Note de cours: Statistique inférentielle $\mathbb{M}_1$ ENSPD

.

.

Titre :	Statistique Inférentielle
Domaine :	Science et technique
Spécialité :	Statistique appliquée
UE :	Inférence et conditions d'application des méthodes statistiques
ECU :	Statistique Inférentielle
Semestre :	1
Nombre d'heures :	50 (dont 40 en présentiel)
Nombre de crédit :	2
Enseignant :	Dr. Koladjo François
Téléphone :	(+229)97188777
Email :	francois.koladjo at gmail.com

Description

Le cours de statistique inférentielle est un module qui initie les étudiants aux techniques d'analyse statistique pour tirer des conclusions sur une population à partir d'échantillons issus de cette population. Ce cours aborde les principes fondamentaux de l'inférence statistique, en mettant plus l'accent sur la compréhension des notions telles que les distributions d'échantillonnage, l'estimation de paramètres dans un modèle et l'aide à la prise de décision éclairée à partir des tests d'hypothèse.

Objectif général de l'ECU

L'objectif général de ce cours est d'initier les apprenants aux différentes méthodes de construction des indicateurs d'analyse statistique d'une population pour en tirer des informations utiles

Objectifs spécifiques de l'ECU

A la fin de ce cours l'apprenant doit être capable de :

- Appliquer la loi faible et la loi forte des grands nombres
- Appliquer le théorème centrale limite

- Définir un estimateur
- construire un estimateur d'un paramètre en utilisant la méthode du maximum de vraisemblance
- construire un estimateur d'un paramètre en utilisant la méthode des moments
- comparer les qualités de deux ou plusieurs estimateurs d'un même paramètre pour en choisir le meilleur
- Formuler correctement une hypothèse alternative, une hypothèse nulle et construire la règle de décision relative à un test statistique.
- Identifier une statistique adéquate pour un test d'hypothèses donné.
- reconnaître un résultat significatif à un seuil donné à l'issue d'un test statistique
- Identifier les différents types d'erreur que l'on peut commettre à l'issue d'un test statistique.
- Calculer la puissance d'un test.

Prérequis

- Mathématiques Générales (intégrales, études de fonctions, dérivations, optimisation, équations différentielles, etc.).
- Maîtriser les lois de probabilité usuelles
- Vecteurs aléatoires et fonction de variables aléatoires
- Le cours de statistiques descriptives est indispensable à la compréhension de ce cours
- Avoir suivi un cours sur les méthodes d'échantillonnage

Contenu de la formation

Chapitre 1 : Rappels sur les lois de probabilité usuelles

Chapitre 2 : Distributions d'échantillonnage

- Généralités.
- Distributions d'échantillonnage.

Chapitre 3 : Convergences des variables aléatoires

- Modes de convergence
- Lois des grands nombres
- Théorème centrale limite

Chapitre 4 : Estimation paramétrique

- Problématique et définition
- Qualités d'un estimateur
- Construction d'un estimateur ponctuel.
- Estimation par intervalle de confiance.

Chapitre 5 : Principes de base d'un test statistique

Chapitre 6 : Tests dans un modèle gaussien.

Chapitre 7 : Test sur variables qualitatives : tests du χ^2 .

- Contexte
- Exemples

Méthodes d'enseignement/apprentissage

- Cours théorique.

- Travaux Dirigés.
- Travaux Personnels d'étudiants.

Méthodes d'enseignement/apprentissage

- Exposés des chapitres par groupes d'étudiants/ jeux de rôle enseignant-apprenants
- Travaux Dirigés
- Travaux pratiques
- Travaux Personnels d'étudiants

Lieu d'apprentissage

- Salle de cours (tables, chaises, tableau).

Matériel pédagogique

- Support polycopié de cours
- Tableau et craie

Compétences visées

- Savoir construire un estimateur par la méthode des moments.
- Savoir construire un estimateur par la méthode du maximum de vraisemblance.
- Connaître les étapes de la procédure de construction d'un intervalle de confiance.
- Savoir utiliser le théorème centrale limite pour obtenir un intervalle de confiance
- Savoir identifier la procédure de test appropriée en fonction de la nature des données et de l'objectif de l'analyse.

- savoir interpréter correctement les tests statistiques.

Mode d'évaluation

- Évaluation diagnostique sous forme de question-réponse au début du cours.

- Contrôle continu : devoirs de tables (30% de la note finale) et/ou travaux pratiques (20% de la note finale)

- Examen final (50% de la note finale).

Bibliographie

- Cantoni E., Huber P., Ronchetti E. (2006). *Maîtriser l'aléatoire : Exercices résolus de probabilités et statistique*. Collection Statistique et probabilités appliquées. Paris : Springer verlag.

- Caumel Y. (2011). *Probabilités et processus stochastiques*. Paris : Springer verlag.

- Lawrence L. (2013). *Exercises and solutions in statistical theory*. Milton Park : Taylor & Francis Group.

- Gauthier C., et al. (2008). *Mathématiques Tout-en-un ECS 2e année*. Paris : Dunod.

- Gilbert Saporta (2006). *Probabilités, Analyse des données et Statistique*. Edition Technip, $2^{\text{ème}}$ édition.

Table des matières

1 Rappels sur quelques lois usuels **9**
- 1.1 Loi normale ou loi de Gauss . . . 9
- 1.2 Loi du χ^2 (Khi-deux) . . . 11
- 1.3 Loi de Student . . . 13
- 1.4 Loi de Fisher-Snedécor . . . 13

2 Distribution d'Échantillonnage **14**
- 2.1 Généralités . . . 14
  - 2.1.1 Rappels : population et échantillon . . . 14
  - 2.1.2 Avantages et inconvénients de l' échantillonnage . . . 15
- 2.2 Distribution d'échantillonnage . . . 16
  - 2.2.1 Étude de la variable aléatoire moyenne d'échantillon . . . 18
  - 2.2.2 Étude de la variance d'un échantillon . . . 19
  - 2.2.3 Distribution de la moyenne d'un échantillon . . . 21
  - 2.2.4 Distribution du S^2 . . . 21
  - 2.2.5 Distribution de la proportion . . . 24

3 Convergence de suites de variables aléatoires **26**
- 3.1 Modes de convergence . . . 26
  - 3.1.1 Convergence Presque sûre . . . 26
  - 3.1.2 Convergence en Probabilité . . . 26
  - 3.1.3 Convergence en Moyenne quadratique . . . 28
- 3.2 Loi des grands nombres . . . 28
  - 3.2.1 Convergence en Loi . . . 29
  - 3.2.2 Fonction caractéristique . . . 30
- 3.3 Théorème centrale limite . . . 33

4 Estimation paramétrique **34**
- 4.1 Estimation : problématique et définitions . . . 34
  - 4.1.1 Données . . . 34
  - 4.1.2 Modèle . . . 34

4.2 Qualités d'un estimateur . . . 35
4.3 Construction d'un estimateur ponctuel . . . 37
4.3.1 Méthode du maximum de vraisemblance . . . 37
4.3.2 Méthode des moments . . . 42
4.4 Estimation par intervalle de confiance . . . 43
4.4.1 Contexte . . . 43
4.4.2 Méthode de construction . . . 43
4.4.3 Quelques exemples de construction . . . 44

5 Principes de base d'un test statistique 48
5.1 Problématique . . . 48
5.2 Concepts et définitions . . . 49
5.3 Nature des hypothèses . . . 50
5.4 Erreur, niveau et puissance d'un test . . . 51
5.5 Construction de la règle de décision . . . 52

6 Tests dans un modèle gaussien 53
6.1 Test sur la moyenne . . . 53
6.1.1 Test de l'hypothèse $H_0 = \{\mu \leqslant \mu_0\}$ contre $H_1 = \{\mu > \mu_0\}$ . . . 54
6.2 Tests sur l'écart-type . . . 55
6.2.1 Test bilatéral sur l'écart-type . . . 55
6.2.2 Test de l'hypothèse $H_0 = \{\sigma \leqslant \sigma_0\}$ contre $H_1 = \{\sigma > \sigma_0\}$ . . . 56
6.3 Tests sur une proportion . . . 56
6.3.1 Test bilatéral sur une proportion . . . 56
6.3.2 Test unilatéral sur une proportion . . . 56
6.4 Exemple de calcul de puissance . . . 57
6.5 Comparaison de populations . . . 57
6.5.1 Contexte et exemples . . . 57
6.5.2 Test de comparaison de deux moyennes . . . 59
6.5.3 Test sur données appariées . . . 61
6.5.4 Test de $H_0 = \{\sigma_1 = \sigma_2\}$ contre $H_1 = \{\sigma_1 \neq \sigma_2\}$ . . . 61
6.5.5 Test de comparaison de deux proportions . . . 62

7 Tests sur variables qualitatives : test du χ^2 63
7.1 Contexte . . . 63
7.2 Exemples de test du χ^2 . . . 63
7.2.1 Adéquation à une loi multinomiale p_0 . . . 63
7.2.2 Test de χ^2 d'indépendance de deux variables catégorielles (ou qualitatives) . . . 64

Chapitre 1

Rappels sur quelques lois usuels

1.1 Loi normale ou loi de Gauss

Une variable aléatoire réelle (v.a.r) suit une loi normale d'espérance μ et d'écart-type σ (strictement positif) si elle admet pour densité de probabilité f avec

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

pour tout réel x.

- Une loi normale sera notée $\mathcal{N}(\mu, \sigma^2)$ car elle dépend de 2 paramètre μ et σ. Donc si $X \sim \mathcal{N}(\mu, \sigma^2)$ on a :

$$\mathbb{E}(X) = \mu \quad et \quad \mathbb{V}ar(X) = \sigma^2.$$

- La variable aléatoire $Z = \dfrac{X-\mu}{\sigma}$ suit la loi $\mathcal{N}(0,1)$. On dit que Z suit la loi centrée réduite.

- On note ϕ la fonction de répartition de la loi normale centrée réduite de densité :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Propriété 1.1.1. .

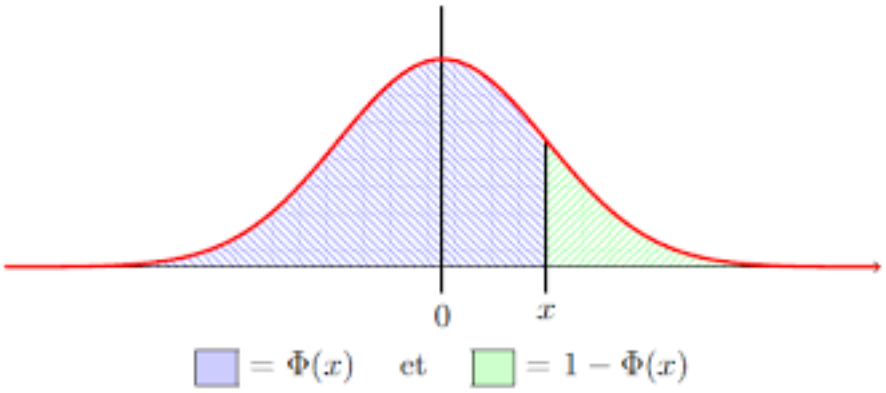

FIGURE 1.1 – Densité de probabilité de la loi normale centrée réduite

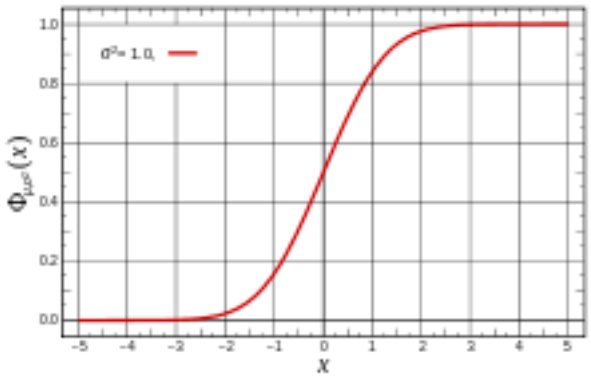

FIGURE 1.2 – Fonction de répartition de la loi normale centrée réduite

- $\phi(-z) = 1 - \phi(z)$
- *Si* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, *et si* X_1 *et* X_2 *sont indépendantes, alors* $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- *Si* $X_1, \cdots, X_n$ *suit la loi* $\mathcal{N}(\mu, \sigma^2)$ *et sont indépendantes alors* $X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$.
- *Si* $X \sim \mathcal{N}(\mu, \sigma^2)$ *et* $Y = aX + b$ $(a > 0)$ *alors* $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- *On notera* $z_{1-\frac{\alpha}{2}}$, *le nombre tel que* $\mathbb{P}(Z > z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$ *lorsque* $Z \sim \mathcal{N}(0, 1)$.

Exercice 1.1.1. .

Déterminer $\mathbb{P}(Z < z_{1-\frac{\alpha}{2}})$, $\mathbb{P}(Z < -z_{1-\frac{\alpha}{2}})$, $\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}})$, $\mathbb{P}(|\ Z\ | > z_{1-\frac{\alpha}{2}})$ *en fonction de* α.

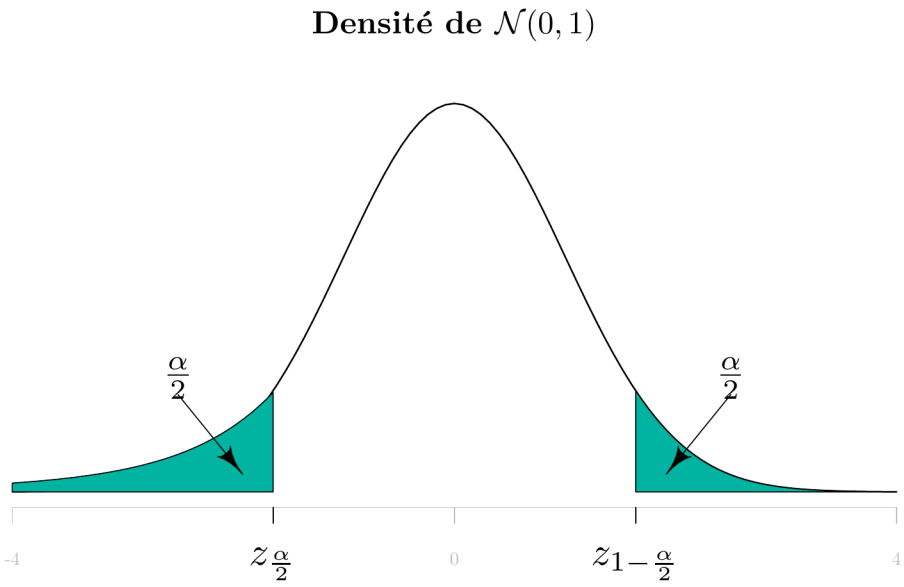

FIGURE 1.3 – Quantile d'ordre $1 - \dfrac{\alpha}{2}$ de la loi normale centrée réduite

1.2 Loi du χ^2 (Khi-deux)

Définitions 1.2.1. . *Soit* $Z_1, \cdots, Z_\nu$ *une suite de variables aléatoires indépendantes de même loi* $\mathcal{N}(0,1)$. *Alors la variable aléatoire* $\sum_{i=1}^{\nu} Z_i^2$ *suit une loi de Khi-deux à* ν *degrés de liberté. On note* $\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2(\nu)$.

