathbb{V}\left[ \frac{\mathbb{1}_{\{X_i=a_k\}} - p_k^{(0)}}{\sqrt{p_k^{(0)}}} \right] &= \frac{\mathbb{V}[\mathbb{1}_{\{X_i=a_k\}}]}{p_k^{(0)}} \\
&= \frac{p_k^{(0)}\left(1-p_k^{(0)}\right)}{p_k^{(0)}} \\
&= 1 - p_k^{(0)}.
\end{aligned}$$

Enfin, pour tout $1 \leq k \neq \ell \leq K$, nous avons :

$$\begin{aligned}
&\mathrm{Cov}\left( \frac{\mathbb{1}_{\{X_i=a_k\}} - p_k^{(0)}}{\sqrt{p_k^{(0)}}}, \frac{\mathbb{1}_{p_\ell^{(0)}}\{X_i=a_\ell\} - p_\ell^{(0)}}{\sqrt{p_\ell^{(0)}}} \right) \\
&= \frac{1}{\sqrt{p_k^{(0)} p_\ell^{(0)}}} \mathrm{Cov}\left( \mathbb{1}_{\{X_i=a_k\}} - p_k^{(0)}, \mathbb{1}_{\{X_i=a_\ell\}} - p_\ell^{(0)} \right) \\
&= \frac{1}{\sqrt{p_k^{(0)} p_\ell^{(0)}}} \left( \mathbb{E}\left[ \left( \mathbb{1}_{\{X_i=a_k\}} - p_k^{(0)} \right) \left( \mathbb{1}_{\{X_i=a_\ell\}} - p_\ell^{(0)} \right) \right] - \underbrace{\mathbb{E}\left[ \mathbb{1}_{\{X_i=a_k\}} - p_k^{(0)} \right]}_{=0} \underbrace{\mathbb{E}\left[ \mathbb{1}_{\{X_i=a_\ell\}} - p_\ell^{(0)} \right]}_{=0} \right) \\
&= \frac{1}{\sqrt{p_k^{(0)} p_\ell^{(0)}}} \mathbb{E}\left[ \mathbb{1}_{\{X_i=a_k\}} \mathbb{1}_{\{X_i=a_\ell\}} - p_k^{(0)} \mathbb{1}_{\{X_i=a_\ell\}} - \mathbb{1}_{\{X_i=a_k\}} p_\ell^{(0)} + p_k^{(0)} p_\ell^{(0)} \right] \\
&= \frac{1}{\sqrt{p_k^{(0)} p_\ell^{(0)}}} \left( \underbrace{\mathbb{E}\left[ \mathbb{1}_{\{X_i=a_k\}} \mathbb{1}_{\{X_i=a_\ell\}} \right]}_{=0 \text{ car } k \neq \ell} - p_k^{(0)} \mathbb{E}\left[ \mathbb{1}_{\{X_i=a_\ell\}} \right] - \mathbb{E}\left[ \mathbb{1}_{\{X_i=a_k\}} \right] p_\ell^{(0)} + p_k^{(0)} p_\ell^{(0)} \right) \\
&= \frac{-p_k^{(0)} p_\ell^{(0)}}{\sqrt{p_k^{(0)} p_\ell^{(0)}}}
\end{aligned}$$

$$= -\sqrt{p_k^{(0)} p_\ell^{(0)}}.$$

Donc la loi commune des $Z_i$ est centrée et matrice de variance covariance :

$$\Gamma = \mathbb{I}_K - \begin{pmatrix} \sqrt{p_1^{(0)}} \\ \vdots \\ \sqrt{p_K^{(0)}} \end{pmatrix} \begin{pmatrix} \sqrt{p_1^{(0)}} & \cdots & \sqrt{p_K^{(0)}} \end{pmatrix}.$$

Par le théorème multivarié de la limite centrale, la loi de la variable $\frac{Z_1+\ldots+Z_n}{n}$ converge vers la loi normale centrée de covariance $\Gamma$. Or, cette loi est le projeté d'un vecteur aléatoire de $\mathbb{R}^K$ suivant une loi normale centrée réduite sur l'hyperplan orthogonal au vecteur colonne $\left(\sqrt{p_1^{(0)}}, \cdots, \sqrt{p_K^{(0)}}\right)^T$.
De plus, nous voyons que la variable $T_n$ est la norme au carré de cette variable donc, d'après le théorème de Cochran, la loi asymptotique est une loi du $\chi^2$ à $K-1$ degrés de libertés.

**Attention au piège**

Quand nous appliquons le théorème, l'erreur la plus classique est de penser que c'est une loi du $\chi^2$ à $K$ degrés de liberté (au lieu de $K-1$). Une aide mémoire possible est de se souvenir que la dernière coordonnée du vecteur est totalement définie une fois les $K-1$ premières sont fixées (nous le voyons notamment dans la matrice $\Gamma$ de la démonstration qui est la matrice identité moins la matrice de rang 1.

