Mensa ACADEMY

IFORD

INSTITUT DE FORMATION & DE RECHERCHE DÉMOGRAPHIQUES

STATISTIQUE - PROBABILITÉS

Cours
Anciens sujets

www.mensa-academy.com

Mensa
ACADEMY

IFORD

INSTITUT DE FORMATION & DE RECHERCHE DÉMOGRAPHIQUES

STATISTIQUE - PROBABILITÉS

+237 655 333 073
+237 671 544 122

Yaoundé

Dédicaces & Remerciements

À

Tous les lauréats des sessions antérieures : Élèves Démographes à l'IFORD depuis l'existence de MENSA Academy. Vous faites notre fierté et êtes la preuve vivante que seul le travail paye.

Les auteurs remercient l'Entreprise MENSA Academy constituée d'une équipe très efficace qui a permis à ce que ce livre soit publié dans plusieurs Maisons d'Édition à l'instar de Horizon + & Optimum, d'avoir mis à la disposition des autres pays africains quelques exemplaires.

La perfection n'étant pas de ce monde, toute suggestion/erreur relevée dans ce livre est la bienvenue. Vous pourrez le faire par **mail : louishenri2018@gmail.com** *ou via le moyen téléphonique suivant - WhatsApp :* ***+237 655 333 073****. De ce fait, le message doit se présenter comme suit :* ***Suggestion(s)/erreur(s) à signaler - Stat & Proba Édition 2026 *** Vous vous présentez et enfin, vous indiquez la page, la suggestion ou l'erreur en question*** **** Toute question est également la bienvenue.*

MENSA Academy
1556 Yaoundé II, IFORD
294 Yaoundé, ISSEA - Rue Pasteur
08 BP 03 Abidjan, ENSEA
45512 Dakar, ENSAE
03 BP 1079 Cotonou, ENEAM
1096 ASECNA, Niamey
2067 EAMAU, Lomé
www.mensaacademy.com

Dans la même collection

Économie
Culture Générale,
Mathématiques ISE
Statistiques descriptives,
Statistiques-Probabilités,
Mathématiques pour économistes

ISBN 13 : x xxxxxxx xxxxxx
ISBN 10 : y yyy yyy yyy
BookEditions
Horizon +
Optimum
MENSA Academy

Avant-Propos

Chaque année, de centaines d'étudiants sont confrontés à des difficultés liées à leur préparation au concours de l'IFORD et ce, pour quelques raisons dont la principale est la non-exhaustivité des cours et la quasi-absence de certaines épreuves passées et évidemment leurs corrigés. MENSA Academy met donc à votre disposition ce document qui va retracer les chapitres importants pour braver son épreuve de statistique et probabilités au concours. Il présente un condensé de cours des chapitres phares, des résumés de certains chapitres et bien-sur, explicite les anciens sujets en statistique et probabilité pour les voies A et B. Cependant, le chapitre sur la statistique descriptive (caractéristiques de tendance centrale, de dispersion et de dispersion relative pour l'analyse univariée et bivariée) n'est pas présentée ici. Pour ceux qui feront les cours, vous aurez évidemment les notes de cours détaillant ligne par ligne ce chapitre, clé pour le concours de l'IFORD puisqu'à ce niveau, les interprétations ont tout leur sens ■

Félicitations aux Lauréats !

1. ISE long/AS (2024)

AGHOKENG Sandra (ISSEA)
BETSI Marc (ISSEA)
DASEBO Etienne (ENSEA)
DJAM A BITANG (ISSEA)
EBALA Quentin (ISSEA)
FIRIDA Aimée Brigitte (ISSEA)
FOTIO Yan Mayel (ISSEA)
GAPILI Kadi (ISSEA)
KAZE Ariane Laure (ENSEA)
LEMA Léocadie (ISSEA)
MBIDA William (ISSEA)
NANTIA Mégane (ENSEA)
NDJOCKO Gaston (ISSEA)
NOUMEN Yvan Rayan (ISSEA)
NZAMBOU Sagesse (ISSEA)
TAMETSA Anderson Bryan (ISSEA)
TIENGA Marc Jorel (ISSEA)
TOUMBA Pascal (ENSEA)
YEMELI Horlus Boban (ISSEA)
YEMELI SAAH Crespo (ENSAE)
YETNA OTTO Polycarpe (ISSEA)
ZEETSOP Boslin (ENSEA)
ZEUNANG Fredy (ENSEA)

2. ISE Option Eco (2024)

ATEBA ONDOUA Aristide (ENSEA)
EWOUNDJO Christian (ISSEA)
DIMI Firsule (ENSEA)
MALGOUBRI Abdou-Aziz (ENSEA)
NGOLLONG Franck Nelson (ENSEA)
NOGA Dilane Cabbel (ISSEA)

3. ISE Option Maths (2024)

FOUDJIN Karlane (ENSEA)
KUATE Cabrel (ISSEA)
MAKPUNE Andre (ISSEA)
NTONGME Arsène (ENSEA)
PAHANE Kerencia (ENSAE)
PELAP Franck Axel (ISSEA)
SOLANGE Difo (ISSEA)
TCHUIDJANG Jorexe (ENSEA)

4. IFORD (2024)

AVOM Paul Antoine (3ème)
KUATE Cabrel (7ème)
NGUIMATSIA Henri.

1. ISE Long/AS (2023)

ABENA BALA Marc Loïc (ISSEA)
AGNANGMA David Landry (ENSAE)
AKONO Joseph Ariel (ISSEA)
DAIFFERLE Marianne (ENSAE)
DONKENG Manuelle Idrisse (ENSEA)
HASSANA Mohamadou (ISSEA)
MFOYET bdellatif (ISSEA)
MOGUEM TABUE Erina Lucelle (ISSEA)
MOUALA NGABILANG Lucrèce (ENSEA)
NGUEAJIO DONGMO Franck (ENSAE)
NKWA TSAMO Leslye Patricia (ENSAE)
ZE II Jonathan Patrick (ISSEA)

2. ISE Option Eco (2023)

ATEBA Joseph Stève (ISSEA)
IKIA AMBOYA Grace (ISSEA)
NJIFON MOUNCHINGAM (ISSEA)
NYANGONO AKONO Jenny (ISSEA)
ONANA ABENA Valery (ENSEA)
SEYDOU Ferdinand (ISSEA)

3. ISE Option Maths (2023)

AWADJOU Rodrigue Pavel (ISSEA)
KENMOGNE Marelle Ingrid (ENSEA)
KEOUL Maab Mara (ISSEA)

MAGUETSWET Rivalien (ISSEA)
MATANG KUETE Josette (ENSAE)
NGUIMGO Mikadeau (ISSEA)
YABOYA Anicet Marius (ISSEA)

4. IFORD (2023)

AWADJOU Rodrigue Pavel (1er)
CAROLINE EBONGUE (Voie A)
KEOUL MAAB MARA
MBENE NGUEDE Jessica (6ème)
METOUKE ADIANG Vidal
YABOYA Anicet Marius (7ème)

1. ISE Long/AS (2022)

DITCHIBE Bertrand (ISSEA)
FARAITINI Christ Justin (ISSEA)
KAPI TCHIAWOU Blanche (ISSEA)
KENFACK Joséphine Elza (ISSEA)
MBA-NZE Stéphane Emmanuel (ISSEA)
NGOUATIO Maguy Bricelle (ISSEA)
NGOUDJO DEUMAGA Kell (ENSEA)
TOUKAM LOWE Sandra (ISSEA)

2. ISE Option Eco (2022)

MISSENGUE Exaucée (ISSEA)
MOSSE Isabelle Danielle (ENSAE)
NGAH KEDE Armel Bruno (ISSEA)
SHERIFFA MVUH (ISSEA)
TAGNE Rinel Valdy (ISSEA)

Table des matières

Chapitre 1

Présentation générale de l'IFORD

L'Institut de Formation et de Recherche Démographiques est un centre d'excellence pour la formation professionnelle des cadres africains et l'appui aux États membres et aux organismes partenaires dans le domaine des sciences de la population et du développement. A ce titre, la formation en Master Professionnel en Démographie à l'IFORD contribue d'une part au renforcement des capacités et des compétences des cadres africains et des administrations afin qu'ils assurent mieux l'élaboration, l'exécution, le suivi et l'évaluation des programmes et projets en matière de population et développement. D'autre part, elle apporte une contribution efficace et efficiente aux cadres stratégiques (nationaux et internationaux) actuels de lutte contre la pauvreté, pauvreté qui touche particulièrement les jeunes. En 40 ans d'existence, l'IFORD a formé près de 800 démographes.

1.1 Conditions d'entrée à l'IFORD

L'entrée à l'IFORD se fait sur concours. A cet effet, un concours international est organisé chaque année dans les 23 pays membres. Il faut noter toutefois qu'un pays peut ne pas éprouver le besoin d'organiser le concours. L'organisation matérielle du concours est assurée généralement par la Direction nationale de la Statistique ou toute autre institution désignée par le pays desservi. Les candidats, quel que soit leur pays d'origine, composent dans les centres ouverts dans leur pays de résidence. Deux types de concours sont organisés et les diplômes requis diffèrent selon le type de concours. Le concours **type A** est destiné aux candidats titulaires d'une licence dans les filières suivantes :

- démographie ;
- Géographie ;
- Sociologie ;
- Anthropologie.

Le concours **type B** quant à lui est ouvert aux candidats titulaires d'un Diplôme d'Ingénieur des Travaux Statistiques ou d'une licence dans les filières suivantes :

— Sciences Économiques ;
— Statistiques ;
— Mathématiques ;
— Informatique.

Peuvent également faire acte de candidature, pour chaque type de concours, les candidats titulaires de tout autre diplôme jugé équivalent par la Direction Exécutive de l'Institut. Le concours porte sur trois épreuves d'une durée de 4 heures chacune et la note de chaque épreuve compte pour 1/3 : une épreuve de culture générale commune à tous les candidats (concours type A et type B), une épreuve de mathématiques et une épreuve de probabilités et statistique pour les candidats de chaque type de concours.

L'admission des candidats au concours est prononcée par un jury qui travaille dans la sérénité et délibère en toute objectivité et seuls les meilleurs candidats sont retenus. L'admission définitive est conditionnée par l'obtention d'une bourse d'études qui peut être financée directement par les gouvernements des pays membres ou indirectement par les organismes tels l 'UNFP A, l'UNICEF, l'ACBF, la Banque Mondiale, la Coopération Française, la Coopération Belge, l'Union Européenne, etc.

1.2 Les débouchés

A leur sortie de l'IFORD, diverses portes sont ouvertes aux Jeunes professionnels démographes :

— Instituts nationaux de statistique ;
— Fonction Publique ;
— Universités t Instituts de Recherche ;
— Organismes Internationaux et ONGs ;
— Toutes structures travaillant dans le domaine de la population et des ressources humaines ;
— Secteurs privés ;
— etc.

1.3 Le programme de culture générale (Types A & B)

L'épreuve de culture générale porte sur un sujet d'ordre général ne nécessitant pas pour le candidat de faire preuve de connaissances techniques particulières. Le candidat aura le choix entre deux sujets : une dissertation ou un commentaire de texte portant sur un sujet d'ordre général touchant aux problèmes de développement.

L'évaluation du candidat est basée sur sa capacité d'analyse et d'argumentation ainsi que sur le degré de connaissance qu'il a des problèmes de développement. Sont aussi prises en compte les aptitudes du candidat à bien rédiger la langue française. Il est recommandé au candidat, pour la préparation de cette épreuve, de lire les ouvrages généraux sur l'actualité africaine et mondiale relative à ces problèmes.

1.4 Le programme de l'épreuve de probabilités et statistique (Concours B)

1.4.1 Statistique descriptive

1. Objet de la statistique descriptive, unités statistiques, caractères qualitatifs, caractères quantitatifs, variables statistiques discrètes, variables statistiques continues.
2. Distributions statistiques à un caractère : tableaux statistiques, représentation graphique.
3. Description numérique d'une variable statistique : caractéristiques de tendance centrale (médiane, mode, moyenne) ; caractéristiques de dispersion (différences, écarts, écart quadratique moyen, quartiles, moments centrés, moments non centrés) ; caractéristiques de forme (coefficient d'asymétrie, coefficient d'aplatissement) ; caractéristiques de concentration (courbe de concentration, indice de concentration médiale).
4. Ajustement d'une distribution observée à une distribution théorique (loi binomiale, loi de Poisson, loi gamma, loi normale, loi lognormale, loi de Pareto).
5. Distributions statistiques à deux caractères : tableaux statistiques, distributions marginales, distributions conditionnelles, indépendance, liaison fonctionnelle, représentation graphique, papiers fonctionnels.
6. Description numérique des séries statistiques à deux caractères quantitatifs ; distributions marginales et conditionnelles, moyennes et variances marginales, moyennes et variances conditionnelles, courbes de régression, rapport de corrélation, coefficient de corrélation linéaire, principe de l'ajustement linéaire et droite des moindres carrés.
7. Séries chronologiques : composantes d'une série chronologique, méthode analytique et méthodes empiriques d'analyse d'une série chronologique.

1.4.2 Probabilités

1. Analyse combinatoire : arrangement avec et sans répétition, permutations avec et sans répétition, combinaisons avec et sans répétition.
2. Notion de probabilité : événements, espace de probabilités, mesure de probabilité.