Propriété 1.2.1. .

a. Sa fonction caractéristique est $\varphi(t) = (1 - 2it)^{-\frac{\nu}{2}}$.

b. Sa densité de probabilité est :

$$f_\nu(x) = \begin{cases} \dfrac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}} & si\ x > 0 \\ 0 \quad sinon \end{cases}$$

$\Gamma(\cdot)$ *est la fonction Gamma d'Euler définir par :*

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$$

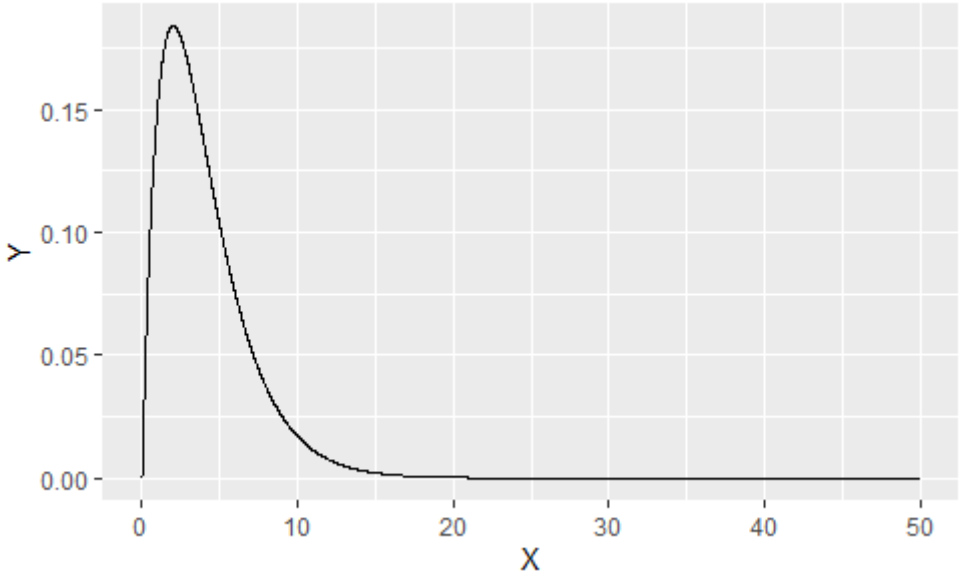

FIGURE 1.4 – Densité de probabilité de la loi χ^2 à 4 ddl.

c. Si $Q \sim \chi^2(\nu)$*, alors* $\mathbb{E}(Q) = \nu$ *et* $\mathbb{V}ar(Q) = 2\nu$*.*

d. Si $Q_1 \sim \chi^2(\nu_1)$ *et* $Q_2 \sim \chi^2(\nu_2)$ *et si* Q_1 *et* Q_2 *sont indépendantes, alors* $Q_1 + Q_2 \sim \chi^2(\nu_1 + \nu_2)$

1.3 Loi de Student

Définitions 1.3.1. .

Soient Z et Q deux variables aléatoires indépendantes telles que $Z \sim \mathcal{N}(0,1)$ et $Q \sim \chi^2(\nu)$. Alors la variable aléatoire $T = \dfrac{Z}{\sqrt{\dfrac{Q}{\nu}}}$ suit la loi de Student à ν degrés de liberté.

Propriété 1.3.1. .

La loi de Student converge en loi vers la loi centrée réduite.

1.4 Loi de Fisher-Snedécor

Définitions 1.4.1. .

Soient Q_1 et Q_2 deux variables aléatoires indépendantes telles que $Q_1 \sim \chi^2(\nu_1)$ et $Q_2 \sim \chi^2(\nu_2)$. Alors, la variable aléatoire $F = \dfrac{\dfrac{Q_1}{\nu_1}}{\dfrac{Q_2}{\nu_2}}$ suit la loi de Fisher-Snedécor à (ν_1, ν_2) degrés de liberté notée $F(\nu_1, \nu_2)$.

Propriété 1.4.1. .

1. *Si F suit une loi de Fisher $F(\nu_1, \nu_2)$, alors $\dfrac{1}{F}$ suit une loi de Fisher $F(\nu_2, \nu_1)$.*

2. *Si T suit une loi de Student à ν degrés de liberté, alors T^2 suit une loi de Fisher $F(1, \nu)$.*

Chapitre 2

Distribution d'Échantillonnage

2.1 Généralités

La procédure d'échantillonnage constitue une étape nécessaire dans toute étude statistique. C'est la façon la plus courante de récolter les données car il est souvent impossible de mesurer une variable sur tous les objets ou individus d'une population.

2.1.1 Rappels : population et échantillon

On suppose l'existence d'éléments appelés unités. Cette notion d'unités est facilement perceptibles dans une population animale ou dans l'industrie si on on s'intéresse à un ensemble de produits (ordinateurs, bic, chaussures, etc).
La définition d'unités devient plus délicate quand il s'agit d'un liquide ou d'un matériel en vrac. Dans le cas des liquides par exemples, on peut considérer les unités physiques comme le mm^3.

- On appelle population, l'ensemble des unités sur lesquels porte l'étude. Avant toute procédure d'échantillonnage, il est recommandé de se demander "quelle est la population étudiée ?".

- Un échantillon est un sous-ensemble d'unités de la population étudiée. Qu'il traite un échantillon ou une population, le statisticien décrit ces ensemble à l'aide de mesures telles que le nombre d'unités, la moyenne, l'écart-type et le pourcentage.

- On appelle base de sondage, la liste des unités de la population. Cette liste est

théoriquement nécessaire pour procéder à des tirages au sort corrects mais on n'en dispose pas toujours en pratique : elle peut être fausse ou incomplète.

- La taille de la population N est le nombre d'unités dans la population. La taille de l'échantillon n est le nombre d'unités de ce dernier. Le taux de sondage ou fraction sondée est le rapport $f_o = \frac{n}{N}$.

- Les mesures utilisées pour décrire une population sont des paramètres. Chaque paramètre désigne une caractéristique de la population.

- Pour décrire un échantillon, on utilise les statistiques. Une statistique est donc une caractéristique de l'échantillon.

2.1.2 Avantages et inconvénients de l' échantillonnage

On distingue plusieurs types d'échantillonnage

- Échantillonnage sur la base du jugement (certaines communes peuvent être représentative de la population en terme d'opinion).

- Échantillonnage aléatoire simple : On appelle échantillonnage aléatoire simple, un échantillonnage obtenu par une méthode qui assure à chaque échantillon de même taille, la même probabilité d'être tiré. On utilise un générateur de nombres aléatoires pour s'assurer que le choix des unités dans la population s'effectue vraiment au hasard.

- Lorsqu'on dispose d'une connaissance à priori sur la population étudiée, par exemple, si elle n'est pas homogène mais est divisée en sous-groupes d'effectifs connus (appelés strates), ayant des caractéristiques assez différentes, il est plus intéressant d'utiliser une procédure d'échantillonnage stratifié qui consiste à prendre un échantillon aléatoire simple dans chaque strate.

Avantages

- Coût moindre.

- Gain de temps.

Inconvénients

L'échantillonnage a pour but de fournir suffisamment d'informations sur les caractéristiques d'une population. Mais en général, les résultats obtenus diffèrent d'un échantillon à l'autre et diffèrent surtout de la caractéristique correspondante dans la population. On dit qu'il y a des fluctuations d'échantillonnage.Comment peut-on dans ce cas tirer des conclusions valables ?

Les lois de probabilité de ces fluctuations permettent d'apprendre plus sur les correspondances entre les statistiques (issues d'échantillon) et les paramètres (caractéristique de la population).

2.2 Distribution d'échantillonnage

Exemple Introductif

Dans une classe constituée de 5 étudiants en statistique, un professeur s'intéresse au temps hebdomadaire consacré à l'étude dans sa matière par chaque étudiant. On a obtenu les résultats suivants :

Étudiant	**Temps d'étude (en heure)**
1	8
2	3
3	5
4	10
5	4
Total	30

La moyenne de la population est $\mu = \dfrac{30}{5} = 6$ heures. Pour évaluer le professeur et la classe, un inspecteur doit choisir trois étudiants et faire la moyenne de leur temps d'étude.

a. Combien a-t-il de possibilités pour faire ce choix ?

b. Quels sont les valeurs possibles de cette moyenne ?

c. Quelle relation existe-il entre la moyenne des échantillons et la moyenne $\mu = 6$ de la population ?

Solution

1 Il y a au total $C_5^3 = 10$ échantillon possibles (i.e 10 possibilités).

2 Les valeurs possibles de $\bar{X}$:

$$\bar{X} \in \{5,33; 7; 5; 7,67; 5,67; 7,33; 6; 4; 6,33\}$$

.

Numéro de l'échantillon	**Échantillon**	**Valeurs du temps d'étude de cet échantillon**	**Moyenne $\bar{X}$ de l'échantillon**
1	1 ;2 ;3	{8 ;3 ;5}	5,33
2	1 ;2 ;4	{8 ;3 ;10}	7,00
3	1 ;2 ;5	{8 ;3 ;4}	5,00
4	1 ;3 ;4	{8 ;5 ;10}	7,67
5	1 ;3 ;5	{8 ;5 ;4}	5,67
6	1 ;4 ;5	{8 ;10 ;4}	7,33
7	2 ;3 ;4	{3 ;5 ;10}	6,00
8	2 ;3 ;5	{3 ;5 ;4}	4,00
9	2 ;4 ;5	{3 ;10 ;4}	5,67
10	3 ;4 ;5	{5 ;10 ;4}	6,33
Total			60

Remarque 2.2.1. .

- *La moyenne des échantillons varie entre 4 et 7,67 (étendue=7,67-4=3,67). La distribution de la moyenne des échantillons est donc moins étalée que celle de la population (étendue=10-3=7).*

- *Il est possible que deux échantillons différents aient la même moyenne (c'est le cas des échantillons 5 et 9).*

3 La moyenne des moyennes d'échantillons est $\mathbb{E}(\bar{X}) = \dfrac{60}{10} = 6$; on constate qu'elle est égale à la moyenne de la population $\mu = 6$. C'est une propriété générale qu'on va montrer un peu plus tard.

Exercice 2.2.1. .

Répondre aux questions 2) et 3) pour l'écart-type σ_o.

On définit sur chaque échantillon, les variables aléatoires $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_{\alpha_i}$ et $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{\alpha_i} - \bar{X})^2$

Ces variables sont aléatoires parce que l'unité prise dans l'échantillon est aléatoire et non pas parce que les mesures X_{α_i} le sont.

2.2.1 Étude de la variable aléatoire moyenne d'échantillon

La situation de l'exemple introductif est idéale car on a accès à tous les échantillons possibles. En pratique, on observe qu'un seul échantillon qui doit servir à caractériser les paramètres de la population.

La moyenne d'un échantillon, c'est-à-dire la valeur prise par la variable aléatoire $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ (on fait abstraction des indices α_i pour simplifier les notations).

On se place ici dans le cas d'une population infinie correspondant à une situation de tirage avec remise des unités dans une population. Cela permet de supposer que les composantes de l'échantillon sont indépendantes et équidistribuées et facilite la preuve des propriétés.

Propriété 2.2.1. .

$$\mathbb{E}(\bar{X}) = \mu \quad et \quad \mathbb{V}ar(\bar{X}) = \frac{\sigma^2}{n}$$

Preuve

- La linéarité de l'espérance permet d'écrire

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\ &= \frac{n \times \mu}{n} \ (car\ les\ variables\ X_i\ ont\ même\ espérance\ \mu) \\ &= \mu\end{aligned}$$

•

$$\begin{aligned}\mathbb{V}ar(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar(X_i)\ (car\ les\ variables\ X_i\ sont\ indépendantes)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_0^2\ (elles\ ont\ même\ variance)\\ &= \frac{\sigma_0^2}{n}\end{aligned}$$

Remarque 2.2.2. .

- *La moyenne de la distribution d'échantillonnage de la moyenne est égale à la moyenne de la population.*

- *Plus n augmente, plus* $\mathbb{V}ar(\bar{X})$ *diminue* $\Leftrightarrow$ *échantillon de taille élevée* $\Rightarrow$ *estimation plus précise de la moyenne.*

Complément

Pour une population de taille finie, on a :

$$\mathbb{V}ar(\bar{X}) = \frac{\sigma^2}{n}(1 - f_0)$$

$1 - f_0$ s'appelle le facteur de correction pour une population finie.

2.2.2 Étude de la variance d'un échantillon

* S_0^2

La variance d'un échantillon est une réalisation de la variable aléatoire $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Possède-t-elle la même propriété que la moyenne de l'échantillon ?

Propriété 2.2.2. . *(espérance de* S_0^2 *)*

$$\mathbb{E}(S_0^2) = \frac{n-1}{n}\sigma_0^2$$

Preuve

$$\begin{aligned} S_0^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - m + m - \bar{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - m)(m - \bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(m - \bar{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (m - \bar{X})^2 \end{aligned}$$

donc

$$\begin{aligned} \mathbb{E}(S_0^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - m)^2 - \mathbb{E}(\bar{X} - m)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma_0^2 - \mathbb{V}ar(\bar{X}) \\ &= \sigma_0^2 - \frac{\sigma_0^2}{n} \\ \mathbb{E}(S_0^2) &= \frac{n-1}{n}\sigma_0^2 \end{aligned}$$

Remarque 2.2.3. .