**Corollaire 35 (Décision)**

À partir de la proposition 34, nous pouvons rejeter l'hypothèse nulle si la statistique $T_n$ dépasse le quantile d'ordre $1-\alpha$ de la loi du $\chi^2$ à $K-1$ degrés de liberté.

**Remarque**

Il existe des versions de ce test pour le cas de données continues en découpant en classes et en calculant les probabilités d'appartenir à une classe (voir par exemple Tenenhaus (2007), chapitre 2, section 3, sous-section 5). L'inconvénient est alors que le résultat dépend du choix de la classe.

**La (mauvaise) utilisation de la statistique à travers les âges**

Dans leurs articles controversés, Reinhart et Rogoff (2010) étudient l'influence de la dette publique sur la croissance d'un pays. En plus d'une erreur Excel, qui a largement été documentée (voir par exemple la vidéo Youtube de Louapre ou son blog lou (2020)), les auteurs choisissent des classes qu'on peut juger arbitraires (notamment la borne critique de 90% du PIB) qui a des conséquences sur beaucoup de politiques d'austérité (dont la France).

### Test de Kolmogorov-Smirnov

Le deuxième test d'adéquation est celui de Kolmogorov-Smirnov et se base cette fois sur la fonction de répartition. Le principe est d'estimer cette dernière.

**Définition 69 (Fonction de répartition empirique)**

Étant donné un $n$-échantillon $\mathbf{X}$ de variables réelles, nous appelons **fonction de répartition empirique** la fonction définie sur $\mathbb{R}$ par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Le principe du test est de calculer le plus grand écart entre la fonction de répartition recherchée et la fonction de répartition empirique.

**Proposition 36 (Test de Kolmogorov-Smirnov)**

Étant donnés un $n$-échantillon $\mathbf{X}$ de variable réelle et une loi de référence de fonction de répartition $F^{(0)}$, alors nous définissons la statistique de test

$$D_n = \sqrt{n}\, sup_{x\in\mathbb{R}} \left| F^{(0)}(x) - F_n(x) \right|$$

et, sous l'hypothèse $\mathcal{H}_0$, nous avons :

$$D_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \sup_{x\in\mathbb{R}} \left| B_{F^{(0)}(x)} \right|$$

où $(B_t)_{t\in[0;1]}$ est un pont brownien ; c'est-à-dire un mouvement brownien conditionné par le fait que $B_0 = B_1 = 0$.

En particulier, dans le cas d'une fonction de répartition continue, nous avons pour toute constante $c$ strictement positive :

$$\mathbb{P}\left(D_n > c\right) \underset{n\to+\infty}{\longrightarrow} \alpha(c) = 2\sum_{r=1}^{+\infty} (-1)^{r-1} e^{-2c^2 r^2}.$$

**Remarque**

Nous remarquons qu'il existe une formule explicite pour calculer la $p$-valeur même si nous ne devrons nous contenter d'approximations numériques dans la pratique. En particulier, $\alpha(1,36) \approx 0,05$. Sur la figure 9.1, nous avons représenté la fonction $c \mapsto \alpha(c)$.

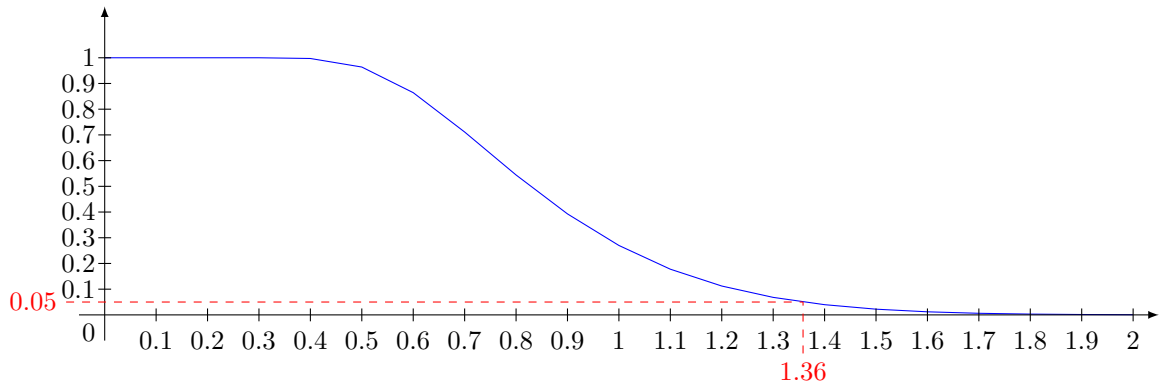



FIGURE 9.1 – Représentation de la fonction $c \mapsto \alpha(c)$ qui représente la probabilité $\mathbb{P}\left(\sup_{t\in[0;1]} |B_t| \geq c\right)$ (donc la $p$-valeur). La ligne hachée rouge permet de retrouver la valeur correspondante à une probabilité de 0.05.