3. Axiome des probabilités totales, axiome des’ probabilités composées (probabilité conditionnelle, indépendance entre événements) ; théorème de Bayes.
4. Les schémas de tirages probabilistes : tirage exhaustif, tirage bernouillien, notion d’échantillon.
5. Variables aléatoires : variables aléatoires discrètes à une et deux dimensions, variables aléatoires continues à une ou deux dimensions, caractéristiques d’une variable aléatoire (moments centrés, moments non centrés), indépendance, liaison fonctionnelle, corrélation, décomposition de la variance.
6. Fonctions génératrices des moments.
7. Principales lois d’usage courant : lois à une dimension (loi uniforme discrète, loi uniforme continue, loi de Bemoulli, loi binomiale, loi de Poisson, loi gamma, loi normale, loi log normale, loi béta, loi du χ^2, loi de Fisher-Snedecor, loi de Student), loi normale à deux dimensions.
8. Fonctions de variables aléatoires : fonction d’une variable aléatoire, fonction de plusieurs variables aléatoires, addition de variables aléatoires.
9. Convergences stochastiques et applications : inégalité de Bienaymé-Tchebychev, convergence en loi, convergence en probabilité, loi faible des grands nombres, théorème central limite.

1.4.3 Statistique mathématique

1. Estimation ponctuelle : estimation d’un ou plusieurs paramètres d’une loi de probabilité ; estimation d’une ou de plusieurs caractéristiques d’une population finie.
2. Estimation par intervalles.
3. Test d’adéquation du χ^2.

1.5 Le programme de Mathématiques (Concours B)

1.5.1 Éléments de la théorie des ensembles

1. Ensemble : inclusion, intersection, réunion, partie ou sous- ensemble, ensemble des parties d’un ensemble, différence, ensemble produit, partition d’un ensemble.
2. Relations binaires : définition, propriétés possibles d’une relation binaire, relations d’ordre et d’équivalence, classes d’équivalence, ensemble ordonné.
3. Applications : définition, injection, surjection, bijection.

4. Lois de composition : définition, propriétés possibles d'une loi de composition interne (commutativité, associativité, élément neutre, élément symétrique, distributivité d'une loi par rapport à une autre) ; loi de composition externe ; groupe ; anneau ; corps.

1.5.2 Analyse

1. Progressions arithmétique et géométrique.
2. Notions sommaires sur la structure de R ; notion de valeur absolue.
3. Nombres complexes ; formule de Moivre.
4. Suites et séries : règles classiques de convergence.
5. Fonctions réelles d'une variable réelle : continuité, limites dérivées, différentielles ; théorème des accroissements finis ; formule de Taylor ; formule de Mac-Laurin.
6. Principales fonctions réelles d'une variable réelle : fonction puissance, fonction logarithme, fonction exponentielle.
7. Fonctions circulaires et principaux résultats de trigonométrie.
8. Développements limités au voisinage d'un point, développements limités à l'infini.
9. Étude de la variation d'une fonction réelle d'une variable et construction des courbes représentatives, extremum, point d'inflexion, asymptote parabolique.
10. Calcul intégral :
    (a) Intégrale définie : méthodes classiques du calcul intégral, interprétation géométrique ;
    (b) Intégrale généralisée : cas où. la fonction à intégrer devient infinie ; principaux critères de convergence.
11. Fonctions réelles de plusieurs variables réelles :
    (a) Limites et continuité (notions sommaires) ;
    (b) Différentielles, dérivées partielles ; élasticités, lignes de niveau ou courbes d'indifférence ;
    (c) Extremum d'une fonction réelle de plusieurs variables réelles, multiplicateurs de Lagrange ;
    (d) Intégrales doubles.

1.5.3 Algèbre linéaire

1. Espaces vectoriels.
2. Applications linéaires.
3. Matrices : définition, produit d'une matrice par un scalaire, somme et produit de matrices.

4. Déterminants, inversion d'une matrice régulière.
5. Résolution des systèmes d'équations linéaires (résolution matricielle).
6. Matrices carrées, diagonales, triangulaires, symétriques, antisymétriques.
7. Puissances successives d'une matrice carrée. Applications à la résolution des équations linéaires récurrentes.
8. Valeurs propres, vecteurs propres ; diagonalisation des matrices carrées ; formes quadratiques.

1.6 Le programme de l'épreuve de probabilités et statistique (Concours A)

1. Espaces probabilisés finis ; axiomes des probabilités, indépendances entre événements ; théorème de Bayes.
2. Schémas de tirage avec remise et sans remise.
3. Variable aléatoire réelle, discrète, finie : définition, fonction de répartition, loi de Bernouilli, loi binomiale.
4. Couple de variables aléatoires réelles, discrètes, finies : loi du couple, lois marginales, lois conditionnelles ; indépendance des deux variables du couple.
5. Espérance mathématique d'une variable aléatoire ; propriétés.
6. Variance, écart-type d'une variable aléatoire.
7. Description statistique d'une population ou d'un échantillon : documents statistiques, représentations graphiques, effectifs, fréquences, moyenne, écart-type.

1.7 Le programme de Mathématiques (Concours A)

1.7.1 Équations et inéquations

1. Équation et inéquation du premier degré à une inconnue dans $\mathbb{R}$.
2. Système de deux équations du premier degré à deux inconnues ; déterminant.
3. Racine carrée d'un nombre positif ou nul ;
4. Équation du second degré à une inconnue dans $\mathbb{R}$, somme et produit des racines.
5. Signe du trinôme du second degré ; position d'un nombre par rapport aux zéros d'un trinôme du second degré ; inéquations du second degré.

1.7.2 Éléments de la théorie des ensembles

1. Ensembles : inclusion, intersection, réunion, partie ou sous-ensemble, ensemble des parties d'un ensemble, ensemble produit, partition d'un ensemble.
2. Relations binaires : définition, propriétés possibles, relations d'ordre et d'équivalence.
3. Applications : définition, injection, surjection, bijection.
4. Lois de composition interne : définition, commutativité, associativité, élément neutre, élément symétrique, distributivité d'une loi par rapport à l'autre.

1.7.3 Fonctions numériques d'une variable réelle

1. Limite, continuité d'une fonction.
2. Dérivée d'une fonction en un point ; dérivée d'une fonction composée de deux fonctions dérivables ; application à l'étude du sens de variation ; représentation graphique.
3. Exemples de fonctions numériques d'une variable réelle.
4. Primitive d'une fonction : définition, propriétés ; application ; calcul d'aire.

Chapitre 2

Lois de probabilité

2.1 Résumé des lois de probabilité

2.1.1 Lois discrètes

Distribution de Bernoulli

- Loi de probabilité : $\mathbb{P}(X = 0) = q$, $\mathbb{P}(X = 1) = p$ où $q = 1 - p$.
- Espérance : $\mathbb{E}(X) = p$; Variance : $\mathrm{Var}(X) = pq$.
- Fonction génératrice : $\mathbb{E}(z^X) = pz + q$

Distribution Binomiale : $\mathcal{B}(n, p)$

- Loi de probabilité : $\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}$, où $q = 1 - p$ et $k = 0, ..., n$.
- Espérance : $\mathbb{E}(X) = np$; Variance : $\mathrm{Var}(X) = npq$.
- Fonction génératrice : $\mathbb{E}(z^X) = (pz + q)^n$

Distribution de Poisson ($\mathcal{P}(\lambda)$)

- Loi de probabilité : $\mathbb{P}(X = k) = \mathrm{e}^{-\lambda} \dfrac{\lambda^k}{k!}$, avec $k = 0, 1, ...$
- Espérance : $\mathbb{E}(X) = \lambda$; Variance : $\mathrm{Var}(X) = \lambda$.
- Fonction génératrice : $\mathbb{E}(z^X) = \mathrm{e}^{\lambda(z-1)}$.

Distribution Géométrique : $\mathcal{G}(p)$

- Loi de probabilité : $\mathbb{P}(X = k) = pq^{k-1}$, où $q = 1 - p$ et $k = 1, ..., n$.
- Espérance : $\mathbb{E}(X) = \dfrac{1}{p}$; Variance : $\mathrm{Var}(X) = \dfrac{q}{p^2}$.
- Fonction génératrice : $\mathbb{E}(z^X) = \dfrac{pz}{1 - qz}$.

Distribution Hypergéométrique : $\mathcal{H}(N, n, p)$

— Loi de probabilité : $\mathbb{P}(X = k) = \dfrac{\binom{N_p}{k}\binom{N_q}{n-k}}{\binom{N}{n}}$, où $q = 1 - p$ et $\max(0, n - N_q) \leqslant k \leqslant \min(N_p, n)$.

— Espérance : $\mathbb{E}(X) = np$; Variance : $\mathrm{Var}(X) = npq\dfrac{N-n}{N-1}$.

— Fonction génératrice : $\mathbb{E}(z^X) = \dfrac{\binom{N_q}{n}}{\binom{N}{n}} F(-n, -N_p; N_q - n + 1; z)$.

Distribution Binomiale négative

— Loi de probabilité : $\mathbb{P}(X = k) = \binom{k+r-1}{r-1} p^r q^k$, où $q = 1 - p$ et $k = 0, 1, ...$.

— Espérance : $\mathbb{E}(X) = \dfrac{rq}{p}$; Variance : $\mathrm{Var}(X) = \dfrac{rq}{p^2}$.

— Fonction génératrice : $\mathbb{E}(z^X) = \left(\dfrac{p}{1 - qz}\right)^r$

Distribution de Pascal

— Loi de probabilité : $\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$, où $q = 1 - p$ et $k = r, r + 1, ...$

— Espérance : $\mathbb{E}(X) = \dfrac{r}{p}$; Variance : $\mathrm{Var}(X) = \dfrac{rq}{p^2}$.

— Fonction génératrice : $\mathbb{E}(z^X) = \left(\dfrac{pz}{1 - qz}\right)^r$

Remarques

- Fonction hypergéométrique : $F(a, b; c; z) = \displaystyle\sum_{n=0}^{+\infty} \frac{a(a+1)...(a+n-1)...(b+n-1)}{c(c+1)...(c+n-1)} \frac{z^n}{n!}$.
- La somme de n v.a. indépendantes suivant la loi de Bernoulli de paramètre p suit une loi binomiale de paramètres n et p : $\mathcal{B}(n, p)$.
- La somme de deux v.a. indépendantes suivant les lois binomiales $\mathcal{B}(m, p)$ et $\mathcal{B}(n, p)$ suit la loi binomiale : $\mathcal{B}(m + n, p)$.
- La somme de deux v.a. indépendantes suivant les lois de Poisson $\mathcal{P}(\lambda)$ et $\mathcal{P}(\mu)$ suit la loi de Poisson : $\mathcal{P}(\lambda + \mu)$.
- La somme de deux v.a. indépendantes suivant les lois binomiales négatives de paramètres (r, p) et (s, p) suit la loi binomiale négative de paramètres $(r + s, p)$.
- La somme de r v.a. indépendantes suivant la loi géométrique $\mathcal{G}(p)$ suit la loi de Pascal de paramètres (r, p).

2.1.2 Lois absolument continues

Distribution uniforme : $\mathcal{U}(a, b)$

— Loi de probabilité : $\dfrac{1}{b-a}\mathbb{1}_{[a,b]}(x)$.

— Espérance : $\mathbb{E}(X) = \dfrac{a+b}{2}$; Variance : $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$.
— Fonction caractéristique : $\mathbb{E}(\mathrm{e}^{itX}) = \dfrac{\mathrm{e}^{bt} - \mathrm{e}^{iat}}{i(b-a)t}$.

Distribution Exponentielle : $\mathcal{E}(\lambda)$

— Loi de probabilité : $\lambda \mathrm{e}^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$.
— Espérance : $\mathbb{E}(X) = \dfrac{1}{\lambda}$; Variance : $\mathrm{Var}(X) = \dfrac{1}{\lambda^2}$.
— Fonction caractéristique : $\mathbb{E}(\mathrm{e}^{itX}) = \dfrac{\lambda}{\lambda - it}$.

Distribution Normale : $\mathcal{N}(m, \sigma^2)$

— Loi de probabilité : $\dfrac{1}{\sqrt{2\pi\sigma}} \exp\left(-\dfrac{(x-m)^2}{2\sigma^2}\right)$.
— Espérance : $\mathbb{E}(X) = m$; Variance : $\mathrm{Var}(X) = \sigma^2$.
— Fonction caractéristique : $\mathbb{E}(\mathrm{e}^{itX}) = \mathrm{e}^{imt - \frac{1}{2}\sigma^2 t^2}$.

Distribution de Weibull : $\mathcal{W}(\lambda, a)$

— Loi de probabilité : $\lambda a x^{a-1} \mathrm{e}^{-\lambda x^a} \mathbb{1}_{]0,+\infty[}(x)$.
— Espérance : $\mathbb{E}(X) = \lambda^{-\frac{1}{a}} \Gamma\left(\dfrac{1}{a} + 1\right)$; Variance : $\mathrm{Var}(X) = \lambda^{-\frac{2}{a}} \left[\Gamma\left(\dfrac{2}{a} + 1\right) - \Gamma\left(\dfrac{1}{a} + 1\right)^2\right]$.