La moyenne des variances d'échantillon n'est pas égale à ma variance de la population. Mais si n est grand, les deux sont très proches.

.

$*S^2$

Par définition, $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \dfrac{nS_0^2}{n-1}$.

On a donc

$$\begin{aligned} \mathbb{E}(S^2) &= \frac{n}{n-1}\mathbb{E}(S_0^2) \\ &= \frac{n}{n-1}\,\frac{n-1}{n}\,\sigma_0^2 \\ &= \sigma_0^2 \end{aligned}$$

On dit que le S^2 est un estimateur sans biais de σ_0^2.

2.2.3 Distribution de la moyenne d'un échantillon

Selon que l'on dispose d'un échantillon de grande taille ($n \geqslant 30$) ou de petite taille ($n < 30$), les résultats peuvent être très différents. On rappelle que les composantes de l'échantillon constituent une suite de variables aléatoires $X_1, \cdots, X_n$ indépendantes et de variance commune σ_0^2.

Cas d'échantillons de grandes tailles ($n \geqslant 30$)
On se trouve

a. En présence de n variables aléatoires indépendantes.

b. Elle suivent la même loi d'espérance μ et de variance σ_0^2.

On peut appliquer le théorème central limite :
La somme $X_1 + \cdots + X_n$ suit approximativement la loi normale $\mathcal{N}(n\mu, n\sigma_0^2)$ d'espérance $n\mu$ et de variance $n\sigma_0^2$. Donc la moyenne

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma_0^2}{n}) \quad ou \quad \frac{\bar{X} - \mu}{\frac{\sigma_0}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

2.2.4 Distribution du S^2

On suppose que chaque variable X_i suit une loi normale de variance σ_0^2. On pose :

$$\begin{aligned} Y &= \frac{nS_0^2}{\sigma_0^2} \\ &= \frac{(n-1)S^2}{\sigma_0^2} \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma_0} \right)^2 \end{aligned}$$

Y est la somme des écarts réduits de variables suivant une loi normale.

Propriété 2.2.3. .

$$Y = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

χ^2_{n-1} est la loi du χ^2 à $n-1$ degré de liberté (ddl).

Cas d'échantillons de grandes tailles $(n \geqslant 30)$

Dans le cas des grands échantillons, la loi du χ^2_n peut être approchée par la loi normale $\mathcal{N}(n, 2n)$. Donc $Y = \dfrac{(n-1)S^2}{\sigma_0^2}$ suit approximativement une loi normale avec $\mathbb{E}(Y) \approx n-1$ et $\mathbb{V}ar(Y) \approx 2(n-1)$.

Propriété 2.2.4. .

Si $n \geqslant 30$, $S^2 \approx \mathcal{N}(\sigma_0^2, \dfrac{2\sigma_0^4}{n-1})$.

Preuve

On sait $Y = \dfrac{(n-1)S^2}{\sigma_0^2}$. Donc $S^2 = \dfrac{\sigma_0^2}{n-1}Y$.

Comme $Y \approx \mathcal{N}(n-1, 2(n-1))$, $\mathbb{E}(S^2) = \sigma_0^2$.

$\mathbb{V}ar(S^2) = \dfrac{\sigma_0^4}{(n-1)^2}\mathbb{V}ar(Y) = \dfrac{2(n-1)\sigma_0^4}{(n-1)^2} = \dfrac{2\sigma_0^4}{n-1}$

Donc $S^2 \approx \mathcal{N}(\sigma_0^2, \dfrac{2\sigma_0^4}{n-1})$

Propriété 2.2.5. .

Si $n \geqslant 30$, $\bar{X}$ *suit approximativement la loi* $\mathcal{N}(\mu, \dfrac{\sigma_0^2}{n})$. *Dans la pratique, la variance* σ_0^2 *est approchée par la statistique* S^2 *qui en est un estimateur sans biais.*

Remarque 2.2.4. .

La variance de l'échantillon $S_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *est aussi une approximation de* σ_0^2, *mais elle est biaisée. Les deux sont liées par :*

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

Cas de petits échantillons $(n < 30)$

Lorsqu'on dispose d'échantillon de petites tailles, il est difficile, voire impossible de faire des approximations. Mais on peut arriver à des résultats tout à fait cohérent dans des cas particuliers. On suppose ici que les variables $X_1, \cdots, X_n$ suivent une loi normale de moyenne μ et de variance σ_0^2. Deux cas de figures peuvent se présenter :

a. Cas où σ_0^2 est connu

Chaque variable X_i suit la loi $\mathcal{N}(\mu, \sigma_0^2)$.

$X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma_0^2)$ car les variables sont indépendantes. Donc

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma_0^2}{n})$ ou $\dfrac{\bar{X}-\mu}{\frac{\sigma_0}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

$$\begin{cases} n < 30 \\ \sigma_0 \text{ connue} \end{cases} \Rightarrow \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma_0^2}{n})$$

b. Cas où σ_0^2 est inconnu

Chaque variable X_i suit une loi normale de variance σ_0^2 inconnue. L'approximation de σ_0^2 par S^2 n'est plus fiable car S^2 varie beaucoup d'un échantillon à un autre. Il faut donc tenir compte de la loi régissant les variations de S^2. On pose

$$T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma_0}}{\frac{S}{\sigma_0}}$$

On sait que $\sqrt{n}\dfrac{(\bar{X}-\mu)}{\sigma_0} = \dfrac{(\bar{X}-\mu)}{\frac{\sigma_0}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

Aussi $\dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$.

Donc $T = \dfrac{\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma_0}}{\sqrt{\frac{S^2}{\sigma_0^2}}} \sim T_{n-1}$

Théorème 2.2.1. .

$$\begin{cases} n < 30 \\ \sigma_0 \text{ inconnue} \end{cases} \Rightarrow T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \text{ suit la loi de student à } n-1 \text{ ddl notée } T_{n-1}$$

Exercice 2.2.2. .

Pour moderniser son département, le ministre de la fonction publique décide de quantifier la valeur de travail de ses administrés en terme de nombre de dossiers traités par semaine. Selon le code du travail, un travailleur doit traiter en moyenne 150 dossiers avec une variance de 100 en situation normale. Une entreprise en est situation régulière si le nombre moyen de dossiers traités par ses travailleurs est compris entre 147 et 153. Le ministre fait une descente inopinée dans une entreprise de 25 salariés.

•

a. *Quelle est la probabilité que cette entreprise soit en situation régulière ?*

b. *Même question qu'en a) si la variance n'étant pas connue, est estimée à 81 à partir de l'échantillon.*

Solution

X_i = nombre de dossiers traités par l'individu i de cette entreprise, $i = 1, \cdots, 25$, $n = 25$
$m = 150, \quad \sigma_0^2 = 100 \rightarrow$ Paramètres de la population.
$X \sim \mathcal{N}(m, \sigma_0^2)$.

a. L'entreprise est en situation régulière si $147 \leqslant \bar{X} \leqslant 153$ avec $\bar{X}$ le nombre moyen de dossiers par individus et par semaine dans cette entreprise.
On cherche $\mathbb{P}(147 \leqslant \bar{X} \leqslant 153)$
On sait que $\dfrac{\bar{X} - m}{\dfrac{\sigma_0}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.
Et

$$\begin{aligned}\mathbb{P}(147 \leqslant \bar{X} \leqslant 153) &= \mathbb{P}\left(\frac{147-150}{\frac{10}{5}} \leqslant \frac{\bar{X}-m}{\frac{\sigma_0}{\sqrt{n}}} \leqslant \frac{\bar{X}-m}{\frac{10}{5}}\right) \\ &= \mathbb{P}\left(-\frac{3}{2} \leqslant \mathcal{N}(0,1) \leqslant \frac{3}{2}\right) \\ &= 0,866.\end{aligned}$$

b. Si σ_0 est inconnu.
$T = \dfrac{\bar{X} - m}{\dfrac{S}{\sqrt{n}}} \sim T_{n-1}$ donc $\mathbb{P}(147 \leqslant \bar{X} \leqslant 153) = \mathbb{P}(\dfrac{-3}{\frac{9}{5}} \leqslant T_{n-1} \leqslant \dfrac{3}{\frac{9}{5}})$
$\mathbb{P}(147 \leqslant \bar{X} \leqslant 153) = \mathbb{P}(-\dfrac{5}{3} \leqslant T_{n-1} \leqslant \dfrac{5}{3}) = 0,8915$

2.2.5 Distribution de la proportion

.

Lorsque le caractère étudié ne prend que deux valeurs, la moyenne représente la proportion d'une modalité. Soit X le nombre de fois où la modalité apparait dans un échantillon de taille n.
$X \sim \mathcal{B}(n, p)$ où p représente la proportion dans la population. On pose $F = \dfrac{X}{n}$.

On sait que $\mathbb{E}(X) = np$ et $\mathbb{V}ar(X) = np(1-p)$.
Donc $\mathbb{E}(F) = p$ et $\mathbb{V}ar(F) = \dfrac{p(1-p)}{n}$.
Quand n tend vers l'infini, on a F converge en loi vers $\mathcal{N}(p, \dfrac{p(1-p)}{n})$. Donc

$$P = \frac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$$

Exercice 2.2.3. .

Selon une étude sur le comportement du consommateur, 30% d'entre eux sont influencés par la marque lors de l'achat d'un téléphone portable. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité qu'au moins 40 d'entre eux se déclarent influencés par la marque ?

Solution

Soit F la variable proportion de ceux qui se déclarent influencés. On cherche $\mathbb{P}(F > 0,4)$.
$p = 0,30$ donc $np = 100 \times 0,3 = 30 > 15$ et $n(1-p) = 100 \times 0,7 = 70 > 15$. On a :

$$F \approx \mathcal{N}(p, \frac{p(1-p)}{n})$$

Donc $P = \dfrac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$
Donc

$$\begin{aligned}
\mathbb{P}(F > 0,4) &= \mathbb{P}\left(\frac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} > \frac{0,4-0,3}{\sqrt{\dfrac{0,3\times 0,7}{100}}}\right) \\
&= \mathbb{P}(\mathcal{N}(0,1) > 2,18) \\
&= 0,01462
\end{aligned}$$

Un peu plus d'une chance sur 100 que 40 d'entre eux se déclarent influencés par la marque.

Chapitre 3

Convergence de suites de variables aléatoires

Les convergences de suites de variables aléatoires sont des outils permettant d'étudier entre autres les propriétés d' estimateurs des paramètres dans un modèle statistique. Ces outils sont également utilisés dans la construction des certaines procédures de test statistique et peuvent servir de critères de choix dans la comparaisons de deux ou plusieurs estimateurs. Il existent différentes formes de convergence de suites de variables aléatoires dont les plus fréquemment rencontrées sont abordées ci-dessous.

3.1 Modes de convergence

3.1.1 Convergence Presque sûre

Définitions 3.1.1. .

Soit $(X_n)_{n\geqslant 1}$ et X des variables aléatoires définies sur un même espace de probabilité $(\Omega, \mathbb{P})$. On dit que $(X_n)_n$ converge presque surement vers X si on a :

$$\mathbb{P}(X_n \to X) = \mathbb{P}(\omega :\ X_n(\omega) \to X(\omega)) = 1$$

3.1.2 Convergence en Probabilité

Définitions 3.1.2. .

Soit $(X_n)_{n\geqslant 1}$ et X des variables aléatoires réelles définies sur un même espace de probabilité $(\Omega, \mathbb{P})$. On dit que :

- *$(X_n)_{n\geqslant 1}$ converge en probabilité vers une constante ℓ si :*

$$\forall \varepsilon > 0, \lim_{n\to+\infty} \mathbb{P}(|X_n - \ell| > \varepsilon) = 0$$

- $(X_n)_{n\geqslant 1}$ *converge en probabilité vers* X *si :*

$$\forall \varepsilon > 0, \lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

Exemple 3.1.1. *On réalise une expérience contenant n épreuves de Bernoulli de probabilité de succès p. Soit X la variable aléatoire égal au nombre de succès dans cette expérience. On sait que X suit une loi binomiale de paramètres n et p :* $X \sim \mathcal{B}(n,p)$.
Montrer que la proportion de succès converge en probabilité vers p.

Solution

On sait que $\mathbb{E}(X) = np$ et $\mathbb{V}ar(X) = np(1-p)$. Donc $\mathbb{E}(\frac{X}{n}) = n$ et $\mathbb{V}ar(\frac{X}{n}) = \frac{p(1-p)}{n}$. Comme la fonction $p \mapsto p(1-p)$ atteint son maximum au point $p = \frac{1}{2}$, on a $p(1-p) \leqslant \frac{1}{4}$.

D'après l'inégalité de Biénaymé-Chebyshev, on a : $\mathbb{P}\left(\left|\frac{X}{n} - \mathbb{E}\left(\frac{X}{n}\right)\right| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{V}ar(\frac{X}{n})}{\varepsilon^2}$ pour tout $\varepsilon > 0$.
Donc $\forall\, \varepsilon > 0$, on a :

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}$$

et $\lim_{n\to+\infty} \frac{1}{4n\varepsilon^2} = 0$. Donc $\forall\, \varepsilon > 0$, $\lim_{n\to+\infty} \mathbb{P}\left(\left|\frac{X}{n} - p\right| \geqslant \varepsilon\right) = 0$

Propriété 3.1.1. .
Soit $(X_n)_{n\geqslant 1}$ *une suite de variables aléatoires réelles définies sur un même espace de probabilité* $(\Omega, \mathbb{P})$ *et vérifiant :*

$$\lim_{n\to+\infty} \mathbb{E}(X_n) = \ell \quad et \quad \lim_{n\to+\infty} \mathbb{V}ar(X_n) = 0$$

Alors $(X_n)_n$ *converge en probabilité vers* ℓ.