**Corollaire 37 (Décision)**

À partir de la proposition 36, nous pouvons rejeter l'hypothèse nulle si la statistique $D_n$ dépasse le quantile d'ordre $1-\alpha$ de la loi présentée dans la proposition.

# Bibliographie


Les politiques d'austérité : à cause d'une erreur excel ? https://scienceetonnante.com/2020/04/17/austerite-excel/, 04 2020. Accessed : 2020-11-26.

P. Besse. Analyse en composantes principales (acp). Wikistat Toulouse. URL http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acp.

G. Biau, J. Droniou, et M. Herzlich. *Mathématiques et statistique pour les sciences de la nature.* EDP Sciences, 2010.

V. Brault. Polycopié pour l'ue statistique. *M1 Mathématiques Générales*, 2021a. URL https://www-ljk.imag.fr/membres/Vincent.Brault/Cours/CoursStatistique.pdf.

V. Brault. Polycopié des cours dispensés dans le cadre de la formation stid grenoble. *Département STID de Grenoble*, 2021b. URL https://www-ljk.imag.fr/membres/Vincent.Brault/Cours/Cours_1A.pdf.

P.-A. Cornillon, A. Guyader, F. Husson, N. Jégou, J. Josse, M. Kloareg, E. Matzner-Løber, et L. Rouvière. Statistique avec r. 2012.

D. Fourdrinier. *Statistique inférentielle : cours et exercices corrigés.* Dunod, 2002.

P. Gautret, J.-C. Lagier, P. Parola, L. Meddeb, M. Mailhe, B. Doudier, J. Courjon, V. Giordanengo, V. E. Vieira, H. T. Dupont, et al. Hydroxychloroquine and azithromycin as a treatment of covid-19 : results of an open-label non-randomized clinical trial. *International journal of antimicrobial agents*, page 105949, 2020.

J. Goetz, S. Lapoix, et H. Poulain. Data gueule, 2014. URL https://www.youtube.com/user/datagueule/.

T. Hamblin. Fake. *British medical journal (Clinical research ed.)*, 283(6307) :1671, 1981.

F. Husson et J. Pagès. *Statistiques générales pour utilisateurs : 2, Exercices et corrigés.* Presses universitaires de Rennes, 2013.

F. Husson, S. Lê, et J. Pagès. *Analyse de données avec R.* Presses universitaires de Rennes, 2016.

F. Husson, P.-A. Cornillon, A. Guyader, N. Jégou, J. Josse, N. Klutchnikoff, E. Le Pennec, E. Matzner-Løber, L. Rouvière, et B. Thieurmel. *R pour la statistique et la science des données.* Presses universitaires de Rennes, 2018.

J.-F. Le Gall. Intégration, probabilités et processus aléatoires. *Ecole Normale Supérieure de Paris*, 2006. URL https://www.imo.universite-paris-saclay.fr/~jflegall/IPPA2.pdf.

L. M. Leemis. Relationships among common univariate distributions. *The American Statistician*, 40(2) : 143–146, 1986.

H. Lehning. Cryptographie et codes secrets. l'art de cacher. 26, 2006.

H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318) :399–402, 1967.

D. Louapre. Les politiques d'austérité : à cause d'une erreur excel ? URL https://www.youtube.com/watch?v=yeX_Zs7zztY&ab_channel=ScienceEtonnante.

P. L. Micheaux, R. Drouilhet, et B. Liquet. *Le logiciel R.* 2011.

C. M. Reinhart et K. S. Rogoff. Growth in a time of debt. *American economic review*, 100(2) :573–78, 2010.

C. Robert. *Le choix bayésien : Principes et pratique*. Springer Science & Business Media, 2006.

J. P. Royston. Algorithm as 181 : the w test for normality. *Applied Statistics*, pages 176–180, 1982.

G. Saporta. *Probabilités, analyse des données et statistique*. Editions technip, 2006.

S. S. Shapiro et M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4) :591–611, 1965.

M. Tenenhaus. *Statistique : méthodes pour décrire, expliquer et prévoir*, volume 680. Dunod Paris, France :, 2007.

N. Uyttendaele, L. Maugeri, et G. Grisi. La statistique expliquée à mon chat, 2016. URL `https://www.youtube.com/channel/UCWty1tzwZW_ZNSp5GVGteaA/featured`.

B. Ycart. Histoires de mathématiques. https ://hist-math.u-ga.fr/, 2017.

# Index