Distribution de Cauchy : $\mathcal{C}(a, b)$

— Loi de probabilité : $\dfrac{a}{\pi(a^2 + (x-b)^2)}$.
— Espérance : non définie ; Variance : non définie.
— Fonction caractéristique : $\mathbb{E}(\mathrm{e}^{itX}) = \mathrm{e}^{ibt - a|t|}$.

Distribution Gamma : $\Gamma(a, \lambda)$

— Loi de probabilité : $\dfrac{\lambda^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-\lambda x} \mathbb{1}_{]0,+\infty[}(x)$.
— Espérance : $\mathbb{E}(X) = \dfrac{a}{\lambda}$; Variance : $\mathrm{Var}(X) = \dfrac{a}{\lambda^2}$.
— Fonction caractéristique : $\left(\dfrac{\lambda}{\lambda - it}\right)^a$.

Distribution Bêta : $B(a, b)$

— Loi de probabilité : $\dfrac{1}{B(a,b)} x^{a-1} (1-x)^b \mathbb{1}_{]0,1[}(x)$.
— Espérance : $\mathbb{E}(X) = \dfrac{a}{a+b}$; Variance : $\mathrm{Var}(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$.
— Fonction caractéristique : $M(a, a+b; it)$.

Distribution du Khi-deux : $\chi^2(n)$

— Loi de probabilité : $\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\mathbb{1}_{]0,+\infty[}(x)$.

— Espérance : $\mathbb{E}(X)=n$; Variance : $\mathrm{Var}(X)=2n$.

— Fonction caractéristique : $(1-2it)^{-\frac{n}{2}}$.

Distribution de Student : $\mathcal{T}(n)$

— Loi de probabilité : $\dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1+\dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$.

— Espérance : 0 si $n>1$; Variance : $\mathrm{Var}(X)=\dfrac{n}{n-2}$ si $n>2$.

— Fonction caractéristique : $\dfrac{2}{\Gamma(n/2)}\left(\dfrac{|t|\sqrt{n}}{2}\right)^{\frac{n}{2}}K_{\frac{n}{2}}\left(|t|\sqrt{n}\right)$.

Distribution de Fisher : $\mathcal{F}(m,n)$

— Loi de probabilité : $\dfrac{m^{\frac{m}{2}}n^{\frac{n}{2}}}{B(\frac{m}{2},\frac{n}{2})}\dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}\mathbb{1}_{]0,+\infty[}(x)$.

— Espérance : $\dfrac{n}{n-2}$ si $n>2$; Variance : $\mathrm{Var}(X)=\dfrac{2n(m+n-2)}{m(n-4)(n-2)^2}$ si $n>4$.

— Fonction caractéristique : $M\left(\dfrac{m}{2};-\dfrac{n}{2};\dfrac{n}{m}it\right)$.

Remarques 1

— Fonction Gamma : $\Gamma(a)=\displaystyle\int_0^{+\infty}x^{a-1}e^{-x}dx$.

— Fonction Bêta : $B(a,b)=\displaystyle\int_0^{1}x^{a-1}(1-x)^{b-1}dx$.

— Fonction de Kummer : $M(a;b;z)=\displaystyle\sum_{n=0}^{+\infty}\dfrac{a(a+1)...(a+n-1)}{b(b+1)...(b+n-1)}\dfrac{z^n}{n!}$.

— Fonction de Bessel modifiée : $K_v(z)=\dfrac{\pi}{2}\dfrac{I_{-v}(z)-I_v(z)}{\sin\pi v}$ où $I_v(z)=\left(\dfrac{z}{2}\right)^v\displaystyle\sum_{n=0}^{+\infty}\dfrac{1}{n!\Gamma(n+v+1)}\left(\dfrac{z^2}{4}\right)^n$.

Remarques 2

— La somme de n v.a. indépendantes suivant la loi exponentielle $\mathbb{E}(\lambda)$ suit la loi Gamma $\Gamma(n,\lambda)$.

— La somme de deux v.a. indépendantes suivant les lois Gamma $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$ suit la loi Gamma $\Gamma(a+b,\lambda)$.

— Si les v.a. indépendantes X et Y suivent les lois Gamma $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$, alors $\dfrac{X}{X+Y}$ suit la loi Bêta : $\mathcal{B}(a,b)$.

— La somme de deux v.a. indépendantes suivant les lois normales $\mathcal{N}(m_1,\sigma_1^2)$ et $\mathcal{N}(m_2,\sigma_2^2)$ suit la loi normale : $\mathcal{N}(m_1+m_2,\sigma_1^2+\sigma_2^2)$.

— Le quotient de deux variables indépendantes suivant la loi normale $\mathcal{N}(0,1)$ suit la loi de

Cauchy $\mathcal{C}(1,0) = \mathcal{T}(1)$.

— La somme des carrés de n v.a. indépendantes suivant la loi normale $\mathcal{N}(0,1)$ suit la loi du Khi-Deux $\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2})$.

— Si les v.a. indépendantes X et Y suivent les lois normale $\mathcal{N}(0,1)$ et du Khi-Deux $\chi^2(n)$, alors $\dfrac{X}{Y/n}$ suit la loi de Student $\mathcal{T}(n)$.

— Si les v.a. indépendantes X et Y suivent les lois du Khi-Deux $\chi^2(m)$ et $\chi^2(n)$ alors, $\dfrac{mX}{nY}$ suit la loi de Fisher $\mathcal{F}(m,n)$.

Approximation

Soit X une v.a de loi $\mathcal{B}(n,p)$ alors si n est assez grand,

$$P(X \leqslant x) \simeq P\left(Y \leqslant x + \frac{1}{2}\right).$$

Où Y est une v.a de loi $\mathcal{N}(np, np(1-p))$.

Ce résultat indique qu'il est possible d'évaluer des probabilités pour une v.a. binomiale en utilisant une table de loi normale. On peut réécrire le résultat directement en fonction de la table de la loi normale :

$$P(X \leqslant x) \simeq P\left(Z \leqslant \frac{x + 0,5 - np}{\sqrt{np(1-p)}}\right).$$

Où $Z \rightsquigarrow \mathcal{N}(0,1)$. Le terme "assez grand" s'interprète en fonction de la valeur de p.

— Si $0,1 \leqslant p \leqslant 0,9$ alors une taille de $n = 20$ est suffisante pour une approximation raisonnable de la probabilité.

— Si $p < 0,1$ il faudra une plus grande taille ,allant jusqu'à plus de 1000 si la valeur p est très petite.

Le terme "$\frac{1}{2}$" est un facteur de correction pour la continuité. Le fait est qu'en utilisant une loi normale pour effectuer une approximation d'une loi binomiale il y a un changement de support : la binomiale a comme support $S_X = \{0, 1, ..., n\}$ tandis que la loi normale a comme support $S_X = \{-\infty, +\infty\}$.

La distribution discrète de la v.a. originale implique que pour chaque valeur du support $P(X = x) > 0$ mais dans un contexte de v.a. continue une des propriétés de ces v.a. est justement que $P(X = x) = 0$. Pour palier ce problème, on considère la correction suivante :

$P(X = x) = P\left(x - \frac{1}{2} \leqslant X \leqslant x + \frac{1}{2}\right)$ lorsqu'on utilise la loi normale. Cela permet d'obtenir une approximation plus près de la réalité.

2.2 Quelques exercices d'application

Exercice 1

On sait que la probabilité qu'une personne choisie au hasard travaille dans le domaine de l'administration ou de la comptabilité est de $\frac{1}{6}$. Si on choisit au hasard 3 personnes, quelle est la probabilité d'avoir au moins 2 personnes sur 3 qui travaillent dans l'administration ou la comptabilité ?

Exercice 2

Dans une entreprise les ressources humaines font passer une entrevue préliminaire aux candidats et on sait par expérience que seulement 50% passent au travers de ce premier tri. Quelle est la probabilité que sur 5 candidats, il y en ait 4 ou plus qui passent la première entrevue ?

Exercice 3

Les données disponibles sur la survie des entreprises démontrent que les nouvelles entreprises du domaine des communications ont une probabilité de passer le cap des 2 ans de 0.20. Si 10 entreprises se sont implantées, quelle est la probabilité d'avoir au moins 4 «survivantes» après 2 ans ?

Exercice 4

Dans l'exemple précédant, si on sait qu'une entreprise en communication qui passe le cap des 2 ans a une probabilité de $\frac{2}{3}$ de devenir une grande entreprise (plus de 50 employés), quelle est la probabilité d'obtenir 4 grandes entreprises en communication sur les 10 qui se sont implantées ?

Exercice 5

Dans une banque les clients arrivent une fréquence moyenne de 10 par heure. Quelle est la probabilité qu'il y ait plus de 2 clients en 10 min ?

Exercice 6

Un cancer rare fait en moyenne 5 victimes par années à Ifordia. Quel est la probabilité que ce cancer fasse plus de 8 victimes en 2009 ?

Exercice 7

Dans la foret tropicale amazonienne il y a en moyenne 0.2 bois serpents (Marmoroxylon racemosum) par ha. C'est un arbre rare qui est très recherché pour la fabrication de bijoux et d'objets décoratifs.

Un exploitant forestier explore 10 ha par jour à la recherche de cet arbre et son exploitation est rentable s'il y a au moins 3 jours par semaine (5 jours) pour lesquels il trouve au moins un bois serpent.

Quelle est la probabilité que son exploitation soit rentable ?

Exercice 8

A une station de métro il y a une rame qui passe à toutes les 10 minutes exactement. Un client qui se présente à a station sans regarder l'heure devra attendre combien de temps ?

Exercice 9

Une centrifugeuse contient une pièce électronique peu fiable qui brise soudainement sans signes précurseurs. De plus, le bris n'est pas détectable sans l'aide d'un technicien puisque la machine fonctionne toujours mais avec une précision moindre. Les inspections de la machine par un technicien se font aux 6 mois. Lors d'une inspection le technicien diagnostique un bris de la pi ce, quelle est la probabilité d'avoir utilisé une machine peu ®able plus de 2 mois ?

Exercice 10

Il y a un tremblement de terre majeur en moyenne aux 150 ans à Ifordia. Quelle est la probabilité qu'il y ait un tremblement de terre majeur dans les prochaines 30 années ?

Exercice 11

La loi exponentielle est aussi une loi permettant de modéliser le temps de vie d'une composante électronique. Une montre digitale a une durée de vie moyenne de 100000 heures. Quelle est la probabilité qu'elle ne fonctionne plus après 5 ans ?

Exercice 12

Soit Z une v.a de loi $\mathcal{N}(0,1)$, évaluer :

$P(Z > 2)$, $P(Z < -2)$, $P(2 < Z)$, $P(-2 < Z < 2)$, $P(1 \leqslant Z < 1,5)$, $P(1 < Z \leqslant 1,5)$, $P(-1,5 < Z < -1)$, $P(-1,5 < Z < 1)$, $P(0,5 > Z)$, $P(Z < -0,5 \text{ ou } Z > 0,5)$, $P(Z < -0,5)$ et $P(Z > 0,5)$ et $P(Z > 3,6)$.

Exercice 13

Un chercheur veut étudier le QI d'une population. L'expérience dans le domaine montre que la v.a. qui donne le QI d'un individu de la population est une v.a. de loi $\mathcal{N}(100, 100)$, quelle est la probabilité d'observer un individu ayant un QI de plus de 120 ?

On veut évaluer la probabilité qu'un individu ait un QI qui se retrouve dans un écart de moins de 10 de la moyenne.

Supposons que le chercheur tire au hasard 10 personnes de la population, quelle est la probabilité d'observer plus de 3 personnes ayant un QI de plus de 110 ?

Exercice 14

Un économiste prédit que le prix de l'essence en été suivra une normale $\mathcal{N}(80, 144)$. On veut obtenir la probabilité que l'essence soit moins de 78 cents.

En reprenant le m !me exemple, on cherche la limite telle que la probabilité d'avoir un prix plus petit est de 20%.

Chapitre 3

Résumé du cours sur les intervalles de confiance et les tests d'hypothèse

Dans la suite, $\overline{x}$ est la moyenne de l'échantillon, f la proportion d'échantillon par rapport à un caractère, μ est la moyenne de la population, p la proportion de la population, M la marge d'erreur, n la taille de l'échantillon, s l'écart-type de l'échantillon et E la marge d'erreur.

Le but d'une estimation par intervalle est de fournir des informations sur l'écart entre l'estimation ponctuelle fournie par l'échantillon et la valeur du paramètre de la population.

3.1 Formules clés - Intervalles de confiance

3.1.1 Moyenne d'une population : cas où σ est connu

$$\mathrm{IC}_{1-\alpha}(\mu) = \overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ est la marge d'erreur, $1-\alpha$ (α étant le risque d'avoir tort) représente le niveau de confiance, $z_{\alpha/2}$ représente la valeur z fournissant une aire égale à $\alpha/2$ dans la queue supérieure de la distribution de probabilité normale centrée réduite.

3.1.2 Moyenne d'une population : cas où σ est inconnu

$$\mathrm{IC}_{1-\alpha}(\mu) = \overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$t_{\alpha/2}\frac{s}{\sqrt{n}}$ est la marge d'erreur, $1-\alpha$ (α étant le risque d'avoir tort) représente le niveau de confiance, $t_{\alpha/2}$ est la valeur t fournissant une aire égale à $\alpha/2$ dans la queue supérieure de la distribution de Student avec $n-1$ ddl. On rappelle que : $s^2 = \dfrac{\sum(x_i-\overline{x})^2}{n-1}$.