Preuve (un peu technique)
Soit $\varepsilon > 0$ et $v_n = \mathbb{E}(X_n) - \ell$. On a : $\lim v_n = 0$. Donc

$$\exists n_0 \in \mathbb{N} \quad tel\ que \quad n > n_0 \Rightarrow |v_n| < \frac{\varepsilon}{2}$$

Aussi $|X_n - \ell| = |X_n - \mathbb{E}(X_n) + \mathbb{E}(X_n) - \ell| \leqslant |X_n - \mathbb{E}(X_n)| + |\mathbb{E}(X_n) - \ell|$
donc pour $n > n_0$, $|X_n - \mathbb{E}(X_n)| < \frac{\varepsilon}{2} \Rightarrow |X_n - \ell| < \varepsilon$

donc $|X_n - \ell| \geqslant \varepsilon \Rightarrow |X_n - \mathbb{E}(X_n)| \geqslant \dfrac{\varepsilon}{2}$

D'où $\mathbb{P}(|X_n - \ell| \geqslant \varepsilon) \leqslant \mathbb{P}(|X_n - \mathbb{E}(X_n)| \geqslant \dfrac{\varepsilon}{2}) \leqslant \dfrac{\mathbb{V}ar(X_n)}{\left(\dfrac{\varepsilon}{2}\right)^2}$ (D'après Bienaymé-Chebychev).

Puisque $\lim \mathbb{V}ar(X_n) = 0$, on a : $\lim\limits_{n\to+\infty} \mathbb{P}(|X_n - \ell| \geqslant \varepsilon) = 0$.

3.1.3 Convergence en Moyenne quadratique

Définitions 3.1.3. .

Une suite de v.a.r $(X_n)_{n\geqslant 1}$ converge en moyenne quadratique vers une v.a.r X si :

$$\lim_{n\to+\infty} \mathbb{E}[(X_n - X)^2] = 0.$$

Propriété 3.1.2. .

1. *Si $(X_n)_n$ converge en moyenne quadratique, alors elle converge en probabilité.*

2. *Si $(X_n)_n$ est telle que $\mathbb{E}(X_n) < +\infty$, $\mathbb{V}ar(X_n) < +\infty \ \forall n$ et $\mathbb{E}(X_n) \to \mu$, $\mathbb{V}ar(X_n) \to 0$, alors X_n converge en moyenne quadratique vers μ.*

Preuve

1. Supposons que $(X_n)_n$ converge en moyenne quadratique. On a d'après l'inégalité de Markov.
   $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^2 > \varepsilon^2) \leqslant \dfrac{\mathbb{E}(|X_n - X|^2)}{\varepsilon^2}$ pour tout $\varepsilon > 0$.
   Et $\lim\limits_{n\to+\infty} \mathbb{E}(|X_n - X|^2) = 0$. Donc $\lim\limits_{n\to+\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$.

2. Sous les hypothèses de la propriété, on a :
   $\mathbb{E}((X_n - \mu)^2)) = \mathbb{E}(X_n^2) - 2\mu\mathbb{E}(X_n) + \mu^2$.
   Comme $\lim\limits_{n\to+\infty} (\mu^2 - 2\mu\mathbb{E}(X_n)) = -\mu^2 = \lim\limits_{n\to+\infty} -(\mathbb{E}(X_n))^2$. Donc $\lim\limits_{n\to+\infty} \mathbb{E}((X_n - \mu)^2)) = \lim\limits_{n\to+\infty} \mathbb{E}(X_n^2) - (\mathbb{E}(X_n))^2 = \lim\limits_{x\to+\infty} \mathbb{V}ar(X_n) = 0$. D'où le résultat.

3.2 Loi des grands nombres

Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles indépendantes définies sur un même espace de probabilité $(\Omega, \mathbb{P})$.

Les lois des grands nombres portent sur le comportement de la moyenne empirique $\overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$.

Proposition 3.2.1. . *(Loi Faible)*

On suppose que les v.a $X_i, i = 1, \cdots, n$ sont indépendantes, ayant même espérance μ et de variance finie vérifiant $\lim_{n\to+\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}ar(X_i) = 0$. Alors, la moyenne empirique $\overline{X}$ converge en probabilité vers la moyenne commune μ.

Preuve

On sait que $\mathbb{E}(\overline{X}) = \mu$ et $\mathbb{V}ar(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}ar(X_i)$. Donc $\lim_{n\to+\infty} \mathbb{V}ar(\overline{X}) = 0$.
Donc $\overline{X}$ converge en probabilité vers μ.
La proposition suivante énonce un résultat plus fort.

Proposition 3.2.2. . *(Loi forte)*

La moyenne empirique $\overline{X}$ d'une suite $(X_n)_{n\geqslant 1}$ de v.a.r iid et d'espérance finie converge presque sûrement vers l'espérance commune ($\overline{X} \xrightarrow{p.s} \mathbb{E}(X_1)$).

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}(X_1)\right) = 1$$

.

.

3.2.1 Convergence en Loi

Définitions 3.2.1. .

Soit $(X_n)_{n\geqslant 1}$ et X des variables aléatoires réelles indépendantes définies sur un même espace de probabilité $(\Omega, \mathbb{P})$, de fonction de répartition respective $\mathbb{F}_n$ et $\mathbb{F}$. On dit que $(X_n)_n$ converge en loi vers X (on note $X_n \xrightarrow{\mathcal{L}} X$) si en tout point x où $\mathbb{F}$ est continue, $\mathbb{F}_n(x)$ converge vers $\mathbb{F}(x) = \mathbb{P}(X \leqslant x)$.

Exemple 3.2.1. .

Soit $X_1, \cdots, X_n$ une suite de v.a.r iid de loi $U[0, 1]$. On définit $M_n = max\{X_1, \cdots, X_n\}$ et $Q_n = n(1 - M_n)$.

a. Déterminer la fonction de répartition de M_n et de Q_n.

b. En déduire la loi limite de Q_n.

.

Solution

a. $\mathbb{F}(x) = \mathbb{P}(X \leqslant x) = \begin{cases} x \ si \ x \in [0,1] \\ 1 \ si \ x \geqslant 1 \end{cases}$.

$$\begin{aligned} \mathbb{F}_{M_n}(x) &= \mathbb{P}(M_n \leqslant x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x) \quad car\ [M_n \leqslant x] \Leftrightarrow X_i \leqslant x\ \forall i = 1, \cdots, n) \\ &= x^n \quad \forall x \in [0,1]. \end{aligned}$$

$$\begin{aligned} \mathbb{F}_{Q_n}(x) &= \mathbb{P}(Q_n \leqslant x) \\ &= \mathbb{P}(n(1 - M_n) \leqslant x) \\ &= \mathbb{P}(M_n \geqslant 1 - \frac{x}{n}) \\ &= 1 - \mathbb{F}_{M_n}(1 - \frac{x}{n}) \\ &= 1 - (1 - \frac{x}{n})^n \quad \forall x \geqslant 0. \end{aligned}$$

b. Comme $\lim\limits_{n \to +\infty} (1 - \frac{x}{n})^n = e^{-x}$ donc $\lim\limits_{n \to +\infty} \mathbb{F}_{Q_n}(x) = 1 - e^{-x}$. Donc $Q_n \xrightarrow{\mathcal{L}} Q \sim \varepsilon(1)$ loi exponentielle de paramètre 1.

NB : La loi exponentielle de paramètre $\lambda > 0$ a pour densité $Q_\lambda(x) = \lambda e^{-\lambda x}$ pour tout $x \geqslant 0$ et $Q_\lambda(x) = 0$ si $x < 0$.

Exercice 3.2.1. .
Déterminer son espérance, sa variance et sa fonction de répartition.

Propriété 3.2.1. .

- $X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{\mathcal{L}} X$.
- $X_n \xrightarrow{\mathcal{L}} X \Rightarrow \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *pour tout fonction* f *continue bornée.*
- $X_n \xrightarrow{\mathcal{L}} c \Rightarrow X_n \xrightarrow{\mathbb{P}} c$ *(c une constante réelle).*

3.2.2 Fonction caractéristique

Définitions 3.2.2. .
Soit X une v.a.r. On appelle fonction caractéristique de X, la fonction :

$$\begin{array}{rccl} \phi_X : \mathbb{R} & \to & \mathbb{C} \\ t & \mapsto & \phi_X(t) = \mathbb{E}(e^{itx}) \end{array}$$

Remarque 3.2.1. .

- $\phi_x(0) = 1$.
- $\forall u \in \mathbb{R},\ \phi_X(-u) = \overline{\phi_X(u)}$

Propriété 3.2.2. . *(Continuité de Levy)*

Soit $(X_n)_{n\geqslant 1}$ une suite de v.a.r de fonctions caractéristiques φ_{X_n} et X une v.a.r de fonction caractéristique φ_X, toutes définies sur un même espace de probabilité.

a. *Si $(X_n)_n$ converge en loi vers X, alors φ_{X_n} converge uniformément vers φ_X sur tout intervalle $[-a, a]$ (i.e $sup|\varphi_{X_n}(t) - \varphi_X(t)| \to 0,\ t \in [-a, a]$.*

b. *Si $(\varphi_{X_n}) \to \varphi$ (de partie réelle continue en 0), alors φ est fonction caractéristique d'une v.a.r vers laquelle X_n converge en loi.*
En résumé : $\{\forall x \in \mathbb{R};\ \varphi_{X_n}(t) \to \varphi_X(t)\} \Leftrightarrow \{X_n \xrightarrow{\mathcal{L}} X\}$.

Exemples de calcul de fonction caractéristique

* Fonction caractéristique de la loi normale.
$X \sim \mathcal{N}(\mu, \sigma^2)$ $\quad$ $\phi_X = ?$
On pose $Y = \dfrac{X - \mu}{\sigma} \Rightarrow X = \sigma Y + \mu$. On a $Y \sim \mathcal{N}(0, 1)$.

$$\begin{aligned}\phi_X(u) &= \mathbb{E}(e^{iuX}) \\ &= \mathbb{E}(e^{iu(\sigma Y+\mu)}) \\ &= e^{iu\mu}\mathbb{E}(e^{iu\sigma Y}) \\ &= e^{iu\mu}\phi_Y(\sigma u)\end{aligned}$$

* Calcul de ϕ_Y

$$\phi_Y(u) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} e^{-\frac{x^2}{2}}\, dx$$

$u \mapsto e^{iux}e^{-\frac{x^2}{2}}$ étant C_1, on la dérive sous le signe somme.

$$
\begin{aligned}
\phi'_Y(u) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} ixe^{iux} e^{-\frac{x^2}{2}}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \left(\int_{\mathbb{R}} (u+ix) e^{iux} e^{-\frac{x^2}{2}}\, dx - u \int_{\mathbb{R}} e^{iux} e^{-\frac{x^2}{2}}\, dx \right) \\
&= \frac{1}{\sqrt{2\pi}} \left(-i \int_{\mathbb{R}} (iu-x) e^{iux-\frac{x^2}{2}}\, dx - u \int_{\mathbb{R}} e^{iux} e^{-\frac{x^2}{2}}\, dx \right) \\
&= \frac{1}{\sqrt{2\pi}} \left(-i \left[e^{iux-\frac{x^2}{2}} \right]_{-\infty}^{+\infty} - u \int_{\mathbb{R}} e^{iux} e^{-\frac{x^2}{2}}\, dx \right) \\
&= -u\phi_Y(u)
\end{aligned}
$$

Donc

$$
\begin{aligned}
\frac{\phi'_Y(u)}{\phi_Y(u)} = -u \Rightarrow \ln|\phi_Y(u)| &= -\frac{1}{2}u^2 + c \\
\Rightarrow \phi_Y(u) &= Ke^{-\frac{1}{2}u^2}
\end{aligned}
$$

$\phi_Y(0) = 1 \Rightarrow K = 1$ donc $\phi_Y(u) = e^{-\frac{1}{2}u^2}$ et $\phi_X(u) = e^{iu\mu - \frac{\sigma^2 u^2}{2}}$.

Autres exemples (les retrouver)

Loi	Fonction caractéristique
Bernoulli $\mathcal{B}(p)$	$(1-p)+pe^{iu}$
Binomiale $\mathcal{B}(n,p)$	$((1-p)+pe^{iu})^n$
$\mathcal{P}$oisson	$e^{\lambda(e^{iu}-1)}$
Exponentielle	$\dfrac{\lambda}{\lambda - iu}$

3.3 Théorème centrale limite

Le théorème centrale limite permet de préciser la vitesse à laquelle la convergence précédente (Loi des grands nombres) a lieu.

Théorème 3.3.1. .

Soit $(X_n)_{n\geqslant 1}$ *une suite de v.a.r iid telles que* $\mathbb{E}(X_1^2) < +\infty$. *On note* $\sigma^2 = \mathbb{V}ar(X_1)$.
Alors, lorsque $n \to +\infty$,

$$\sqrt{n}\left(\frac{\bar{X_n} - \mathbb{E}(X_1)}{\sigma}\right) \xrightarrow{\mathcal{L}} Y \sim \mathbb{N}(0,1)$$

.

Exemple 3.3.1. .

Si $X_1, \cdots, X_n$ *sont iid* $\sim \mathcal{B}(p)$ *(loi de Bernoulli de paramètre* p*),* $X_1 + \cdots + X_n \sim \mathcal{B}(n,p)$. $\bar{X_n} = \dfrac{X_1 + \cdots + X_n}{n}$ *est donc telle que* $\mathbb{E}(\bar{X}) = p$ *et* $\mathbb{V}ar(\bar{X}) = \dfrac{p(1-p)}{n}$.
Pour n *suffisamment grand, l'application de TCL donne :*

$$\sqrt{n}\left(\frac{\bar{X_n} - p}{\sqrt{p(1-p)}}\right) \sim \mathbb{N}(0,1) \quad \left(ou\ \frac{n\bar{X_n} - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1)\right)$$

Dans les applications, on peut assimiler la loi binomiale à une loi normale dès que $np > 15$ *et* $n(1-p) > 15$ *ou si* $n > 30, np > 5$ *et* $n(1-p) > 5$.

Exercice 3.3.1. .

Soit $(X_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires de loi exponentielle de paramètre* λ_n. *Étudier la convergence en loi dans les trois cas suivants :*

a. $\lim\limits_{n\to+\infty} \lambda_n = \lambda, \quad \lambda \in]0, +\infty[$.

b. $\lim\limits_{n\to+\infty} \lambda_n = +\infty$.

c. $\lim\limits_{n\to+\infty} \lambda_n = 0$.

Chapitre 4

Estimation paramétrique

En statistique, une méthode d'estimation est toute procédure permettant de construire une fonction de variables aléatoires (données) destinée à approcher un paramètre inconnu dans un modèle. lorsque ce paramètre est de dimension finie, on parle d'estimation paramétrique qui peut être une estimation ponctuelle ou par intervalle de confiance selon le contexte. Une estimation est dite ponctuelle lorsque le paramètre qu'on cherche à approcher est un point. Lorsqu'on s'intéresse plutôt à un intervalle dans lequel se situe le paramètre inconnu, on parle d'estimation par intervalle de confiance.

4.1 Estimation : problématique et définitions

4.1.1 Données

A partir d'un échantillon de taille n, les données sont les valeurs $x_1, \cdots, x_n$, réalisations de n variables aléatoires aléatoires $X_1, \cdots, X_n$.

4.1.2 Modèle

C'est la famille des lois de probabilité suivies par les variables aléatoires. Dans le cas des modèles paramétriques, cette famille est caractérisée par un paramètre $\theta \in \Theta \subset \mathbb{R}^d$.

Souvent, les variables aléatoires $X_1, \cdots, X_n$ sont iid $\sim f_\theta \in \mathcal{F}$ où $\mathcal{F}$ désigne une famille paramétrique.

Exemple 4.1.1. .

- $\{\mathcal{B}(p),\ p \in]0,1[\}$
- $\{\mathcal{P}(\lambda),\ \lambda \in \mathbb{R}_+^*\}$

- $\{\mathcal{N}(\mu, \sigma^2),\ (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$

Problème

- On suppose que f_θ est complètement définie par un paramètre θ inconnu.
- Soit $x_1, \cdots, x_n$, des réalisations de $X \sim f_\theta$.
- On cherche à estimer la valeur θ_0 de θ qui a généré les données $x_1, \cdots, x_n$.

Définitions 4.1.1. .

Soit $(X_1, \cdots, X_n)$, un échantillon (des variables aléatoires iid) issu de f_θ. Une statistique est une fonction (ou application) mesurable T de X ne dépendant pas de θ et à valeur dans $\mathbb{R}^d$.

- *Intuitivement, toute fonction de l'échantillon est une statistique.*
- *Toute statistique est elle même une variable aléatoire avec sa propre loi.*

Exemple 4.1.2. .

- $T(X) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
- $T(X) = max(X_1, \cdots, X_n)$.

Définitions 4.1.2. .

Soit $\{f_\theta,\ \theta \in \Theta\}$, un modèle paramétrique où $\Theta \in \mathbb{R}^d$ et soit $X = (X_1, \cdots, X_n) \in f_{\theta_0}$ pour un certain $\theta_0 \in \Theta$.

Un estimateur $\widehat{\theta}$ de θ_0 est une statistique à valeurs dans Θ. Ainsi, toute statistique $T : \mathbb{R}^n \to \Theta$ est un potentiel estimateur.

4.2 Qualités d'un estimateur

Soit $T_n = f(X_1, \cdots, X_n)$, un estimateur de θ. On cherche un critère permettant de mesurer la qualité de l'estimateur. Ce critère doit être fondé sur l'écart entre l'estimateur T_n et la vraie valeur θ : $T_n - \theta$ s'appelle erreur d'estimation. On a la décomposition suivante :

$$T_n - \theta = \underbrace{(T_n - \mathbb{E}(T_n))}_{Erreur\ aléatoire} + \underbrace{(\mathbb{E}(T_n) - \theta)}_{Erreur\ systématique\ ou\ biais}$$

Définitions 4.2.1. .

- *On appelle biais de T_n pour θ, la valeur $b_\theta(T_n) = \mathbb{E}(T_n) - \theta$.*

- *Un estimateur T_n de θ est dit sans biais si $b_\theta(T_n) = 0$, c'est-à-dire $\mathbb{E}(T_n) = \theta$.*

Définitions 4.2.2. .

- *Un estimateur T_n de θ est dit convergent si $\lim_{n\to+\infty} \mathbb{E}(T_n) = \theta$.*

- *T_n sera dit consistant si T_n convergent en probabilité vers θ, c'est-à-dire*

$$\lim_{n\to+\infty} \mathbb{P}(|T_n - \theta| > \varepsilon) = 0$$

ou

$$\forall \beta > 0,\ \exists\, n_0,\ n \geqslant n_0 / \ \mathbb{P}(|T_n - \theta|) > \varepsilon) < \beta$$

Théorème 4.2.1. .

Si T_n est convergent et de variance tendant vers 0 quand $n \to +\infty$, alors T_n est consistant.

Définitions 4.2.3. .

On appelle erreur quadratique moyenne de T_n, la quantité

$$EQM(T_n) = \mathbb{E}((T_n - \theta)^2)$$

$\sqrt{EQM(T_n)}$ mesure l'erreur moyenne de l'estimation et donc la précision de l'estimation.

On a une bonne précision si $EQM(T_n)$ est faible.

Théorème 4.2.2. .

Soit T_n un estimateur de θ, on a :

$$\mathbb{E}((T_n - \theta)^2) = \mathbb{V}ar(T_n) + \underbrace{(\mathbb{E}(T_n) - \theta)^2}_{(b_\theta(T_n))^2}$$

Remarque 4.2.1. .

Entre deux estimateurs sans biais, le meilleur sera celui dont la variance est minimale (on parle d'efficacité).
Un estimateur efficace est donc un estimateur sans biais et de variance minimale parmi les estimateurs sans biais.

4.3 Construction d'un estimateur ponctuel

4.3.1 Méthode du maximum de vraisemblance

Démarche

Soit X un v.a.r de loi paramétrique (discrète ou continue) dont on veut estimer le paramètre θ. Alors, on définit une fonction f telle que :

$$f(x,\theta) = \begin{cases} f_\theta(x) \text{ si } X \text{ est une v.a continue à densité } f_\theta \\ \mathbb{P}_\theta(X = x) \text{ si } X \text{ est une v.a discrète de probabilité ponctuelle } \mathbb{P}_\theta \end{cases}$$

Définitions 4.3.1. .

On appelle vraisemblance de θ pour une réalisation $(x_1, \cdots, x_n)$ d'un échantillon, la fonction de θ :

$$L : \Theta \to [0;1]$$
$$\theta \mapsto L(x_1, \cdots, x_n, \theta) = f(x_1, \cdots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Remarque 4.3.1. . *La vraisemblance est définie aux points d'observation.*

Définitions 4.3.2. .

La méthode qui consiste à estimer θ par la valeur maximisant L (vraisemblance) s'appelle méthode du maximum de vraisemblance.

$$\widehat{\theta} = \{\theta \,/\, L(\widehat{\theta}) = sup\ L(\theta),\ \theta \in \Theta\}$$

Déterminer $\widehat{\theta}$ est un problème d'optimisation. On utilise le fait que si L est dérivable et si L admet un maximum global en une valeur $\widehat{\theta}$, alors la dérivée première s'annule en $\widehat{\theta}$ et que la dérivée seconde est négative en ce point.

Réciproquement, si la dérivée première s'annule et que la dérivée seconde est négative en $\theta = \widehat{\theta}$, alors $\widehat{\theta}$ est le maximum local (et non global) de L. Il est nécessaire de vérifier qu'il s'agit bien d'un maximum global.

La vraisemblance étant positive et la fonction ln croissante, il est équivalent et souvent plus simple de maximiser le logarithme népérien de la vraisemblance.

Ainsi :

a. La condition nécessaire : $\dfrac{\partial L(x_1, \cdots, x_n, \theta)}{\partial \theta} = 0$ ou $\dfrac{\partial ln(L(x_1, \cdots, x_n, \theta))}{\partial \theta} = 0$ permet de trouver des valeurs de $\widehat{\theta}$ appelées points critiques.

b. $\theta = \widehat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique.

Exemple 4.3.1. .

On suppose que $X \sim \mathcal{P}(\lambda)$ et $X_1, \cdots, X_n$ de même loi que X. On a :

$$\begin{aligned} L(x_1, \cdots, x_n, \lambda) &= \prod_{i=1}^{n} \mathbb{P}_\lambda(X = x_i) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \\ l_n(\lambda) &= ln(L) \\ &= -n\lambda + (ln(\lambda)) \sum_{i=1}^{n} x_i - ln(\prod_{i=1}^{n} x_i!) \\ \frac{\partial l_n(\lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^n x_i}{\lambda} \\ \frac{\partial l_n(\lambda)}{\partial \lambda} = 0 \Rightarrow \widehat{\lambda} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \frac{\partial^2 l_n(\lambda)}{\partial \lambda} &= -\frac{\sum_{i=1}^n x_i}{\lambda^2} < 0 \quad \forall\, \lambda \end{aligned}$$

donc $\widehat{\lambda}$ est un maximum local de $l_n(\lambda)$.

Exemple 4.3.2. .

Dans les modèles suivants :

- $\{\mathcal{N}(\mu, 1),\ \mu \in \mathbb{R}\}$.
- $\{\mathcal{N}(0, \sigma^2),\ \sigma^2 \in \mathbb{R}_+^*\}$.
- $\{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \quad \sigma^2 \in \mathbb{R}_+^*\}$

Déterminer les estimateurs du maximum de vraisemblance.

Propriétés de l'EMV

Information de Fisher

Soit Θ un intervalle ouvert de $\mathbb{R}$ (ou de $\mathbb{R}^d$ si θ est un vecteur à coordonnées) et un ensemble χ ne dépendant pas de θ. On suppose que les conditions suivantes sont vérifiées :

- $\forall\, \theta \in \Theta,\ \forall\, x \in \chi,\ f_\theta(x) > 0$.

- $\forall\ x \in \chi,\ \theta \mapsto f_\theta(x)$ est une fonction deux fois différentiable.

- $\forall\ \theta \in \Theta,\ \forall\ x \in \chi$, on peut dériver deux fois $f_\theta(x)$ par rapport à θ sous le signe intégrale.

Définitions 4.3.3. .

- *On appelle fonction score (ou score), la dérivée du logarithme de la vraisemblance.*
$$\mathcal{S}_n(\theta) = \frac{\partial l_n(\lambda)}{\partial \lambda}$$

- *Sous les conditions de régularité énumérées ci-dessus, on définit l'information de Fisher notée $I(\theta)$ pour une observation par :*
$$I(\theta) = \mathbb{V}ar(\mathcal{S}_1(\theta))$$

Propriété 4.3.1. .

a. $\mathbb{E}(\mathcal{S}_1(\theta)) = 0$.

b. $I(\theta) = -\mathbb{E}\left(\dfrac{\partial^2 l_n(\theta)}{\partial \theta^2}\right)$.
Puisque le score est centré, on redéfinit l'information de Fisher de la manière suivante :
$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \lambda} ln(f_\lambda(x))\right)^2\right]$$
et pour une variable aléatoire discrète, on a :
$$I(\theta) = \sum\left(\frac{\partial}{\partial \theta} ln(\mathbb{P}_\theta(X = x))\right)^2 \mathbb{P}_\theta(X = x)$$

c. Si θ est unidimensionnelle, les dérivées partielles deviennent des dérivées simples.

d. Si θ est un vecteur, $I(\theta)$ est une matrice de terme général :
$$I_{ij}(\theta) = \int_{X(\Omega)}\left(\frac{\partial}{\partial \lambda_i} ln(f_\lambda(x))\left(\frac{\partial}{\partial \lambda_j} ln(f_\lambda(x))\right)\right) f_\lambda(x)dx$$

Exemple 4.3.3. .
Information de Fisher pour :
$$\{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \quad \sigma^2 \in \mathbb{R}_+^*\}$$

Définitions 4.3.4. .

- *On définit également l'information de Fisher associée à un échantillon de taille n, on a :*

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \lambda} ln(L(x_1, \cdots, x_n, \theta)) \right)^2 \right] \\ &= \mathbb{E}_\theta \left[\left(\frac{\partial^2}{\partial \lambda^2} ln(L(x_1, \cdots, x_n, \theta)) \right) \right] \end{aligned}$$

- *Dans le cas où les variables sont indépendantes et de même loi, on a :*

$$I_n(\theta) = nI_1(\theta)$$

Exemple 4.3.4. .

Déterminer $I_n(\theta)$ dans les modèles :

- $\{\mathcal{B}(p),\ p \in]0,1[\}$
- $\{\mathcal{N}(\mu, 1),\ \mu \in \mathbb{R}\}$
- $\{\mathcal{N}(0, \sigma^2),\ \sigma^2 \in \mathbb{R}_+^*\}$

Propriété 4.3.2. .