3.1.3 Proportion d'une population

$$\mathrm{IC}_{1-\alpha}(p) = f \pm z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$$

$z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$ est la marge d'erreur, $1-\alpha$ (α étant le risque d'avoir tort) représente le niveau de confiance et $z_{\alpha/2}$ représente la valeur z fournissant une aire égale à $\alpha/2$ dans la queue supérieure de la distribution de probabilité normale centrée réduite.

3.1.4 Taille d'échantillon de l'intervalle de confiance pour la moyenne d'une population

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2}$$

3.1.5 Taille d'échantillon de l'intervalle de confiance pour la proportion d'une population

$$n = \frac{\left(z_{\alpha/2}\right)^2 p^*(1-p^*)}{E^2}$$

La valeur de p^* peut être déterminé par plusieurs procédures : en recourant à la proportion d'échantillon, mener une étude pilote, intuition ou encore, prendre $p = 0,5$.

3.2 Formules clés - Tests d'hypothèse

TABLE 3.1 – Résumé des tests d'hypothèses relatifs à la moyenne d'une population : cas où σ est connu.

	Test unilatéral inférieur	Test unilatéral supérieur	Test bilatéral
Hypothèses	$H_0 : \mu \geqslant \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leqslant \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$
Statistique	$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$
Règle de rejet	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$
Zone de rejet	RH_0 si $z \leqslant -z_\alpha$	RH_0 si $z \geqslant z_\alpha$	RH_0 si $z \leqslant z_{\alpha/2}$ ou $z \geqslant z_{\alpha/2}$

Note : RH_0 : Rejet de H_0 et p est la p-value.

TABLE 3.2 – Résumé des tests d'hypothèses relatifs à la moyenne d'une population : cas où σ est inconnu.

	Test unilatéral inférieur	Test unilatéral supérieur	Test bilatéral
Hypothèses	$H_0 : \mu \geqslant \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leqslant \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$
Statistique	$z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$	$z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$	$z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$
Règle de rejet	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$
Zone de rejet	RH_0 si $t \leqslant -t_\alpha$	RH_0 si $t \geqslant t_\alpha$	RH_0 si $t \leqslant t_{\alpha/2}$ ou $z \geqslant t_{\alpha/2}$

Note : RH_0 : Rejet de H_0 et p est la p-value.

TABLE 3.3 – Résumé des tests d'hypothèses relatifs à la proportion d'une population

	Test unilatéral inférieur	Test unilatéral supérieur	Test bilatéral
Hypothèses	$H_0 : \mu \geqslant f_0$ $H_1 : \mu < f_0$	$H_0 : \mu \leqslant \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu = f_0$ $H_1 : \mu \neq \mu_0$
Statistique	$z = \dfrac{f - f_0}{\sqrt{\frac{f(1-f)}{n}}}$	$z = \dfrac{f - f_0}{\sqrt{\frac{f(1-f)}{n}}}$	$z = \dfrac{f - f_0}{\sqrt{\frac{f(1-f)}{n}}}$
Règle de rejet	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$	RH_0 si $p \leqslant \alpha$
Zone de rejet	RH_0 si $z \leqslant -z_\alpha$	RH_0 si $z \geqslant z_\alpha$	RH_0 si $z \leqslant z_{\alpha/2}$ ou $z \geqslant z_{\alpha/2}$

Note : RH_0 : Rejet de H_0 et p est la p-value.

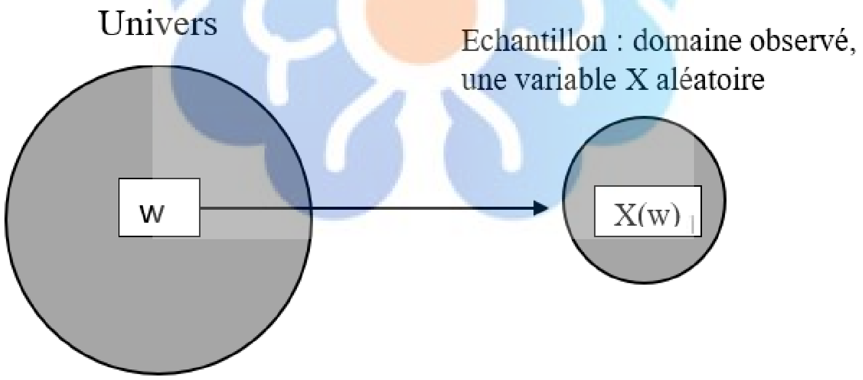
Chapitre 4

Généralités sur les tests d'hypothèse

4.1 Introduction générale

Les phénomènes sur lesquels sont faites les études statistiques sont considérés comme des évènements aléatoires. L'étude est faite sur un échantillon de réalisation c'est-à-dire des valeurs observées ou des résultats numériques.

FIGURE 4.1 – **Schéma illustratif**



Ω := Univers, X := Variable aléatoire ; P : probabilité de distribution inconnue.
P^X : distribution de probabilité inconnue a priori.
X étant aléatoire, l'information $X(\omega)$ est emprunte d'incertitude.

Comme il est difficile, voire impossible de remonter l'information jusqu'au niveau de l'univers Ω afin de connaître tous les paramètres de la distribution de probabilité P, on est emmené à prendre une décision en se basant sur un aspect partiel (l'échantillon observé) du phénomène aléatoire. Il faut donc évaluer l'incertitude pour savoir jusqu'à quel point compter sur les conclusions tirées des observations c'est-à-dire il faut évaluer le risque d'erreur de se tromper

4.2 Exemples de problèmes de tests :

Exemple 4.2.1. *La norme ISO est le système de certification des produits.*

Une usine fabrique des pièces cylindriques qui seront commercialisées si le diamètre d'une pièce est égal à d_0, ce qui correspond à la norme (valeur ISO acceptée dans le commerce). Le service de contrôle de fabrication de l'usine doit contrôler chaque lot de pièces fabriqué avant de mettre en vente. Ne pouvant contrôler l'ensemble des pièces des lots, le service va prélever un échantillon de pièces dans un lot et procéder au contrôle ; Le lot contient N pièces, n pièces. Après observation d'un échantillon, on peut obtenir le diamètre moyen $\overline{d}$, la proportion des pièces defectueuses p (ne répondant pas à la norme). La décision de commercialiser ou non les pièces fabriquées sera basée sur les seules observations des échantillons et non sur toutes les pièces fabriquées dans un lot. Par exemple, commercialiser si la proportion $p < p_0$ fixé. On peut aussi dire $\overline{d} \in \left[d_0 - \epsilon\ ,\ d_0 + \epsilon\right]$, avec ϵ suffisamment petit.
Si d'aventure après décision il s'avère que la vraie proportion p est supérieure à p_0 aurait pris un risque en décidant de mettre en vente ce lot. Le risque étant inévitable, car l'observation est faite sur des échantillons, il faut donc mesurer et minimiser ce risque découlant de la prise de décision.
On vient de poser un problème de test d'hypothèse paramétrique car l'hypothèse porte sur le paramètre p ou le diamètre moyen $\overline{d_0}$ de l'expérience aléatoire.

Exemple 4.2.2. *Le département des ressources humaines d'une grande boîte veut vérifier l'hypothèse selon laquelle la participation à une formation peut améliorer le rendement au travail. Il constitue deux groupes :*

- ***Groupe 1 :*** *Ceux qui participent à la formation (groupe expérimental) ;*
- ***Groupe 2 :*** *Ceux qui ne participent pas à la formation (groupe témoin)*

Le rendement de travail est mesuré dans chaque groupe sur un échantillon aléatoire. Le DRH compare les résultats des deux groupes en se basant sur l'hypothèse formulée au départ (avant l'expérience). Les résultats peuvent confirmer ou infirmer l'hypothèse. Il s'agit ici d'un problème de test consistant à comparer la valeur d'un paramètre dans deux échantillons issus d'une même population. C'est un test de comparaison.

Exemple 4.2.3. *Pour établir la période de garantie d'un appareil, un fabriquant observe un échantillon de 200 appareils ? Il constate que la courbe de la durée de vie ressemble à celle d'une loi exponentielle de paramètre λ. Il fait donc l'hypothèse que la durée est constante et est égale à λ. L'hypothèse ne concerne pas un paramètre de la loi suivi par la population d'appareils mais plutôt à l'adéquation d'une loi théorique connue aux observations (Loi appartenant à la famille exponentielle).*

Remarque 4.2.1. *Dans chacune de ces situations on a formulé clairement une hypothèse daont on se demande si les observations sont compatibles avec ou au contraire semblent l'infirmer. Cette hypothèse sera appelée hypothèse nulle ou hypothèse principale.*

Remarque 4.2.2. *Il faut noter qu'il ne s'agit pas a priori de confirmer à partir des observations l'hypothèse à tester, mais de juger si les observations l'infirmeent suffisamment. En d'autres termes, à l'issue d'un test si l'hypothèse à tester est acceptée c'est faute d'avoir pu suffisamment l'infirmer*

4.3 Cadre général des tests paramétriques

On considère un modèle statistique paramétrique

$$\left(\Omega_X, \left\{P_\theta;\ \theta \in \Theta\right\}\right) \quad X \in \Omega_X, \quad X = (X_1, \cdots, X_n)$$

Θ est l'ensemble des valeurs du paramètre θ

On fait l'hypothèse nulle $H_0 : \theta \in \Theta$. Cette hypothèse paramétrique

$$\Theta = \Theta_0 \cup \Theta_1, \quad avec \quad \Theta_0 = \left\{\theta \in \Theta \quad v\acute{e}rifiant \quad H_0\right\} \quad \Theta_0 \cap \Theta_1 = \emptyset$$

On a $H_0 \Longleftrightarrow \theta \in \Theta_0$ et $H_1 \Longleftrightarrow \theta \in \Theta_1$

Définition 4.3.1. *H_0 : Hypothèse nulle (ou principale) et H_1 : l'hypothèse alternative.*

Définition 4.3.2. *Si $\Theta = \{\theta\}$ est un singleton, l'hypothèse $H_0 : \theta = \theta_0$ est dite simple. Dans les autres cas, l'hypothèse est dite composite.*

Exemple 4.3.1. *X est la variable extraite d'une population atteinte d'une certaine maladie. $X = (X_1, \cdots, X_n)$ est i.i.d. On veut tester si la proportion p de personnes atteintes ne dépasse pas un seuil p_0 donné.*
Dans le présent cas, on a :

$$H_0 : p \leqslant p_0 \ \ et \ \ H_1 : p_1 > p_0 \ ; \ \ \Theta_0 = \left[0, p_0\right] \ \ et \ \ \Theta_1 = \left]p_0, 1\right]$$

Soit $x = (x_1, \cdots, x_n)$ une réalisation de l'échantillon. $\Phi : \mathbb{R}^n \longrightarrow [0, 1]$ une mesure de probabilité.

Définition 4.3.3. *Φ est la fonction du test pour tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. Si on rejette l'hypothèse H_0 avec la probabilité $\Phi(x)$ et on rejette H_1 avec la probabilité $1 - \Phi(x)$.*

Si on prend $\Phi(x) = \mathbf{1}_{\Theta_1}(\theta)$, on définit les ensembles :

$$W = \{x : \Phi(x) = 1\} \quad et \quad \overline{W} = \mathbb{R}^n - W$$

W est appelé région critique du test ou région de rejet de l'hypothèse nulle H_0. $\overline{W}$ est par contre appelé région d'acceptation de l'hypothèse nulle H_0. La fonction Φ et W caractérisent un problème un problème de test.

Erreurs de décision :

	H_0 vraie	H_0 fausse
On accepte H_0	Bonne décision $(1-\alpha)$	Erreur de deuxième espèce β
On rejette H_0	Erreur de première espèce α	Bonne décision $(1-\beta)$

$\alpha = P($rejeter H_0 sachant H_0 est vraie, ou à tort)
$\beta = P($ Ne pas rejeter H_0 sachant que H_0 est fausse, ou à tort).

Définition 4.3.4. *Le risque est défini comme la probabilité de commettre une erreur.*

Définition 4.3.5. *On appelle risque de première espèce, la probabilité de commettre une erreur de première espèce :* $\alpha(\theta) = P_\theta(W) = P_{\theta_0}(W)$ *avec* $\theta \in \Theta_0$

Définition 4.3.6. *On appelle risque de deuxième espèce, la probabilité de commettre une erreur de deuxième espèce*

$$\beta(\theta) = P_\theta(\overline{W}) = 1 - P_\theta(W) = 1 - P_{\theta_1}(W) \ \ avec \ \ \theta \in \Theta_1$$

Définition 4.3.7. *On appelle niveau (ou seuil du test), la valeur* $\alpha = sup\left\{\alpha(\theta) : \theta \in \Theta_0\right\}$.

Définition 4.3.8. *On appelle puissance du test la probabilité de rejeter H_0 sachant que H_0 est fausse. On la note π_θ, avec*

$$\pi_\theta = E_\theta(\Phi(x)) = 1 - \beta(\theta), \quad \theta \in \Theta_1$$

Le test est puissant lorsque β est faible, c'est-à-dire si le risque de deuxième espèce est faible.