Soit T_n un estimateur du maximum de vraisemblance sous les conditions de régularité énumérées plus haut, on a :

a. T_n est un estimateur consistant.

b. $\lim\limits_{n \to +\infty} b_\theta(T_n) = 0$. On dit que T_n est asymptotiquement sans biais.

c. T_n est asymptotiquement gaussien. Dans le cas où les X_i sont indépendantes et de même loi, on a :

$$\sqrt{n}(T_n - \theta) \xrightarrow{L} Y \sim \mathcal{N}(0, I_1^{-1}(\theta))$$

Dans le troisième point de la propriété précédente, on peut voir que la variance asymptotique de T_n est bien :

$$I_n^{-1}(\theta) = \frac{I^{-1}(\theta)}{n}$$

Borne de Cramer Rao (ou FDCR)

Soit $X = (X_1, \cdots, X_n)$ un échantillon et $T(X)$ une statistique telle que $\mathbb{E}(T) < +\infty$ *et* $g(\theta) = \mathbb{E}(T)$ est dérivable avec $g^{'}(\theta) = \mathbb{E}_\theta \left(T \frac{\partial ln(f(X,\theta))}{\partial \theta} \right) \ \forall \ \theta.$
Ceci est vrai car sous les conditions de régularité du modèle, on a :

$$\begin{aligned} g^{'}(\theta) &= \frac{\partial}{\partial \theta} \int T(x) \cdot f(x,\theta) dx \\ &= \int T(x) \cdot \frac{\partial}{\partial \theta} f(x,\theta) dx \\ &= \int T(x) \cdot \frac{\frac{\partial}{\partial \theta} f(x,\theta)}{f(x,\theta)} f(x,\theta) dx \\ &= \int T(x) \cdot \frac{\partial ln(f(x,\theta))}{\partial \theta} f(x,\theta) dx \\ &= \mathbb{E} \left(T(x) \cdot \frac{\partial}{\partial \theta} ln(f(x,\theta)) \right) \end{aligned}$$

Théorème 4.3.1. . *(Borne de FDCR)*
Soit $X = (X_1, \cdots, X_n)$ *un vecteur aléatoire de densité conjointe* $f(X,\theta)$ *vérifiant les conditions de régularité énumérées plus haut.*
Si T_n *est une statistique vérifiant* $g^{'}(\theta) = \mathbb{E}_\theta \left(T_n(X) \frac{\partial}{\partial \theta} ln(f(X,\theta)) \right) \ \forall \ \theta$, *alors :*

$$\mathbb{V}ar(T_n) \geqslant \frac{[g^{'}(\theta)]^2}{I_n(\theta)}$$

Preuve

$I_n(\theta) = \int \left(\frac{\partial}{\partial \theta} ln(f(x,\theta)) \right)^2 f_\theta(x,\theta) dx$ car $\mathbb{E} \left(\frac{\partial}{\partial \theta} ln(f(x,\theta)) \right) = 0.$
$\mathbb{C}ov \left(T_n, \frac{\partial}{\partial \theta} ln(f(x,\theta)) \right) = \mathbb{E} \left(T_n, \frac{\partial}{\partial \theta} ln(f(x,\theta)) \right) = g^{'}(\theta).$
D'après l'égalité de Cauchy-Schwatz, on a :

$$g^{'}(\theta) = \mathbb{C}ov \left(T_n, \frac{\partial}{\partial \theta} ln(f(x,\theta)) \right) \leqslant \sqrt{\mathbb{V}ar(T_n) \cdot I_n(\theta)}$$

D'où $(g^{'}(\theta))^2 \leqslant \mathbb{V}ar(T_n) \cdot I_n(\theta)$. et $\mathbb{V}ar(T_n) \geqslant \frac{(g^{'}(\theta))^2}{I_n(\theta)}$.

4.3.2 Méthode des moments

Construction

La méthode des moments consiste à trouver une fonction h bijective continue avec h^{-1} (bijection réciproque) continue et une fonction φ mesurable telle que $\mathbb{E}(\varphi(X)) < +\infty$ et $h(\theta) = \mathbb{E}_\theta(\varphi(X)),\ \theta \in \Theta$ (1) où X est une v.a de densité (ou masse de probabilité) f_θ dépendante de θ.

On peut à partir de (1) écrire :

$$\theta = h^{-1}(\mathbb{E}_\theta(\varphi(X))) \quad (2)$$

Un estimateur des moments est donc obtenu en remplaçant les moments dans (2), par leurs versions empiriques. On obtient :

$$\widehat{\theta} = h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right)$$

où $X_i,\ i = 1, \cdots, n$ sont iid $\sim f_\theta$.

Exemple 4.3.5. .

Déterminer deux estimateurs des moments convergents de λ dans le modèle :

- $\{\mathcal{E}(\lambda),\ \lambda \in \mathbb{R}^+\}$.
- $\{\mathcal{U}_{[a,b]},\ a, b \in \mathbb{R}\}$.

Propriétés

Propriété 4.3.3. .

Par sa définition, un estimateur obtenu par la méthode des moments est un estimateur convergent.

EXERCICE

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées dont la densité est définie par

$$f_\theta(x) = \frac{x}{\theta} exp\left\{-\frac{x^2}{2\theta}\right\} \text{ si } x > 0 \text{ et } 0 \text{ sinon .}$$

1/ Déterminer l'estimateur des moments de θ.

2/ Déterminer θ_{mv}, l'estimateur du maximum de vraisemblance de θ puis montrer que c'est un estimateur sans biais de θ.

3/ Pour $n = 10$, existe-t-il un autre estimateur sans biais de θ dont la variance est strictement plus petite que celle de θ_{mv} ?

4.4 Estimation par intervalle de confiance

4.4.1 Contexte

Au lieu de donner une valeur estimée du paramètre θ, on choisit deux bornes B_1 et B_2 entre lesquelles on espère que se trouve la valeur du paramètre. Les bornes B_1 et B_2 sont fonctions de variables aléatoires.

Définitions 4.4.1. .

Soit $B_1 = f_1(X_1, \cdots, X_n)$ et $B_2 = f_2(X_1, \cdots, X_n)$. On appelle intervalle de confiance de θ, le couple (B_1, B_2).

- *$\mathbb{P}(B_1 < \theta < B_2)$ s'appelle la probabilité de recouvrement ou niveau de confiance.*
- *Si $\mathbb{P}(B_1 < \theta < B_2) = 1-\alpha$, $[B_1; B_2]$ est appelé intervalle de probabilité $1-\alpha$.*
- *En général, on choisit une valeur faible pour α (environ 0,05 ou 0.01) et on construit l'intervalle de sorte que la probabilité de recouvrement soit égale à $1-\alpha$.*

Définitions 4.4.2. .

On appelle intervalle de confiance $1-\alpha$ ($IC_{1-\alpha}$) pour θ, toute réalisation $[b_1; b_2]$ d'un intervalle de probabilité de recouvrement égale à $1-\alpha$.

4.4.2 Méthode de construction

Une façon de construire un intervalle de confiance $IC_{1-\alpha}$ consiste à utiliser une statistique dite pivotable T_n dépendant de $X_1, \cdots, X_n$ et de θ, mais dont la loi de probabilité est indépendante de θ.

On désigne par $t_{\frac{\alpha}{2}}$ et $t_{1-\frac{\alpha}{2}}$ deux quantiles de la loi de T_n. On a :

$$\mathbb{P}\left(t_{\frac{\alpha}{2}} < T_n < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Exemple 4.4.1. .

Soit $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. On note $\bar{X} = \frac{1}{n}X_i$ la moyenne empirique.

4.4.3 Quelques exemples de construction

Intervalle de confiance pour une moyenne μ

Si σ^2 est connu, on a : $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \to \sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$.

La probabilité de recouvrement est :

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

d'où

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\,; \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$$

Si σ^2 n'est pas connu, un estimateur de σ^2 est donné par $S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \sim \mathcal{N}(0,1) \quad et \quad \frac{(n-1)S^2_{n-1}}{\sigma^2} \sim \chi^2_{(n-1)}$$

donc

$$\frac{\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)}{\sqrt{\frac{(n-1)S^2_{n-1}}{(n-1)\sigma^2}}} \sim \mathcal{T}_{(n-1)}$$

donc

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{S_{n-1}}\right) \sim \mathcal{T}_{(n-1)}$$

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2},n-1} < \sqrt{n}\left(\frac{\bar{X}-\mu}{S_{n-1}}\right) < t_{1-\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

$$\mathbb{P}\left(\bar{X} - \frac{S_{n-1}}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} < \mu < \bar{X} + \frac{S_{n-1}}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

d'où

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - \frac{S_{n-1}}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}; \bar{X} + \frac{S_{n-1}}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}\right]$$

Intervalle de confiance pour la variance σ^2

Sous les mêmes hypothèses que dans le paragraphe précédent, on a :

$$\frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

$$\mathbb{P}\left(q_{\alpha_1} < \frac{(n-1)S_{n-1}^2}{\sigma^2} < q_{1-\alpha_2}\right) = 1-\alpha \quad avec\ \alpha_1 + \alpha_2 = \alpha$$

donc

$$\mathbb{P}\left(\frac{q_{\alpha_1}}{(n-1)S_{n-1}^2} < \frac{1}{\sigma^2} < \frac{q_{1-\alpha_2}}{(n-1)S_{n-1}^2}\right) = 1-\alpha$$

d'où

$$\mathbb{P}\left(\frac{(n-1)S_{n-1}^2}{q_{1-\alpha_2}} < \sigma^2 < \frac{(n-1)S_{n-1}^2}{q_{\alpha_1}}\right) = 1-\alpha$$

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S_{n-1}^2}{q_{1-\alpha_2}}; \frac{(n-1)S_{n-1}^2}{q_{\alpha_1}}\right]$$

q_t est le quantile d'ordre t d'une loi de χ^2 à $(n-1)$ ddl.

Intervalle de confiance pour une proportion

Soit X le nombre d'individus d'un échantillon de taille n possédant un caractère étudié : $X \sim \mathcal{B}(n,p)$. Il existe plusieurs intervalles de confiance pour p dont en voici quelques uns.

a. On pose $T_n = \dfrac{X-np}{\sqrt{np(1-p)}}$.

$T_n \xrightarrow{L} \mathcal{N}(0,1)$. (Théorème Central Limite)
donc

$$\mathbb{P}\left(\left|\frac{X-np}{\sqrt{np(1-p)}}\right| < z_{1-\frac{\alpha}{2}}\right) \to 1-\alpha$$

$$\Rightarrow \mathbb{P}\left((X-np)^2 < np(1-p)z^2_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(X^2 - 2npX + n^2p^2 < npz^2_{1-\frac{\alpha}{2}} - np^2z^2_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha$$

$$\mathbb{P}\left(p^2(n^2+nz^2_{1-\frac{\alpha}{2}}) - (2nX+nz^2_{1-\frac{\alpha}{2}})p + X^2 < 0\right) \simeq 1-\alpha$$

On en déduit que

$$IC_{1-\alpha}(p) = \left[\frac{\frac{X}{n} + \frac{z^2_{1-\frac{\alpha}{2}}}{2n} \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{z^2_{1-\frac{\alpha}{2}}}{4n} + \frac{X}{n}\left(1-\frac{X}{n}\right)}}{1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}}\right]$$

b. On pose $\widehat{p} = \dfrac{X}{n}$.

$$\sqrt{n}\frac{\left(\frac{X}{n}-p\right)}{\sqrt{p(1-p)}} \xrightarrow{L} \mathcal{N}(0,1)$$

$$\sqrt{n}\frac{\left(\frac{X}{n}-p\right)}{\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}} = \sqrt{n}\frac{\sqrt{p(1-p)}}{\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}} \times \frac{\left(\frac{X}{n}-p\right)}{\sqrt{p(1-p)}} \xrightarrow{L} \mathcal{N}(0,1)$$

$$\mathbb{P}\left(\sqrt{n}\left|\frac{\left(\frac{X}{n}-p\right)}{\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}}\right| \leqslant z_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha$$

donc

$$\mathbb{P}\left(\frac{X}{n} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)} < p < \frac{X}{n} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}\right) = 1-\alpha$$

d'où

$$IC_{1-\alpha}(p) = \left[\frac{X}{n} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}; \frac{X}{n} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)} \right]$$

Chapitre 5

Principes de base d'un test statistique

5.1 Problématique

La méthodologie des tests statistiques consiste à répondre (à l'aide de résultats expérimentaux) à une question concernant les paramètres de la loi de probabilité des variables aléatoires.

Exemple 5.1.1. .

- *Est-ce que le processus de remplissage des bouteilles de la SOBEBRA est bien réglé ?*
- *Est-ce que la nouvelle variété de soja est plus productive que l'ancienne ?*
- *Le revenu annuel des ménages dans la ville de Parakou est-il conforme à la moyenne nationale ?*
- *Y-a-t-il disparité entre les revenus annuels des ménages dans les grandes villes du Bénin ?*
- *Les bouteilles d'eau en plastique sont t-elles vraiment potables ?*

Toutes les questions de ce type ne sont pas susceptibles d'obtenir des réponses par un test statistique. Pour que ça soit le cas, il faut réunir quatre conditions :

a. La question doit être posée de telle sorte qu'il n'y ait que deux réponses possibles : Oui ou Non.

b. On doit avoir des données chiffrées résultant d'un échantillon ou d'une expérimentation.

c. Ces données doivent pouvoir être considérées comme réalisation de variables aléatoires dont la loi de probabilité est connue par l'intermédiaire du modèle statistique.

d. La question posée doit concerner un ou plusieurs paramètres de cette loi.

5.2 Concepts et définitions

On considère $X = (X_1, \cdots, X_n)$ un échantillon de variables aléatoires iid dans le modèle $P = \{P_\theta,\ \theta \in \Theta\}$. Soit $\{H_0,\ H_1\}$ une partition de l'espace des paramètres.

- On appelle hypothèse, une assertion concernant θ.

- Un test d'hypothèse (paramétrique) est constitué de quatre éléments :
  - Des données : $x_1, \cdots, x_n$ réalisation de n variables aléatoires $X_1, \cdots, X_n$.
  - Un modèle statistique : la loi de probabilité. $P_\theta \in P$ dépendant d'un ou de plusieurs paramètres.
  - Une hypothèse prioritaire, appelée hypothèse nulle parce qu'elle s'écrit souvent sous la forme $f(\theta) = 0$. On la note H_0.
  - Une règle de décision :
    Soit $T = f(X_1, \cdots, X_n)$ une statistique appelée "statistique de test" et W un sous-ensemble de l'ensemble des valeurs possibles de T, composée des valeurs de T "improbables sous l'hypothèse H_0". L'ensemble W est appelé "région de rejet" du test. La règle de décision est la suivante : Si $T \in W$, on rejette H_0.

- La réponse du test est :
  - Soit on accepte l'hypothèse H_0, ce qui signifie que les données ne sont pas en contradiction avec l'hypothèse.
  - Soit on rejette l'hypothèse H_0, ce qui signifie qu'il est très peu probable d'obtenir les résultats que l'on a trouvés si l'hypothèse est vraie. Dans ce cas, on peut dire que les données sont en contradiction avec l'hypothèse.

Exemple 5.2.1. .

1 Doit-on régler le processus de fabrication de la bière "La Béninoise" ?
On étudie la contenance d'une bouteille de la bière "La Béninoise". On suppose que la contenance X d'une bouteille $66cl$ est normalement distribuée d'espérance μ. On veut savoir si les bouteilles contiennent réellement $66cl$. On prélève un échantillon de taille n et l'on effectue les n mesures.

- *Données : $x_1, \cdots, x_n$ mesures de la contenance de chacune des n bouteilles : réalisations des variables aléatoires $X_1, \cdots, X_n$.*

- *Modèle statistique : Les X_i sont indépendantes et suivent la même loi $\mathcal{N}(\mu, \sigma^2)$.*

- *Hypothèse nulle : $H_0 = \{\mu = 66cl\}$.*

- *Règle de décision : Soit la statistique de test $T = \dfrac{1}{n}\sum_{i=1}^{n} X_i$. La région de rejet ou région critique est de la forme $W = \{t;\ |t - 66| > \ell\}$ où ℓ est un seuil à fixer.*
  Si $T \in W$, on rejette H_0.

2 Un lot de graines destinées à la vente est considéré comme conforme si le taux de germination (proportion de graines donnant un germe normal) est supérieure à 0,85. Afin de décider de la certification d'un lot, on prélève un échantillon de n graines que l'on met à germer.

- *Données : x= nombre de germes normaux sur n graines, réalisations de la variable aléatoire X.*

- *Modèle statistique : $X \sim \mathcal{B}(n, p)$ où p est le "taux" de germination du lot.*

- *Hypothèse nulle : $H_0 = \{p \geqslant 0,85cl\}$.*

- *Règle de décision : La statistique utilisée est $T = X$. La région de rejet ou région critique est de la forme $W = \{t;\ t < \ell\}$, ℓ étant un seuil à fixer.*
  Si $T \in W$, on rejette H_0.

5.3 Nature des hypothèses

Il est à noter que H_0 est l'hypothèse privilégiée : c'est celle que l'on garde si le résultat de l'expérience n'est pas clair. En ce sens, il y a une analogie entre un test

d'hypothèse et un procès : tout suspect est présumé innocent et l'accusation doit apporter la preuve de sa culpabilité avant que la justice ne décide de le condamner.

Quand on accepte H_0, on ne prouve pas qu'elle est vraie. On accepte de conserver H_0 parce qu'on n'a pas accumulé suffisamment d'éléments matériels contre elle pour la rejeter. Accepter H_0, c'est "acquitter faute de preuve".

Définitions 5.3.1. .

On appelle hypothèse alternative que l'on note H_1, une hypothèse différente de H_0. C'est souvent le contraire de H_0. On dit que l'on teste H_0 contre H_1.

5.4 Erreur, niveau et puissance d'un test

A l'issu d'un test, les quatre situations possibles sont résumées dans le tableau suivant :

	Accepeter H_0	Rejeter H_0
H_0 vraie	Bonne décision	Erreur de première espèce
H_1 vraie	Erreur de $2^{ième}$ espèce	Bonne décision

Soit $H_0 = \{\theta \in \Theta_0\}$.

Définitions 5.4.1. .

- *On appelle risque de première espèce, que l'on note $\alpha(\theta)$, la probabilité de rejeter H_0 alors qu'elle est vraie.*

$$\alpha(\theta) = \mathbb{P}(T \in W/\ \theta \in \Theta_0)(\theta \in \Theta_0 \mapsto \mathbb{P}_\theta(T \in W))$$

- *On appelle niveau que l'on note α, la valeur la plus élevée du risque de première espèce quand θ parcourt Θ_0.*

$$\alpha = \sup_{\theta \in \Theta_0}[\alpha(\theta)]$$

- *Soit $\Theta_1 = \Theta_0^c$. On appelle risque de deuxième espèce, et on note $\beta(\theta)$, la probabilité de ne pas rejeter H_0 alors que c'est H_1 qui est vraie.*

$$\beta(\theta) = \mathbb{P}(T \notin W/\ \theta \in \Theta_1)(\theta \in \Theta_1 \mapsto \mathbb{P}_\theta(T \notin W))$$

- *On appelle puissance d'un test que l'on note $\pi(\theta)$, la probabilité de rejeter H_0 alors qu'elle est fausse. La puissance d'un test mesure donc son aptitude à rejeter une hypothèse fausse. On a :*

$$\pi(\theta) = 1 - \beta(\theta)$$

.

Définitions 5.4.2. .

Soit $(W_n)_n$ une suite de régions critiques où n désigne la taille de l'échantillon. La suite de tests de régions critiques $(W_n)_n$ est dite :

- *Convergente si* $\forall\, \theta \in \Theta_1 \lim\limits_{n\to+\infty} \mathbb{P}_\theta(T_n \in W_n) = 1.$

- *de niveau asymptotique* α *si* $\sup\limits_{\theta\in\Theta_0} \lim\limits_{n\to+\infty} \mathbb{P}_\theta(T_n \in W_n) = \alpha.$

Définitions 5.4.3. .

- *Un test est dit sans biais si sa fonction puissance est toujours supérieure au niveau sous l'hypothèse alternative.*

$$i.e \quad \pi(\theta) \geqslant \alpha \,\forall\, \theta \in \Theta_1$$

- *Lorsque l'on observe* $x = (x_1, \cdots, x_n)$*, on appelle p-valeur du test, la probabilité sous* H_0 *que la statistique de test notée* T_n *prenne des valeurs plus défavorables pour l'acceptation de* H_0 *que celle* t_n^{obs} *observée.*

5.5 Construction de la règle de décision

Supposons que l'on dispose d'une statistique de test pertinente. Par exemple, on utilise souvent un estimateur T_n du paramètre θ pour tester l'hypothèse $H_0 = \{\theta \leqslant \theta_0\}$ contre $H_1 = \{\theta > \theta_0\}$. La règle de décision est alors du type suivant : si $T_n - \theta_0 > \ell$, on rejette H_0. Il faut donc choisir ℓ.

Principe : On choisit ℓ de sorte que le niveau du test soit égal à une valeur (faible) fixée à priori α. On prend souvent $\alpha = 5\%,\ 1\%$ *ou* $0,1\%$.
On veut donc que $\mathbb{P}(T_n\theta_0 > \ell) = \alpha$ sous H_0 (i.e $\theta \in \Theta_0$).
Dans le cas où on connait la loi de $T-\theta_0$ et en particulier sa fonction de répartition F, on doit choisir ℓ de telle sorte que $1 - F(\ell) = \alpha$. Soit $\ell = F^{-1}(1-\alpha)$.

Remarque 5.5.1. .

- *On choisit* ℓ *de façon à contrôler l'erreur de première espèce.*

- *L'erreur de deuxième espèce n'entre pas en compte dans la détermination de* ℓ*. C'est en ce sens que l'hypothèse* H_0 *est privilégiée. On veut avant tout contrôler le risque de rejeter* H_0 *à tort. (quand on va faire un test médical pour détecter une maladie, il est préférable de donner un résultat positif au test à tort que le contraire qui laisse la maladie vous abattre)*

Chapitre 6

Tests dans un modèle gaussien

considérons le modèle gaussien

$$P = \{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}\ et\ \sigma^2 > 0\}$$

Les variables $X_1, \cdots, X_n$ sont iid $\sim \mathcal{N}(\mu, \sigma^2)$

6.1 Test sur la moyenne

Soit $\mu_0 \in \mathbb{R}$. On désire tester $H_0 = \{\mu = \mu_0\}$ contre $H_1 = \{\mu \neq \mu_0\}$ (test bilatéral) au niveau $\alpha \in]0;1[$.
On note $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
On rappelle que $\bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$ sous H_0.
Donc $\sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0,1)$ sous H_0.
$\sqrt{n}\frac{\bar{X} - \mu_0}{S} \sim t(n-1)$ sous H_0.

Sous H_1 par la loi forte des grands nombres, $\bar{X} - \mu_0$ converge presque sûrement vers $\mu - \mu_0$ et S converge presque sûrement vers σ.
Donc T_n tend p.s vers $+\infty$ ou $-\infty$ lorsque $n \to +\infty$ (selon que $\mu > \mu_0$ ou $\mu < \mu_0$
Si σ est connue
Pour un test de niveau α, on rejette H_0 si $\frac{|\bar{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} > z_{1-\frac{\alpha}{2}}$.
C'est-à-dire si $\bar{X} > \mu_0 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ ou $\bar{X} > \mu_0 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

Dans l'exemple de la contenance des bouteilles de la SOBEBRA, on a vu qu'on rejette H_0 si $|\bar{X} - \mu_0| > \ell$. On fixe alors un niveau α et on choisit ℓ de sorte que le niveau du test soit égal à α.

$$\alpha = \mathbb{P}(|\bar{X} - \mu_0| > \ell) = \mathbb{P}\left(\frac{|\bar{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} > \frac{\ell}{\frac{\sigma}{\sqrt{n}}}\right)$$

Donc $\mathbb{P}\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{\ell}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2} \Rightarrow \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\ell}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2}$

D'où $1 - F\left(\frac{\ell}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2}$ i.e $F\left(\frac{\ell}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$

$$\ell = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Si σ est inconnue

On mène le même raisonnement que précédemment avec $T_n = \sqrt{n}\frac{\bar{X} - \mu_0}{S}$.
Pour un test de niveau α, on rejette H_0 si

$$\frac{|\bar{X} - \mu_0|}{\frac{S}{\sqrt{n}}} > t_{n-1,1-\frac{\alpha}{2}}.$$

car sous H_0, $\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t(n-1)$.

6.1.1 Test de l'hypothèse $H_0 = \{\mu \leqslant \mu_0\}$ contre $H_1 = \{\mu > \mu_0\}$

Si σ est connue

$$T_n = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Pour un test de niveau α, on rejette H_0 si

$$\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} \Rightarrow \bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

Si σ est inconnue

Statistique du test : $T_n = \dfrac{\bar{X} - \mu_0}{\dfrac{S}{\sqrt{n}}}$

Règle de décision : Rejeter H_0 si $\dfrac{\bar{X} - \mu_0}{\dfrac{S}{\sqrt{n}}} > t_{n-1,1-\alpha}$

6.2 Tests sur l'écart-type

6.2.1 Test bilatéral sur l'écart-type

(compléter avec le cas où μ est connu)

On souhaite maintenant tester $H_0 = \{\sigma = \sigma_0\}$ contre $H_1 = \{\sigma \neq \sigma_0\}$.

Statistique : $T_n = \dfrac{(n-1)S^2}{\sigma_0^2}$ (μ est inconnu)

Loi : Sous H_0, $\dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

Pour un test de niveau α, on rejette H_0 si $\dfrac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\frac{\alpha}{2}}$ ou $\dfrac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,\frac{\alpha}{2}}$.

Ou si $S^2 > \dfrac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}}$ ou $S^2 < \dfrac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}}$

Pour mieux comprendre la règle de décision, on écrit :

$$T_n = \frac{(n-1)S^2}{\sigma_0^2} = \frac{\sigma^2}{\sigma_0^2} * \frac{(n-1)S^2}{\sigma^2}$$

On sait déjà que $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

T_n prend des valeurs de plus en plus grandes lorsque σ^2 croit et des valeurs de plus en plus petites lorsque σ^2 décroit.

6.2.2 Test de l'hypothèse $H_0 = \{\sigma \leqslant \sigma_0\}$ contre $H_1 = \{\sigma > \sigma_0\}$

Statistique de test : $T_n = \dfrac{(n-1)S^2}{\sigma_0^2}$ (μ est inconnu)

Règle de décision : Rejeter H_0 si $\dfrac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}$, i.e $S^2 > \dfrac{\sigma_0^2 \chi^2_{n-1,1-\alpha}}{n-1}$.

Exercice 6.2.1. . *(Voir TD)*

6.3 Tests sur une proportion

6.3.1 Test bilatéral sur une proportion

Si X est le nombre d'individus d'un échantillon de taille n possédant un caractère étudié, on a $X \sim \mathcal{B}(n,p)$ où p est la proportion d'individus possédant ce caractère dans la population. On souhaite tester $H_0 = \{p = p_0\}$ contre $H_1 = \{p \neq p_0\}$. On fait l'approximation suivante sous H_0

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{Loi} \mathcal{N}(0,1)$$

Cette approximation est d'autant meilleure que np_0 et $n(1-p_0)$ sont grands. Dans la pratique, on suppose que cette approximation est bonne lorsque np_0 et $n(1-p_0)$ sont supérieurs à 10.

Pour un test de niveau approximativement α, on rejette H_0 si $\dfrac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > z_{1-\frac{\alpha}{2}}$.

6.3.2 Test unilatéral sur une proportion

On veut maintenant tester $H_0 = \{p \leqslant p_0\}$ contre $H_1 = \{p > p_0\}$.

Statistique de test : $T_n = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$

Règle de décision : Rejeter H_0 si $\dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha}$ où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$.

6.4 Exemple de calcul de puissance

Soit $H_0 = \{\mu = \mu_0\}$ contre $H_1 = \{\mu \neq \mu_0\}$.

$$\pi(\mu) = 1 - \beta(\mu)$$

$$\begin{aligned}
\beta(\mu) &= 1 - \mathbb{P}_{H_1}\left(\frac{|\bar{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} > z_{1-\frac{\alpha}{2}}\right) \\
&= 1 - \mathbb{P}\left[\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} + \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\frac{\alpha}{2}}\right) \cup \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} + \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_{1-\frac{\alpha}{2}}\right)\right] \\
&= 1 - \mathbb{P}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > -\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < -\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}}\right)
\end{aligned}$$

comme $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ *sous* H_0, *on a* :

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}\left(\mathcal{N}(0,1) \leqslant z_{1-\frac{\alpha}{2}} - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) - \mathbb{P}\left(\mathcal{N}(0,1) < -z_{1-\frac{\alpha}{2}} - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= \phi\left(z_{1-\frac{\alpha}{2}} - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) - \phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)
\end{aligned}$$

μ *parcourant* H_1

$$\pi(\mu) = 1 - \beta(\mu).$$

6.5 Comparaison de populations

6.5.1 Contexte et exemples

A l'aide de deux échantillons issus de deux populations différentes, on souhaite comparer ces populations.

Données : $x_1, \cdots, x_{n_1}$ sont les résultats obtenus dans la population 1, réalisations des v.a $X_1, \cdots, X_{n_1}$. $y_1, \cdots, y_{n_2}$ sont les résultats obtenus dans la population 2, réalisations de v.a $Y_1, \cdots, Y_{n_2}$. Les variables X_i et Y_j sont supposées indépendantes.
Modèle statistique : $X_i \sim Loi(\theta_1)$ et $Y_j \sim Loi(\theta_2)$
Hypothèse testée : $H_0 = \{\theta_1 = \theta_2\}$.
Statistique de test : $T_n = f(X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2})$.
On rejette H_0 si $T \in W$; W étant la région de rejet.

Exemple 6.5.1. .