Problématique : Quelque soit la décision prise au vue des observations, il subsiste un risque de se tromper :

— Rejeter l'hypothèse alors qu'elle est vérifier (rejeter H_0 à tort);
— Accepter l'hypothèse nulle alors qu'elle n'est pas vérifier (accepter H_0 à tort).

Le problème revient donc à quantifier et à minimiser les deux risques. Or minimiser l'un des risques tend à faire augmenter l'autre, ce qui conduit à un problème d'optimalité. En l'absence d'une procédure de décision optimale, il va falloir faire recours à des principes de selection afin de chercher le test optimal parmi un ensemble de tests

Choix d'un test : Le choix d'un test sera fondé sur la fonction de risque $R(W,\theta)$ définie à partir de la fonction de perte $L(\theta,d)$ où d est la décision prise.

$$R(W,\theta) = E_\theta\left(L(\theta,d)\right) = L(\theta,d_1)P_\theta(W) + L(\theta,d_0)\left[1 - P_\theta(W)\right]$$

avec $L(\theta,d_0) = 0,\ \ \forall\theta \in \Theta_0\ \ et\ \ L(\theta,d_1) = 0,\ \ \forall\theta \in \Theta_1$
En d'autres termes : si on a deux régions critiques W_1 et W_2 telles que $W_1 >> W_2$ (W_1 est meilleure que W_2 en terme de risque)

$$\alpha_\theta(W_1) \leqslant \alpha_\theta(W_2),\ \ \forall\theta \in \Theta_0\ \ et\ \ \beta_\theta(W_1) \leqslant \beta_\theta(W_2),\ \ \forall\theta \in \Theta_1$$

Le choix du test est fondé sur $\alpha(\theta)$ et $\beta(\theta)$.

4.3.1 Principe de NEYMAN

Il consiste à essayer de minimiser le risque de deuxième espèce en ayant bloqué celui de première espèce à un niveau maximal acceptable. On suppose que le risque de deuxième espèce est lié à l'erreur de première espèce considérée comme la plus grave. On voit donc que les hypothèses H_0 et H_1 ne jouent pas des rôles symétriques. Le seuil α étant fixé, on cherche sous cette contrainte s'il n'existe pas un test ayant pour risque de deuxième espèce inférieur à tous les autres de même niveau, c-à-d un test plus puissant que tous les autres.

4.3.2 Test uniformément le plus puissant

A priori un test de région critique W^* peut être plus puissant que tout autre test de région critique W pour certaines valeurs du paramètre θ et moins puissant pour d'autres valeurs de θ.

$$Sch\div ma1$$

Le test de région critique W est plus puissant que le test W', $\forall\theta \in \Theta$. Mais W' et $W\varepsilon$ ne sont pas comparables.
Le problème de NEYMAN est le suivant :

$$\begin{cases} Min\ \beta_\theta(W) & \forall\theta \in \Theta_1 \\ Sc\ \ \alpha_0(W) \leqslant \alpha_0 & \alpha_0\ donn\div,\ \ \forall\theta \in \Theta_0 \end{cases}$$

Définition 4.3.9. *Un test W^* solution du programme de NEYMAN est appelé* ***test uniformément le plus puissant de niveau*** *α, noté UPP_α, et vérifiant :*

1. *W^* est de niveau α ;*
2. *Pour tout autre test W de même niveau, on a :*

$$\beta_\theta(W^*) \leqslant \beta_\theta(W), \quad \forall\theta \in \Theta_1 \quad \text{ou encore} \quad \pi_\theta(W^*) \geqslant \pi_\theta(W) \quad \forall\theta \in \Theta_1$$

Remarque 4.3.1. *Le test UPP_α est tel que la contrainte de niveau est saturé.*

dans la suite on aura deux grandes familles de tests d'hypothèses : **Les tests d'hypothèses paramétriques et les tests d'hypothèses non paramatriques**.

Chapitre 5

Test d'une hypothèse contre une autre

5.1 Cadre du problème

- L'hypothèse est dite simple si les ensembles Θ_0 et Θ_1 sont réduits à des singletons.
- Le modèle général d'un test d'hypothèses simples est donc

$$\left(\Omega_X, \left\{P_\theta, \theta \in \Theta = \left\{\theta_0, \theta_1\right\}\right\}\right)$$

avec P_{θ_0} et P_{θ_1} parfaitement connues.

Remarque 5.1.1. *P_{θ_0} et P_{θ_1} ne sont pas toujours identiques.*

En adoptant le principe de NEYMAN, le niveau (ou seuil) α du test est borné.

$$\forall \alpha \in]0, 1[\quad \alpha(W) = Sup_{\theta_0} P_\theta(W) \leqslant \alpha$$

Sous cette contrainte, on chercher à maximiser la puissance du test, c-à-d

$$\begin{cases} Max P_{\theta_1}(W) \\ W \ / \ P_{\theta_0}(W) \leqslant \alpha \end{cases}$$

5.2 Théorème de NEYMAN - PEARSON

Jersey NEYMAN (Moldavie 1894 - Californie 1981) et Karl PEARSON (Londres 1857 - 1936) Biometrica.

Théorème 5.2.1. $\forall \alpha \in [0, 1]$ *pour le problème de test*

$$\begin{cases} H_0 : P_\theta = P_{\theta_0} \\ H_1 : P_\theta = P_{\theta_1} \end{cases}$$

Il existe $k_\alpha \in \mathbb{R}^+$ et $\gamma_\alpha \in [0,\ 1]$ tels que le test le plus puissant de seul α pour tester H_0 contre H_1 est donné par :

$$\varphi(x) = \begin{cases} 1 \ si \ L(P_{\theta_1},\ \tilde{x}) > k_\alpha L(P_{\theta_0}, \tilde{x}) \\ \gamma_\alpha \ si \ L(P_{\theta_1},\ \tilde{x}) = k_\alpha L(P_{\theta_0}, \tilde{x}) \\ 0 \ \ L(P_{\theta_1},\ \tilde{x}) < k_\alpha L(P_{\theta_0}, \tilde{x}) \end{cases}$$

$\tilde{x} = (x_1, \cdots, x_n)$

La région critique d'un tel test est donné

$$W_\alpha = \left\{ \tilde{x} \ / \ \frac{L(P_{\theta_1},\ \tilde{x})}{L(P_{\theta_0,\ \tilde{x})}} \geqslant k_\alpha \right\} \quad avec \ k_\alpha \ tel \ que \ \ P_{\theta_0}(W_\alpha) = \alpha$$

Intuitivement, si ce rapport est grand, cela signifie que $L(P_{\theta_1}, \tilde{x})$ est plus vraisemblable que $L(P_{\theta_0}, \tilde{x})$.

Démonstration. Soit φ' une autre fonction du test de niveau $\alpha' < \alpha$. Il faut montrer que φ' est moins puissant que φ.
Soit $A(x) = \varphi(x) - \varphi'(x)$, $B(x) = L(P_{\theta_1}, \tilde{x}) - k_\alpha L(P_{\theta_0}, \tilde{x})$ *et* $g(x) = A(x)B(x)$
Si $B(x) > 0$ alors $\varphi(x) = 1$, donc $A(x)1 - \varphi'(x) \geqslant 0$ et donc $g(x) \geqslant 0$
Par conséquent,$\forall x \in \Omega_X$, $g(x) \geqslant 0$ cela implique que $\int_{\Omega_X} g(x) d P_\theta \geqslant 0$ or

$$\int_{\Omega_X} g(x) d P_\theta = \int_{\Omega_X} \varphi(x) L(P_{\theta_1}, \tilde{x}) d P_\theta - \int_{\Omega_X} \varphi'(x) L(P_{\theta_1}, \tilde{x}) d P_\theta - \int_{\Omega_X} \varphi(x) L(P_{\theta_1}, \tilde{x}) k_\alpha d P_\theta + \int_{\Omega_X} \varphi'(x) L(P_{\theta_1}, \tilde{x}) k_\alpha d P_\theta$$

$$\int_{\Omega_X} g(x) d P_\theta = \Pi_{\theta_1}(\varphi) - \Pi_{\theta_1}(\varphi') + k_\alpha \left(\Pi_{\theta_0}(\varphi') - \Pi_{\theta_0}(\varphi) \right) \geqslant 0$$

Ainsi, $1 - \beta(\varphi) - 1 + \beta(\varphi') + k_\alpha(\alpha' - \alpha) \geqslant 0$. Autrement,

$$\beta(\varphi') - \beta(\varphi) \geqslant k_\alpha(\alpha - \alpha') > 0 \implies \beta(\varphi') > \beta(\varphi)$$

Donc φ est plus puissant que φ'. □

Un test ayant une région critique de cette forme est appelé **test de NEYMAN**

5.3 Exemples

Exemple 5.3.1. ***Test de la valeur de la moyenne dans le modèle gaussien, la variance étant connue.***
Les hypothèses étant simple, le modèle est

$$\left(X,\ \left\{\mathcal{N}(m,\ \sigma^2);\ \ m \in \left\{m_0,\ m_1\right\},\ \ m_0 < m_1\right\}\right)$$

Échantillon $(X_1, \cdots, X_n)$ i.i.d de X. on veut tester les hypothèses suivantes :

$$\begin{cases} H_0 \ : \ m = m_0 \\ Contre \\ H_1 \ : \ m = m_1 \ \ au \ seuil \ \alpha \ donn\dot{-} \end{cases}$$

Le théorème de Neyman - Pearson implique qu'il existe un test UPP_α de niveau α dont la région critique est de la forme :

$$W_\alpha = \left\{ (X_1, \cdots, X_n) \ / \ \frac{L(x_1, \cdots, x_n;\ m_1)}{L(x_1, \cdots, x_n;\ m_0)} \geqslant k_\alpha \right\} \quad avec\ k_\alpha\ tel\ que\ \ P_{m_0}(W_\alpha) = \alpha$$

Dans ce cas L désigne la vraisemblance. Autrement dit,

$$L(x_1, \cdots, x_n;\ m) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \times exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \right)$$

Après remplacement, on obtient l'expression suivante :

$$W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ 2(m_1 - m_0) \sum_{i=1}^{n} x_i + n(m_0^2 - m_1^2) \geqslant k'_\alpha \right\} = \left\{ (x_1, \cdots, x_n) \ / \ 2(m_1 - m_0)\overline{x}_n + (m_0^2 - m_1^2) \geq k''_\alpha \right\}$$

Comme $m_0 < m_1$, alors $(m_1 - m_0)\overline{x}_n$ est une fonction croissante de $\overline{x}_n$,

$$W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ \overline{x}_n \geq \frac{k''_\alpha + (m_1^2 - m_0^2)}{2(m_1 - m_0)} \right\} = \left\{ (x_1, \cdots, x_n) \ / \ \overline{x}_n \geq k'''_\alpha \right\}$$

Détermination de k'''_α :

On a :

$$P_{m_0}(W_\alpha) = \alpha = P_{m_0} \left\{ (x_1, \cdots, x_n) \ / \ \overline{X}_n \geq k'''_\alpha \right\} = \left\{ (x_1, \cdots, x_n) \ / \ \sqrt{n}\frac{\overline{X}_n - m_0}{\sigma} \geq \sqrt{n}\frac{k'''_\alpha - m_0}{\sigma} \right\}$$

Donc $\alpha = P_{m_0} \left\{ (x_1, \cdots, x_n) \ / \ U \geq \sqrt{n}\frac{k'''_\alpha - m_0}{\sigma} \right\}$ ce qui signifie que $\sqrt{n}\frac{k'''_\alpha - m_0}{\sigma} = u_{1-\alpha}$ on en déduit que

$$k'''_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \ et\ W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ \overline{x}_n \geq m_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \right\}$$

Remarque 5.3.1. *1. Intuitivement, on rejette l'hypothèse que la moyenne théorique m soit la plus petite des deux moyennes lorsque la moyenne empirique est trop élevée.*

2. La borne inférieure de la région critique W_α se rapproche définiment de m_0 lorsque n tend vers $+\infty$. Dans ces conditions on rejette H_0 dès que $\overline{x}_n$ dépasse de très peu m_0. En effet, $\overline{X}_n$ étant un estimateur convergent de m, m_0 constitue un point d'accumulation de la fonction densité de $\overline{X}_n$ et une valeur légèrement différente de m_0 devient atypique.

3. La région critique W_α ne dépend de m_1 (la valeur de l'hypothèse alternative). Ainsi, W_α est la région critique de tous les problèmes de tests de la forme

$$\begin{cases} H_0 \ : m = m_0 \\ Contre \\ H_1 \ : m = m_1, \ \ m_1 \geqslant m_0 \end{cases}$$

Par conséquent W_α est la région critique des problèmes de tests de la forme

$$\begin{cases} H_0 \ : m = m_0 \\ Contre \qquad W_\alpha\ est\ donc\ UPP_\alpha \\ H_1 \ : m = m_1, \ \ m_1 > m_0 \end{cases}$$

Puissance du test : α étant fixé, il reste à évaluer β. $\beta(m_1) = 1 - \beta(m_1)$ avec

$$\Pi_{(m_1)} P_{m_1}(W_\alpha) = P_{m_1}\left(\left\{\overline{X}_n \geqslant m_0 + \frac{\sigma}{\sqrt{n}}\mu_{1-\alpha}\right\}\right)$$

On obtient alors

$$\Pi_{(m_1)} = P_{m_1}\left(\left\{\sqrt{n}\frac{\overline{X}_n - m_1}{\sigma} \geqslant \sqrt{n}\frac{m_0 - m_1)}{\sigma} + \mu_{1-\alpha}\right\}\right) = P_{m_1}\left\{\mathcal{N}(0,\ 1) \geqslant \sqrt{n}\frac{(m_0 - m_1)}{\sigma} + \mu_{1-\alpha}\right\} = 1 - F_{\mathcal{N}(0,\ 1)}\left(\sqrt{n}\frac{m_0 - m_1}{\sigma} + \mu_{1-\alpha}\right)$$

On obtient enfin

$$\beta(m_1) = F_{\mathcal{N}(0,\ 1)}\left(\sqrt{n}\frac{m_0 - m_1}{\sigma} + \mu_{1-\alpha}\right)$$

Remarque 5.3.2. *La puissance du test, et donc le risque de deuxième espèce $\beta(m_1)$ dépend de m_1. Plus m_1 est élevé plus β est faible, et donc plus le test est puissant. Lorsque n tend vers $+\infty$, β tend vers 0 (Pas de risque de deuxième espèce).*

Exemple 5.3.2. ***Dans le modèle, on veut tester la valeur de la variance (la moyenne*** m ***supposée connue).*** *Le modèle est le suivant :* $\left(X,\ \left\{\mathcal{N}(m,\ \sigma^2);\ \sigma_0^2 < \sigma_1^2\right\}\right)$ *On teste :*

$$\begin{cases} H_0\ :\ \sigma^2 = \sigma_0^2 \\ Contre \qquad au\ seuil\ \alpha \\ H_1\ :\ \sigma^2 = \sigma_1^2 \end{cases}$$

Un raisonnement analogue au précédent permet d'obtenir l'égalité suivante :

$$W_\alpha = \left\{(x_1, \cdots, x_n)\ /\ \sum_{i=1}^{n}(x_i - m)^2 \geqslant k_\alpha\right\} \quad avec\ k_\alpha\ tel\ que\ \ P_{\sigma_0^2}(W_\alpha) = \alpha$$

Sous $H_0\ :\ \sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma_0}\right)^2 \rightsquigarrow \mathcal{X}^2_{(n)}$

$$\alpha = P_{\sigma_0^2}\left(\left\{(X_1, \cdots, X_n)\ /\ \sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma_0}\right)^2 \geqslant \frac{k_\alpha}{\sigma_0^2}\right\}\right) \qquad \frac{k_\alpha}{\sigma_0^2} = C_{1-\alpha}(n)$$

où $C_{1-\alpha}(n)$ est le fractile d'ordre $1 - \alpha$ de la loi du $\mathcal{X}^2_{(n)}$ Donc

$$W_\alpha = \left\{(x_1, \cdots, x_n)\ /\ \sum_{i=1}^{n}(x_i - m)^2 \geqslant \sigma_0^2 C_{1-\alpha}(n)\right\}$$

Toujours par analogie, on obtient le risque de deuxième espèce donnée par la relation

$$\beta(\sigma_1^2) = F_{\mathcal{X}^2_{(n)}}\left(\frac{\sigma_0^2}{\sigma_1^2} C_{1-\alpha}(n)\right)$$

5.4 Probabilité critique

Au lieu de décider en fonction du niveau du test choisi, les logiciels statistiques fournissent des probabilités critiquespour le rejet de H_0.
Si $\varphi(X)$ est la statistique du test, t la valeur observée de $\varphi(X)$, on aura

Forme de la région critique W_α	Probabilité critique
$\{\varphi(X) \geqslant k\}$	$P_{H_0}(\varphi(X) \geqslant t)$
$\{\varphi(X) \leqslant k\}$	$P_{H_0}(\varphi(X) \leqslant t)$

α_0 étant le seuil choisi, si la probabilité critique α est tel que $\alpha < \alpha_0$, la valeur t observée sur l'échantillon se trouve dans W_α. Donc on doit rejeter H_0. α **est appelé " p - value "**.

5.5 Lien entre test UPP de Neyman et statistique exhaustive

Théorème 5.5.1. .

1. *Si S est une statistique exhaustive de l'échantillon, la rrégion critique du test UPP_α ne dépend que de S.*
2. *Cette région critique peut être entièrement déterminée dans le modèle de S.*

Démonstration. D'après le critère de factorisation, on a :

$$f(x,\ \theta) = g(x) \times h(S(x),\ \theta) \quad g \geq 0 \ \ et \ \ h > 0$$

Le rapport des vraisemblances donne :

$$\frac{L(\tilde{x},\ \theta_1)}{L(\tilde{x},\ \theta_0)} = \frac{h\left(S(x),\ \theta_1\right)}{h\left(S(x),\ \theta_0\right)}$$

Donc W_α est fonction des observations uniquement à travers la statistique S. □

Chapitre 6

Tests d'hypothèses composites

6.1 Tests d'hypothèses unilatérales

Les hypothèses sont composites lorsque Θ_0 et Θ_1 ne sont pas réduits à des singletons mais sont plutôt des sous ensembles de $\mathbb{R}$, intervalles fermés ou non.
Le modèle étant :

$$\left(X,\ \left\{P_\theta,\ \theta \in \Theta = \Theta_0 \cup \Theta_1\right\}\right)$$

Les hypothèses à tester sont les suivantes :

$$(1)\begin{cases} H_0\ :\ \theta \leq \theta_0 \\ Contre \\ H_1\ :\ \theta > \theta_0 \end{cases} \quad ou \quad (2)\begin{cases} H_0\ :\ \theta \geq \theta_0 \\ Contre \\ H_1\ :\ \theta < \theta_0 \end{cases}$$

Remarque 6.1.1.

1. *Pour résoudre le problème de test* (2)*, il suffit de prendre* $\lambda = -\theta$ *et se ramener au problème* (1).
2. *L'hypothèse alternative dans ce cadre n'est pas réduite à une seule loi.*
3. *Rien n'indique a priori l'existence d'un test* UPP_α. *Il faut donc trouver des conditions pour rechercher des tests* UPP.

6.1.1 Famille de lois à rapport de vraisemblances monotones (RVM)

Définition 6.1.1. *Une famille de lois* $\left(P_\theta;\ \theta \in \Theta\right)$ *de densité* $f(x,\ \theta)$ *est le rapport de vraisemblances monotones ssi :*

$$\forall \theta_1 < \theta_2,\ \frac{L(\tilde{x},\ \theta_2)}{L(\tilde{x},\ \theta_1)}$$

est une fonction strictement croissante d'une statistique S *de l'échantillon.*

Si ce rapport est strictement décroissant, on prendra $T = -S$.
Si P_θ appartient à l famille exponentielle, la densité s'écrirant

$$f(x,\ \theta) = C(\theta) \times exp\left[\alpha(\theta)T(x) + \beta(\theta)\right]$$

alors β est à RMV ssi $\alpha(\theta)$ conserve un signe constant $T(x)$.

Démonstration.

$$\frac{f(x,\theta_2)}{f(x,\theta_1)} = exp\left[\left(\alpha(\theta_2)-\alpha(\theta_1)\right)\times T(x) + \left(\beta(\theta_2)-\beta(\theta_1)\right)\right]$$

est monotone de $T(x)$ ssi $\alpha(\theta_2)-\alpha(\theta_1)$ conserve un signe constant. □

Exemple 6.1.1. *Dans le modèle* $\left(X,\ \left\{\mathcal{N}(m,\ \sigma^2),\ \sigma^2\ connu\right\}\right)$ *Pour* $m > m_1$ *on a :*

$$\frac{L(x_1,\cdots,x_n;\ m_2)}{L(x_1,\cdots,x_n;\ m_1)} = exp\left[\frac{1}{\sigma^2}(m_2-m_1)\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\left(m_2^2+m_1^2\right)\right]$$

qui est une fonction croissante la statistique $T=\sum_{i=1}^{n} X_i$ *ou* $S=\overline{X}_n$.

6.1.2 Théorème de Lehman

Dans le cas où la loi de P_θ du modèle est une famille à RMV croissante pour la statistique T, il existe, pour le problème de test

$$\begin{cases} H_0\ :\ \theta \leq \theta_0 \\ Contre \\ H_1\ :\ \theta > \theta_0,\quad au\ seuil\ \ \alpha \end{cases}$$

un test UPP_α dont la région critique est donnée par :

$$W_{UPP_\alpha} = \{T(X)\geq k\}\quad avec\ P_{\theta_0}(W_{UPP_\alpha}) = \alpha$$

Remarque 6.1.2. *La contrainte de niveau* $P_{\theta_0}(W_{UPP_\alpha}) = \alpha$ *est satisfaite aux frontières de l'hypothèse* $H_0\ :\ \theta=\theta_0$.
Si la loi est discrète, on prendra le niveau $\alpha_0 \leq \alpha,\ \ \alpha_0$ *le plus proche de* α.

Pour le problème :

$$\begin{cases} H_0\ :\ \theta \geq \theta_0 \\ Contre \\ H_1\ :\ \theta < \theta_0 \end{cases}$$

on reparamètre en prenant $\lambda=-\theta$. Le rapport de vraisemblances est monotone décroissante en T et croissante en $S=-T$. On a :

$$W_{UPP_\alpha} = \{S(X)\geq k\} = \left\{T(X)\leq k'\right\}$$

Par conséquent changer le sens des inégalités dans les hypothèses du test implique de changer le sens de l'inégalité dans la région critique.

Remarque 6.1.3.

1. *Comme dans le cas du test de Neyman, la région critique W_{UPP_α} étant définie en termes de vraisemblances, elle ne dépend que des statistiques exhaustives.*
2. *L'inégalité de l'hypothèse nulle est large. Elle garantie que le niveau α est effectivement atteint sur l'ensemble des valeurs de Θ_0 de l'hypothèse nulle.*

Exemple 6.1.2. $\left(X, \left\{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}\right\}\right)$

1. $\begin{cases} H_0 : m \leq m_0 \\ Contre \\ H_1 : m > m_0 \end{cases}$ *le modèle est à RMV croissante de $T(X) = \overline{X}_n$, le théorème de Lehman assure que :* $W_{UPP_\alpha} = \{T(X) \geq k\} = \left\{\overline{X}_n \geq k\right\}$ *avec*

$$P_{m_0}\left(W_{UPP_\alpha}\right) = \alpha \quad W_{UPP_\alpha} = \left\{\overline{X}_n \geq m_0 + \frac{\sigma}{\sqrt{n}}\mu_{1-\alpha}\right\}$$

<u>***Fonction puissance :***</u> .

$\forall m > m_0$,

$$\Pi_m = P_m\left(W_{UPP_\alpha}\right) = P_m\left(\left\{\overline{X}_n \geq m_0 + \frac{\sigma}{\sqrt{n}}\mu_{1-\alpha}\right\}\right) = 1 - F_{\mathcal{N}(0,\,1)}\left(\sqrt{n}\frac{m_0 - m}{\sigma} + \mu_{1-\alpha}\right)$$

Dans le même modèle, pour le problème

$$\begin{cases} H_0 : m \geq m_0 \\ Contre \\ H_1 : m < m_0 \end{cases}$$

on a $W_{UPP_\alpha} = \left\{\overline{X}_n \leq m_0 - \frac{\sigma}{\sqrt{n}}\mu_{1-\alpha}\right\}$

6.2 Tests d'hypothèses bilatérales

On pose les problèmes de test dans lesquels l'une des hypothèses n'est pas un intervalle mais une réunion d'intervalles disjoints de $\mathbb{R}$. $\left(\Theta_0 \text{ ou } \Theta_1 \text{ sont non convexes}\right)$

6.2.1 Test de l'extérieur d'un intervalle

On prend $\theta_1 < \theta_2$, on teste les hypothèses suivantes :

$$\begin{cases} H_0 : \theta \leq \theta_1 \text{ ou } \theta \geq \theta_1 \\ Contre \\ H_1 : \theta_1 < \theta < \theta_2 \end{cases}$$

Théorème 6.2.1. *Si le modèle $\left(X,\ \left\{P_\theta,\ \theta \in \Theta\right\}\right)$ appartirnt à la famille exponentielle, il existe un test UPP_α de région critique $W_\alpha = \left\{k_1 \leq \Pi(X) \leq k_2\right\}$ avec k_1 et k_2 tels que $P_{\theta_1}(W_\alpha) = P_{\theta_2}(W_\alpha) = \alpha$ (dans le cas d'une loi discrète, on prendre $\alpha_0 \leq \alpha$, avec α_0 le plus petit proche de α.*

Remarque 6.2.1. *Le niveau α du test est atteint sur la frontière des valeurs des hypothèses.*

Exemple 6.2.1. $X \rightsquigarrow \mathcal{N}(m,\ \sigma^2)$

$$\begin{cases} H_0\ :\ m \leq m_1\ ou\ \ m \geq m_2 \\ Contre \\ H_1\ :\ m_1 < m < m_2 \end{cases}$$