1. *Choix entre deux placements.*
   - *Données : 10 placements de type A ont donné les pourcentages annuels d'intérêt suivant :* $2\%, 8\%, 4.2\%, 6.3\%, 9.6\%, 10.2\%, 11\%, 8\%, 12.4\%$ *et* 13.1%. *Ce sont des réalisations de dix variables aléatoires définies sur la population A :* $X_1, \cdots, X_{10}$.
     Les résultats donnés par un placement de type B sont les suivants : $4\%, 6\%, 8\%, 11.3\%, 12.3\%, 8\%, 2.6\%, 14.7\%, 5.7\%$ *et* 16% *réalisations de dix variables aléatoires définies sur la population B :* $Y_1, \cdots, Y_{10}$.
     On souhaite comparer les deux placements.
   - *Modèle :* $X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ *et* $Y_j \sim \mathcal{N}(\mu_2, \sigma_2)$
   - *Hypothèse :* $H_0 = \{\mu_1 = \mu_2\}$
   - *Statistique :* $T = \bar{X} - \bar{Y}$ *où* $\bar{X} = \dfrac{1}{10}\sum_{i=1}^{10} X_i$ *et* $\bar{Y} = \dfrac{1}{10}\sum_{i=1}^{10} Y_i$.
   - *Région de rejet :* $W = \{t;\ |t| > \ell\}$ ℓ *étant un seuil à fixer.*
     RD : rejeter H_0 *si* $T \in W$.
2. *Efficacité d'un médicament*
   - *Données : Sur 40 malades traités par un placébo, on a obtenu 14 guérisons, réalisation de X. Sur 50 malades traités par le médicament testé, on a obtenu 22 guérisons, réalisation de Y. On souhaite comparer les deux placements.*
   - *Modèle :* $X \sim \mathcal{B}(n_1, p_1)$ *où* p_1 *est le taux de guérison lié au placébo ; et* $Y \sim \mathcal{B}(n_2, p_2)$ *où* p_2 *est le taux de guérison lié au médicament.*
   - *Hypothèse :* $H_0 = \{p_1 = p_2\}$ *contre* $H_1 = \{p_1 < p_2\}$

- *Statistique : $T = f(X, Y)$.*

- *Région de rejet : on rejette H_0 si $T \in W$.*

6.5.2 Test de comparaison de deux moyennes

(représenter les boites à moustaches)

On admet le modèle statistique et notations suivants :

$X_1, \cdots, X_{n_1}$ iid, $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$

$Y_1, \cdots, Y_{n_2}$ iid, $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i. \quad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_j. \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2$$

$$S^2 = \frac{(n_1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Test de $H_0 = \{\mu_1 = \mu_2\}$ contre $H_1 = \{\mu_1 \neq \mu_2\}$

- Si σ_1 et σ_2 sont connus, on a :

  $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n_1})$ et $\bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{n_2})$

  $\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

  Donc sous H_0, $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\mathbb{V}ar(\bar{X} - \bar{Y})}} = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$

  Ainsi pour un test de niveau α, on rejette H_0 si :

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}} > z_{1-\frac{\alpha}{2}}$$

- Si σ_1 et σ_2 sont inconnus et $\sigma_1 = \sigma_2 = \sigma$, on a :

  Sous H_0 $\quad \bar{X} - \bar{Y} \sim \mathcal{N}\left[0, \sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\right]$.

  Donc $\dfrac{\bar{X} - \bar{Y}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim \mathcal{N}(0, 1).$

Puisque $\frac{(n_1+n_2-2)S^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$, on a :

$$\frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

Pour un test de niveau α, on rejette H_0 si

$$\frac{|\bar{X}-\bar{Y}|}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2,1-\frac{\alpha}{2}}$$

- Cas où $\sigma_1 \neq \sigma_2$ (obtenir par le test de Fisher)
  On souhaite tester $H_0 = \{\mu_1 = \mu_2\}$ contre $H_1 = \{\mu_1 \neq \mu_2\}$
  — Le test de Fisher permet de vérifier d'abord si $\sigma_1 = \sigma_2$ ou pas. Si le test de Fisher conduit au rejet de l'hypothèse $\{\sigma_1 = \sigma_2\}$, on procède à une correction de degré de liberté.
  — Statistique de test

$$T_n = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \underset{approx}{\sim} \mathcal{T}(\nu) \quad sous\ H_0$$

Car $\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2} \underset{approx}{\sim} \chi^2(\nu)$ avec

$$\nu = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}}$$

Pour un test approximatif de niveau α :
RD : rejeter H_0 si $|T_n| > t_{\nu,1-\frac{\alpha}{2}}$

Note :

Si $S_1^2, \cdots, S_K^2$ sont des estimateurs de variance $\sigma_1^2, \cdots, \sigma_K^2$ respectivement à partir des échantillons indépendants de taille respectives $n_1, \cdots, n_K$. Pour tous nombres réels $a_1, \cdots, a_K$, la statistique

$$L = \frac{\nu \sum_{k=1}^K a_k S_k^2}{\sum_{k=1}^K a_k \sigma_k^2}$$

suit approximativement la loi de χ^2 à ν ddl avec :

$$\nu = \frac{\left(\sum_{k=1}^{K} a_k S_k^2\right)^2}{\sum_{k=1}^{K} \frac{(a_k S_k^2)^2}{n_k - 1}}$$

6.5.3 Test sur données appariées

On mesure deux variables X et Y sur une même unité statistique ou sur deux unités statistiques appariées.

Exemple 6.5.2. .

- *rendement de deux variétés de blé cultivées sur la même parcelle ou sur des parcelles adjacentes.*

- *L'effet de deux traitements A et B sur le même sujet, mais à deux instants différents.*

Soit X_i et Y_i les mesures effectuées sur l'unité commune i ou sur deux unités statistique appariées. On a :

- $X_1, \cdots, X_{n_1}$ sont des v.a d'espérance μ_1 et $Y_1, \cdots, Y_{n_2}$ sont des v.a d'espérance μ_2.

- On note $D_i = X_i - Y_i$ et on suppose que ces variables sont iid avec $D_i \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2)$.
  Soit $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ et $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \quad (n_1 = n_2 = n)$
  On souhaite tester $H_0 = \{\mu_1 = \mu_2\}$ contre $H_1 = \{\mu_1 \neq \mu_2\}$, i.e $H_0 = \{\mu_1 - \mu_2 = 0\}$ contre $H_1 = \{\mu_1 - \mu_2 \neq 0\}$
  Sous $H_0 \quad \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} \sim t(n-1)$.
  RD : rejeter H_0 si $|\bar{D}| > \frac{S_D}{\sqrt{n}}.t_{\,n-1,1-\frac{\alpha}{2}}$

6.5.4 Test de $H_0 = \{\sigma_1 = \sigma_2\}$ contre $H_1 = \{\sigma_1 \neq \sigma_2\}$

Deux variances empiriques permettent de bâtir ce test. En cas d'égalité des deux variances σ_1^2 et σ_2^2, les deux variances empiriques sont très proches, donc la statistique $\frac{S_1^2}{S_2^2}$ est très proche de 1. Si cette statistique est trop grande ou trop

petite, on rejette H_0.
On sait que $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$ et $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$.
Donc sous H_0 $\quad \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \sim \mathcal{F}(n_1-1, n_2-1)$.
RD : rejeter H_0 si $\frac{S_1^2}{S_2^2} > f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ ou $\frac{S_1^2}{S_2^2} < f_{n_1-1,n_2-1,\frac{\alpha}{2}}$
où $f_{n,n',p}$ est le quantile d'ordre p de la loi de Fisher à n et n' ddl.

6.5.5 Test de comparaison de deux proportions

Test de $H_0 = \{p_1 = p_2\}$ contre $H_1 = \{p_1 \neq p_2\}$
$X \sim \mathcal{B}(n_1, p_1)$ et $Y \sim \mathcal{B}(n_2, p_2)$
Soit $P = \frac{X+Y}{n_1+n_2}$ et $S_d^2 = P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.
On sait que

$$\frac{X}{n_1} - \frac{Y}{n_2} \underset{approx}{\sim} \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Sous H_0 si $p_1 = p_2 = p$, on a : $\frac{X}{n_1} - \frac{Y}{n_2} \underset{approx}{\sim} \mathcal{N}\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$
Un estimateur de p est donné par $P = \frac{X+Y}{n_1+n_2}$.
Pour un test approché de niveau α, on rejette H_0 si

$$\frac{\left|\frac{X}{n_1} - \frac{Y}{n_2}\right|}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_{1-\frac{\alpha}{2}}$$

Chapitre 7

Tests sur variables qualitatives : test du χ^2

7.1 Contexte

Soit X une variable aléatoire qualitative à c modalités. On s'intéresse à la distribution des c modalités de la variable X dans une population donnée. Pour $j = 1, \cdots, c$, on pose $p_j = \mathbb{P}(X = j)$ la probabilité que la variable X prenne la modalité j. On a $\sum_{j=1}^{c} p_j = 1$. Cette population est donc caractérisée par $c - 1$ paramètres libres.
On choisit dans cette population, un échantillon de n individus et on note $N_1, \cdots, N_c$ les variables aléatoires correspondant aux fréquences d'apparition des c modalités. La loi conjointe de $N_1, \cdots, N_c$ est une loi multinomiale et on a :

$$\mathbb{P}(N_1 = n_1, \cdots, N_c = n_c) = \begin{cases} \dfrac{n!}{\prod_{j=1}^{c} n_j} \prod_{j=1}^{c} p_j^{n_j!} & si \ \sum_{j=1}^{c} n_j = n \\ 0 \ sinon \end{cases}$$

Les $N_1, \cdots, N_c$ ne sont pas $\perp$ (car $\sum_{j=1}^{c} N_j = n$).

7.2 Exemples de test du χ^2

7.2.1 Adéquation à une loi multinomiale p_0

On souhaite tester $H_0 : \{p_1 = p_{01}, \cdots, p_c = p_{0c}\}$ contre $H_1 : \ \exists j \in \{1, \cdots, c\}; \ p_j \neq p_{oj}$.

On sait que :

$$\mathbb{P}(N_1 = n_1, \cdots, N_c = n_c) = \frac{n!}{n_1! n_2! \cdots n_c!} p_1^{n_1} \cdots p_c^{n_c}$$

si $\sum_{j=1}^{c} n_j = n$ et 0 sinon. $N_1, \cdots, N_c$ ne sont pas indépendantes car leur somme doit être égale à n. On s'intéresse à l'hypothèse $H_0 : \{p_1 = p_{01}, \cdots, p_c = p_{0c}\}$ où les p_{0j} sont des valeurs de probabilités spécifiées telles que $\sum_{j=1}^{c} p_{0j} = 1$. L'hypothèse alternative étant qu'il existe au moins une catégorie j telle que $p_j \neq p_{0j}$.
En utilisant une approximation gaussienne, on montre que :

$$Q = \sum_{j=1}^{c} \frac{(N_j - np_{0j})^2}{np_{0j}} \xrightarrow[n \to +\infty]{Loi} \chi^2(c-1) \quad \textit{sous } H_0$$

C'est pour quoi Q est appelée statistique du Khi-deux (ou statistique de Pearson).

Remarque 7.2.1. .

Sous H_0, $\dfrac{N_j - np_{0j}}{\sqrt{np_{0j}(1-p_{0j})}}$ est la variable centrée réduite de N_j suivant la loi $\mathcal{B}(n, p_{0j})$.
Asymptotiquement, cette variable aléatoire suit la loi $\mathcal{N}(0,1)$ et son carré est un $\chi^2(1)$.
Intuitivement, on voit que la valeur prise par Q est d'autant plus petite que les fréquences observées sont proches des np_{oj} appelées fréquences attendues(ou fréquences théoriques) sous H_0. La règle de décision consiste donc à rejeter H_0 pour les grandes valeurs de réalisation q de Q.
***RD :** Rejeter H_0 si $q > \chi^2_{1-\alpha}(c-1)$ pour un test de niveau asymptotique α.*

7.2.2 Test de χ^2 d'indépendance de deux variables catégorielles (ou qualitatives)

On considère un couple de variables aléatoires qualitatives l'une à I catégories (ou modalités), l'autre à J catégories observables sur toute unité statistique sélectionnée dans un échantillon. Le croisement des deux variables donne lieu à une variable qualitative à $I \times J$ modalités avec $I \times J - 1$ paramètres libres. A la modalité obtenue par croisement des modalités i et j respectives de chaque variable, on associe la probabilité p_{ij}. On a donc $\sum_{j=1}^{J} \sum_{i=1}^{I} p_{ij} = 1$.
On montre que l'indépendance des deux variables qualitatives est équivalente à

l'indépendance de chaque modalité de l'une avec toutes les modalités de l'autre. Donc tester l'hypothèse H_0 : "Les deux variables sont indépendantes" revient à tester $H_0 : p_{ij} = p_{i.}p_{.j}$ *pour* $i = 1, \cdots, I$ *et* $j = 1, \cdots, J$.
L'hypothèse alternative est qu'il existe au moins un coupe (i, j) de modalités tel que $p_{ij} \neq p_{i.}p_{.j}$.
Pour un échantillon aléatoire de taille n, on observe les fréquences au croisement des deux variables résumées dans le tableau de contingence suivant :

	1	$\cdots$	j	$\cdots$	J	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

En utilisant la statistique Q de Pearson obtenue en estimant les fréquences attendues sous H_0 par le maximum de vraisemblance, soit $n\widehat{p}_{i.}\widehat{p}_{.j} = n\dfrac{N_{i.}}{n}\dfrac{N_{.j}}{n} = \dfrac{N_{i.}N_{.j}}{n}$, on a

$$Q = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(N_{ij} - \dfrac{N_{i.}n_{.j}}{n}\right)^2}{\dfrac{N_{i.}n_{.j}}{n}}$$

Sous $H_0 \quad Q \underset{n\to+\infty}{\overset{Loi}{\to}} \chi^2_{((I-1)(J-1))}$.
La règle de décision consiste donc à rejeter H_0 si la réalisation q de Q est telle que

$$q = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(n_{ij} - \dfrac{n_{i.}n_{.j}}{n}\right)^2}{\dfrac{n_{i.}n_{.j}}{n}} > \chi^2_{((I-1)(J-1)),1-\alpha}.$$

Remarque 7.2.2. .
Cette procédure est identique à celle du test de comparaison de plusieurs lois multinomiales.