On suppose que σ^2 connu. On a vu que dans ce modèle $\dfrac{m}{\sigma^2}\sum X_i$ est une fonction strictement croissante de m. D'après le théorème il existe un test UPP_α de région critique $W_\alpha = \left\{k_1 \leq \sum_{i=1}^{n} X_i \leq k_2\right\}$ *avec $P_{m_1}(W_\alpha) = P_{m_2}(W_\alpha) = \alpha$ après transformation, les deux conditions de niveau donnent*

$$P_{m_1}\left(\sqrt{n}\frac{k_1' - m_1}{\sigma} \leq \sqrt{n}\frac{\overline{X}_n - m_1}{\sigma} \leq \sqrt{n}\frac{k_2' - m_1}{\sigma}\right) = \alpha \quad P_{m_2}\left(\sqrt{n}\frac{k_1' - m_2}{\sigma} \leq \sqrt{n}\frac{\overline{X}_n - m_2}{\sigma} \leq \sqrt{n}\frac{k_2' - m_2}{\sigma}\right) = \alpha$$

on a deux intervalles de même longueur qui doivent avoir la même probabilité α.
Soit ϵ tel que $u_\epsilon = \sqrt{n}\dfrac{k_1' - m_1}{\sigma}$ on a

$$P_{m_1}\left(u_\epsilon \leq \mathcal{N}(0,\ 1) \leq \sqrt{n}\frac{k_2' - m_1}{\sigma}\right) = \alpha \Longleftrightarrow F_{\mathcal{N}(0,1)}\left(\sqrt{n}\frac{k_2' - m_1}{\sigma}\right) - F(u_\epsilon) = \alpha$$

On a ainsi, $\sqrt{n}\dfrac{k_2' - m_1}{\sigma} = u_{\epsilon+\alpha}$ *donc*

$$k_1' = m_1 + \frac{\sigma}{\sqrt{n}}u_\epsilon \quad et \quad k_2' = m_1 + \frac{\sigma}{\sqrt{n}}u_{\epsilon+\alpha}$$

En injectant dans (2) on trouve

$$P_{m_2}\left(\sqrt{n}\frac{m_1 - m_2 + \frac{\sigma}{\sqrt{n}}u_\epsilon}{\sigma} \leq \mathcal{N}(0,\ 1) \leq \sqrt{n}\frac{m_1 - m_2 + \frac{\sigma}{\sqrt{n}}u_{\epsilon+\alpha}}{\sigma}\right) = \alpha$$

Les intervalles $\left[u_\epsilon,\ u_{\epsilon+\alpha}\right]$ et $\left[\sqrt{n}\dfrac{m_1 - m_2 + \frac{\sigma}{\sqrt{n}}u_\epsilon}{\sigma},\ \sqrt{n}\dfrac{m_1 - m_2 + \frac{\sigma}{\sqrt{n}}u_{\epsilon+\alpha}}{\sigma}\right]$ *Soit de même longueur et de même probabilité $\mathcal{N}(0,\ 1)$ étant syumétrique, $\sqrt{n}\dfrac{m_2 - m_1}{\sigma} + u_\epsilon = -u_{\epsilon+\alpha}$ donc $u_\epsilon + u_{\epsilon+\alpha} = \sqrt{n}\dfrac{m_2 - m_1}{\sigma}$. Ce qui détermine ϵ de manière unique. On en déduit k_1' et k_2' et la région critique W_α*

6.2.2 Test de l'intérieur d'un intervalle

On considère toujours $\theta_1 < \theta_2$. On teste les hypothèses suivantes :

$$\begin{cases} H_0 \,:\, \theta_1 \leq \theta\theta_2 \\ Contre \\ H_1 \,:\, \theta < \theta_1 \;\; ou \;\; \theta > \theta_2 \end{cases}$$

Définition 6.2.1. *Un test de région critique W est dit sans biais si*

$$\left(\forall\theta \in \Theta_0, \;\; P_\theta(W) \leq \alpha\right) \quad et \quad \left(\forall\theta \in \Theta_1 \;\; P_\theta(W) \geq \alpha\right)$$

Ces deux inégalités signifient que

$$\forall\theta \in \Theta_0, \; \forall\theta' \in \Theta_1 \;\; 1 - P_\theta(W) \geq 1 - P_{\theta'}(W)$$

Donc si le test est sans biais, la probabilité d'une bonne décision est toujours la plus grande (quelle que soit). Donc la puissance du test est toujours supérieure au niveau du test.

Théorème 6.2.2. *Si la famille de lois du modèle est de type exponentiel avec $\alpha(\theta)$ strictement croissante, alors pour le problème de test posé ; il existe un test UPP_α de région critique $W_\alpha = \left\{T(X) \leq k_1 \; ou \; T(X) \geq k_2\right\}$ avec $P_{\theta_1}(W_\alpha) = P_{\theta_2}(W_\alpha) = \alpha$*

Remarque 6.2.2. *La double condition est souvent difficile à obtenir dans ce cas. On traite alors le cas particulier suivant.*

6.2.3 Cas particulier

$$\begin{cases} H_0 \,:\, \theta = \theta_0 \\ Contre \\ H_1 \,:\, \theta \neq \theta_0 \end{cases}$$

On a considéré $\theta_1 = \theta_2 = \theta_0$. La région critique
$W_\alpha = \left\{T(X) \leq k_1 \; ou \;\; T(X) \geq k_2\right\}$ avec $P_{\theta_0}(W_\alpha) = \alpha$

$$Sch\dot{-}ma \qquad Svch\dot{-}ma$$

la fonction $\theta \longmapsto P_\theta(W)$ atteint son minimum en $\theta = \theta_0$, la fonction densité étant dérivable (car exponentielle) $P_\theta(W_\alpha) = \int_{W_\alpha} f(x, \; \theta)dx$ on a

$$\frac{\partial P_\theta(W_\alpha)}{\partial\theta} = 0 \; en \; \theta = \theta_0, \;\; f(x, \; \theta) = C(x) exp\left[\alpha(\theta)T(x) + \beta(\theta)\right]$$

On a

$$\frac{\partial P_\theta(W_\alpha)}{\partial\theta} = \int_{W_\alpha} \frac{\partial f(x, \; \theta)}{\partial\theta} dx = \int_{W_\alpha} \frac{\partial log f(x, \; \theta)}{\partial\theta} dx \int_{W_\alpha} \left(\alpha'(\theta)T(x) + \beta'(\theta)\right) f(x, \; \theta)dx = 0$$

Il vient que $\alpha'(\theta)T(x) + \beta'(\theta) = 0$ en $\theta = \theta_0$ on a
$\alpha'(\theta)\int_{W_\alpha} T(x)f(x,\theta_0)dx = -\beta'(\theta_0)\int_{W_\alpha} f(x,\theta_0)dx$ or $\int_{W_\alpha} f(x,\theta_0)dx = P_{\theta_0}(W_\alpha) = \alpha$
$\alpha'(\theta_0)E(T) = -\beta'(\theta_0)\alpha$ Or d'après le théorème de Koopman, si le modèle appartient à la famille exponentielle

$$E(T) = -\frac{\beta'(\theta)}{\alpha'(\theta)} \quad , \quad \int_{W_\alpha} T(x)f(x,\theta)dx = \alpha E_{\alpha_0}(T) \quad , \quad E_{\theta_0}(T\mathbf{1}_{W_\alpha}) = \alpha E_{\theta_0}(T) = E_\theta(T)E_{\theta_0}(\mathbf{1}_{W_\alpha})$$

Dans la pratique, si la loi P_θ est symétrique par rapport à une valeur α donnée $W_\alpha = \left\{T(X) \leq k_1 \; ou \; T(X) \geq k_2\right\}$ doit être symétrique par rapport à α et vérifiant $P_{\theta_0}(W_\alpha) = \alpha$. La deuxième condition est automatiquement vérifiée.

Exemple 6.2.2. *Pour le problème*

$$\begin{cases} H_0 \; : \; m = m_0 \\ Contre \\ H_1 \; : \; m \neq m_0 \end{cases}$$

$$W_\alpha = \left\{ \overline{X}_n \leq m_0 - \frac{\sigma}{\sqrt{n}}u_{1-\alpha} \;\; ou \;\; \overline{X}_n \geq m_0 - \frac{\sigma}{\sqrt{n}}u_{1-\alpha} \right\}$$

6.3 Lien entre région critique du test et région de confiance du paramètre

$$\begin{cases} H_0 \; : \; \theta = \theta_0 \\ Contre \\ H_1 \; : \; \theta = \theta_1 \end{cases}$$

Si on note $W_\alpha(\theta_0)$ la région d'acceptation de H_0

$$IC_{1-\alpha}(\theta) = \left\{\theta \;/\; \tilde{X} \in \overline{W}_\alpha(\theta_0)\right\} \quad P_\theta\left(\theta \in IC_{1-\alpha}(\theta)\right) = P_\theta\left(\tilde{X} \in \overline{W}_\alpha(\theta_0)\right) = 1 - \alpha$$

Résumé sur la démarche générale d'un test :

1. Définir l'hypothèse nulle
2. Choisir la statistique du test et définir sa distribution de probabilité sous H_0
3. Choisir le niveau de significativité du test α et la région critique W_α
4. Calculer à partir des données de l'échantillon la valeur de la statistique du test
5. Prendre une décision concernant l'hypothèse H_0

Chapitre 7

Tests en présence de paramètres nuisibles et tests asymptotiques généraux

7.1 Paramètre scalaire en présence d'un paramètre nuisible

Le modèle est $\left(X,\ \left\{P_\theta, \theta \in \Theta\right\}\right)$; $\theta = \begin{pmatrix} \lambda \\ \delta \end{pmatrix}$. On veut tester les hypothèses sur λ. La région critique du test va dépendre de δ, δ étant inconnu. L'idée est de remplacer δ par un estimateur convergent. Ceci ne peut se faire que si on peut appliquer rigoureusement les théorèmes de convergence, en particulier le théorème central limite. A défaut, on va choisir une "assez bonne" statistique (ne faisant pas perdre beaucoup d'information). On se place dans le modèle de cette statistique dont la loi ne dépend pas du paramètre nuisible δ mais seulement du paramètre testé λ et on procède au test.

Exemple 7.1.1. *(Test de la moyenne dans le modèle gaussien, variance connue)*

$$\begin{cases} H_0 \ : \ m \leq m_0 \\ Contre \\ H_1 \ : \ m > m_0 \end{cases}$$

On a vu que si σ est connu, la région critique du test est $W_\alpha = \left\{\overline{X}_n \geq m_0 + \frac{\sigma}{\sqrt{n}}\mu_{1-\alpha}\right\}$

où $\mu_{1-\alpha}$ est le fractile d'ordre α de $\mathcal{N}(0,1)$. Si σ^2 n'est pas connu, on le remplace par son esimateur convergent $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. On sait que $\sqrt{n}\frac{\overline{X}_n - m}{\sigma}$ converge en loi vers $\mathcal{N}(0,1)$ et $\frac{\sigma}{S_n}$ converge presque sûrement vers 1. Par ailleurs, le théorème de Fisher indique que $U = \sqrt{n}\frac{\overline{X}_n - m}{\sigma}$ et $Z = \frac{(n-1)S_n^2}{\sigma^2}$ sont deux variables indépendantes suivant respectivement $\mathcal{N}(0,1)$ et $\mathcal{X}^2_{(n-1)}$. Par conséquent

$$\sqrt{n}\frac{\overline{X}_n - m}{S_n} \rightsquigarrow T(n-1)$$

La vraie loi de $\sqrt{n}\frac{\overline{X}_n - m}{S_n}$ étant connue, $W_\alpha = \left\{\sqrt{n}\frac{\overline{X}_n - m_0}{S_n} \geq t_{1-\alpha}(n-1)\right\}$

Exemple 7.1.2. *(Test de la variance dans le modèle gaussien $\mathcal{N}(m, \sigma^2)$, m étant inconnue.)*
$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ Contre \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$ *. Pour m connue, on avait* $W_\alpha = \left\{ \sum_i \frac{(X_i - m)^2}{\sigma^2} \geq \mathcal{X}^2_{1-\alpha}(n) \right\}$
Aucun théorème de convergence ne permet de remplacer m par $\overline{X}_n$. On va passer par S_n^2 qui est un estimateur sans biais optimal de σ^2. La loi S_n^2 est à RVM pour S_n^2 elle-même. Il existe un test UPP_α solution du problème. La région critique est donnée par $W_\alpha = \left\{ (n-1)\frac{S_n^2}{\sigma^2} \geq \mathcal{X}^2_{1-\alpha}(n-1) \right\}$

Remarque 7.1.1. *S_n^2 n'est pas une statistique exhaustive, donc en passant de l'échantillon au modèle de S_n^2 on perd de l'information.*

7.2 Généralisation : test sur une fonction scalaire des paramètres

Ce test généralise le problème précédent. Dans la pratique, on va utiliser un bon estimateur (l'EMV par exemple) de la fonction à tester et on se place dans le modèle de cet estimateur.

Exemple 7.2.1. *Une norme sanitaire impose qu'un produit alimentaire ne doit pas contenir un taux d'une substance supérieure à un seuil s pour une quantité N donnée (le seuil s ne doit pas être dépassé dans 1 cas sur N). On prélève un échantillon de taille suffisament grande n de sorte que la variable représentant le taux de substance suive une loi normale. Le test consiste à vérifier la conformité à la norme. On aura les hypothèses :*

$$\begin{cases} H_0 : P(X > s) \geq \frac{1}{N} \\ Contre \\ H_1 : P(X > s) < \frac{1}{N} \end{cases} \quad ou \quad \begin{cases} H_0 : P\left(\frac{X-m}{\sigma} > \frac{s-m}{\sigma}\right) \geq \frac{1}{N} \\ Contre \\ H_1 : P\left(\frac{X-m}{\sigma} > \frac{s-m}{\sigma}\right) < \frac{1}{N} \end{cases}$$

Ce qui équivaut à $\begin{cases} H_0 : s - m - \sigma\mu_{1-\alpha} \leq 0 \\ Contre \\ H_1 : s - m - \sigma\mu_{1-\alpha} > 0 \end{cases}$ *On connait de bons estimateurs de m et σ que sont $\overline{X}_n$ et S_n. On sait que* $\sqrt{n}\left[\begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} m \\ \sigma^2 \end{pmatrix}\right]$ *converge en loi vers* $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$
Posons $g(y,z) = y + \mu_{1-\frac{1}{N}}\sqrt{z} - s$*, on a* $\nabla g(y,z) = (1, \ \frac{1}{2\sqrt{z}}\mu_{1-\frac{1}{N}})$ *ce qui signifie que* $\sqrt{n}\left[(\overline{X}_n + \mu_{1-\frac{1}{N}}S_n - s) - (m + \sigma\mu_{1-\frac{1}{N}} - s)\right]$ *converge en loi vers*

$$\mathcal{N}\left(0\,,\ \left(1 \quad \frac{1}{2\sigma}\mu_{1-\frac{1}{N}}\right)\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\begin{pmatrix} 1 \\ \frac{1}{2\sigma}\mu_{1-\frac{1}{N}} \end{pmatrix}\right) = \mathcal{N}\left(0\ ,\ \sigma^2\left(1 + \frac{1}{2}\mu^2_{1-\frac{1}{N}}\right)\right)$$

En posant $\lambda = m + \sigma\mu_{1-\frac{1}{N}} - s$, $\hat{\lambda}_n = \overline{X}_n + \mu_{1-\frac{1}{N}} S_n - s$, *il vient que* $\sqrt{n}(\hat{\lambda}_n - \lambda)$ *converge en loi vers* $\mathcal{N}\left(0\ ,\ \sigma^2\left(1+\frac{1}{2}\mu^2_{1-\frac{1}{N}}\right)\right)$. *on peut donc tester les hypothèses dans le modèle de* $\hat{\lambda}_n$.

$$\begin{cases} H_0 : & \lambda \geq 0 \\ Contre & \\ H_1 : & \lambda < 0 \end{cases} \qquad W_\alpha = \left\{ \sqrt{n}\frac{\hat{\lambda}_n - \lambda}{\sigma\sqrt{1+\frac{1}{2}\mu^2_{1-\frac{1}{N}}}} \leq -\mu_{1-\alpha} \right\},$$

σ étant inconnue, si on sature la contrainte pour H_0 $(\lambda = 0)$ *on aura :*

$$W_\alpha = \left\{ \sqrt{n}\frac{\hat{\lambda}_n}{S_n\sqrt{1+\frac{1}{2}\mu^2_{1-\frac{1}{N}}}} \leq -\mu_{1-\alpha} \right\} = \left\{ \overline{X}_n + S_n\left(\mu_{1-\frac{1}{N}} + \mu_{1-\alpha}\sqrt{1+\frac{1}{2}\mu^2_{1-\frac{1}{N}}}\right) - s \leq 0 \right\}$$

7.3 Tests asymptotiques généraux

7.3.1 Fondement

Notons $\hat{\theta}_n$ l'estimateur de vraisemblance de $\theta \in \Theta \subset \mathbb{R}^p$.
La normalité asymptotique de $\hat{\theta}_n$ implique que $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ converge en loi vers $\mathcal{N}\left(O,\ I_1^{-1}(\theta)\right)$

Lemme 7.3.1. *(**Slutsky**) Pour* $\hat{\theta}_n$ *asymptotiquement efficace, on* $n^t\left(\hat{\theta}_n - \theta\right) I^{-1}(\theta)\left(\hat{\theta}_n - \theta\right)$ *converge en loi vers* $\mathcal{X}^2_{(q)}$ *Sit g la fonction vectorielle définie de* $\Theta \subset \mathbb{R}^p$ *dans* $\mathbb{R}^p$. *g est donc une fonction régulière. On aura*

$$\sqrt{n}\left(g(\hat{\theta}_n - g(\theta))\right) \ converge\ loi\ vers\ \mathcal{N}\left(0,\ \frac{\partial g(\theta)}{\partial\theta} I^{-1}(\theta)\frac{\partial g(\theta)^t}{\partial\theta}\right)$$

Soit Θ_0 une partie de Θ telle que

$$\Theta_0 = \left\{\theta \ / \ g_i(\theta) = 0\ ,\ 1 \leq i \leq q \leq p\right\} \quad On\ veut\ tester \quad \begin{cases} H_0 : & \theta \in \Theta_0 \\ Contre & \\ H_1 : & \theta \in \Theta_1 \end{cases}$$

Cette situation est fréquente en économétrie lorsqu'il s'agit de tester la nullité de l'effet d'un bloc de variables exogènes sur la variable endogène . On se place dans le modèle $\left(X;\ \left\{P_{\underline{\theta}},\ \underline{\theta} \in \Theta\right\}\right)$

Définition 7.3.1. *Dans ce modèle, un test de région critique* $W_{\alpha n}$, *de niveau asymptotique α est dit convergent ssi :*

$$\forall\theta \in \overline{\Theta}_0,\ P_\theta(W_{\alpha n})\ tend\ vers\ 1\ lorsque\ n\ tend\ vers\ +\infty$$

Le risque de deuxième espèce tend vers 0 *lorsque n tend vers* $+\infty$.

Pour résoudre le problème de test posé, on utilise trois tests équivalents qui sont convergents

7.3.2 Test de WALD

La fonction $g(\theta) = \begin{pmatrix} h(\theta) \\ k(\theta) \end{pmatrix}$, $H_0 : h(\theta) = \begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}$ $On\ a\ \ \theta = \theta_0 \Longrightarrow h(\theta_0) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$,
g étant régulière, on a $\widehat{g(\theta)} = g(\hat{\theta}_n)$. Par conséquent, sous H_0, on écrit le Lemme de Slutsky comme précédemment et obtient que $\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta_0)\right)$ converge en loi vers $\mathcal{N}_q\left(0,\ M(\theta_0)\right)$ $M(\theta_0)$ est une matrice $q \times q$ de variance - covariance

$$M(\theta_0) = \left(\frac{\partial g(\theta)}{\partial \theta}\right)_{\theta=\theta_0} I^{-1}(\theta_0) \left(\frac{\partial g(\theta)}{\partial \theta}\right)^t_{\theta=\theta_0}$$

On a ainsi

$$n^t\left(g(\hat{\theta}_n) - g(\theta_0)\right)\left(M(\theta_0)\right)^{-1}\left(g(\hat{\theta}_n) - g(\theta_0)\right)\ qui\ converge\ vers\ \mathcal{X}^2_{(q)}$$

sous H_0, la région critique du test de WALD est

$$W_{\alpha n} = \left\{n^t\left(g(\hat{\theta}_n) - g(\theta_0)\right)\left(M(\theta_0)\right)^{-1}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \geq \mathcal{X}^2_{(q),1-\alpha}\right\}$$

Sous H_1, on a

$$^t\left(g(\hat{\theta}_n) - g(\theta_0)\right)\left(M(\theta_0)\right)^{-1}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \geq 0$$

Donc $n^t\left(g(\hat{\theta}_n) - g(\theta_0)\right)\left(M(\theta_0)\right)^{-1}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \to +\infty$. Ainsi, $\lim_{n\to+\infty} P_{\theta\in\overline{\Theta}_0}(W_{\alpha n}) = 1$ Le test de WALD est convergent.

7.3.3 Test du rapport des maxima de vraisemblance

Soit $\mathcal{L}_X(\hat{\theta})$ la vraisemblance sous l'hypothèse $H_0 : \theta \in \Theta_0$,

$$\mathcal{L}_X(\hat{\theta}) = Max\left(\mathcal{L}_X(\theta),\ \theta \in \Theta_0\right)$$

Or $\hat{\theta}$ est le $Max\left(\mathcal{L}_X(\theta),\ \theta \in \Theta\right)$ Par conséquent, si H_0 est vraie, $\mathcal{L}_X(\tilde{\theta})$ doit être proche de $\mathcal{L}_X(\hat{\theta}_n)$. Le test du rapport des maxima de vraisemblance est basée sur la région critique :

$$W_\alpha = \left\{X \in \mathbb{R}^n \ / \ \frac{\mathcal{L}_x(\tilde{\theta})}{\mathcal{L}_x(\hat{\theta})_n} \geq k_\alpha\right\}\ ou\ \left\{X \in \mathbb{R}^n \ / \ \frac{\mathcal{L}_X(\hat{\theta}_n)}{\mathcal{L}_X(\tilde{\theta})} < k_\alpha\right\}$$

Pour trouver k_α, il faut connaître la loi asymptotique du rapport $\frac{\mathcal{L}_X(\hat{\theta}_n)}{\mathcal{L}_X(\tilde{\theta})}$.

On va faire un développement limité de Taylor à l'ordre 1 de $log\left(\frac{\mathcal{L}_X(\hat{\theta}_n)}{\mathcal{L}_X(\tilde{\theta})}\right)$ qui donne

$$log\left(\mathcal{L}_X(\hat{\theta}_n)\right) - log\left(\mathcal{L}_X(\tilde{\theta})\right) \simeq \left(\frac{\partial log \mathcal{L}_X(\theta)}{\partial \theta}\right)^t_{\theta=\hat{\theta}_n}(\hat{\theta}_n - \tilde{\theta}) - \frac{1}{2}(\hat{\theta}_n - \tilde{\theta})I^{-1}(\theta_n)(\hat{\theta}_n - \tilde{\theta}) + \mathcal{O}(\hat{\theta}_n - \tilde{\theta})$$

Or $\dfrac{\partial log\mathcal{L}_X(\theta)}{\partial\theta} = S_X(\theta)$ (la fonction score) et $S_X(\hat{\theta}_n) = 0$. On définit la statistique $Dev = 2\left[log\mathcal{L}_X(\hat{\theta}_n) - log\mathcal{L}_X(\tilde{\theta})\right]$ qui asymptotiquement a la même loi que la statistique $T = (\hat{\theta}_n - \theta)^t I^{-1}(\hat{\theta}_n)(\hat{\theta}_n - \theta)$. Dev est appelé **la deviance**. Par conséquent, en vertu des résultats précédents, on a :

Théorème 7.3.1. *$Dev = 2\left[log\mathcal{L}_X(\hat{\theta}_n) - log\mathcal{L}_X(\tilde{\theta})\right]$ converge en loi vers $\mathcal{X}^2_{(q)}$. La région critique est donc :*

$$W_\alpha = \left\{ Dev \geq \mathcal{X}^2_{(q),1-\alpha} \right\}$$

Remarque 7.3.1. *La statistique du test de rapport des maxima de vraisemblance est plus facile à calculer que celle du test de WALD.*

7.3.4 Test des scores(Test du multiplicateur de Lagrange)

On considère $\tilde{\theta}$; l'estimateur du maximum de vraisemblance de θ sous la contrainte $g_i(\theta) = 0,\ 1 \leq i \leq q \leq p$ $\begin{cases} \tilde{\theta} = MaxLog\mathcal{L}_X(\theta) \\ S.C \quad h(\theta) \end{cases}$ $\tilde{\theta}$ est solution de l'équation $\dfrac{\partial}{\partial\theta}\left(Log\mathcal{L}_X(\theta) - \lambda h(\theta)\right) = 0$ sous les conditions de normalité asymptotique de l'EMV. Sous l'hypothèse H_0, la statistique $\sqrt{n}\dfrac{\partial log\mathcal{L}_X(\tilde{\theta})}{\partial\theta}$ converge en loi vers $\mathcal{N}\left(0\,,\ \dfrac{\partial g(\theta)}{\partial\theta}M^{-1}(\theta)\dfrac{\partial g(\theta)^t}{\partial\theta}\right)$
On obtient le théorème sous H_0. La statistique

$$S = n\left[\frac{\partial}{\partial\theta}log\mathcal{L}_X(\tilde{\theta})\right]^t I(\tilde{\theta})\left[\frac{\partial}{\partial\theta}log\mathcal{L}_X(\tilde{\theta})\right]$$

converge en loi vers $\mathcal{X}^2_{(q)}$. On en déduit la région critique du test des scores

$$W_\alpha = \left\{(x_1,\cdots,x_n)\ /\ S \geq \mathcal{X}^2_{(q)}\right\}$$

Ces trois tests sont équivalents. On s'attend à ce que les conclusions obtenues des trois tests soient les mêmes.

Chapitre 8

Tests de comparaison d'échantillons (Tests empiriques)

8.1 Position du problème

On dispose de deux échantillons etraits a priori de loi différentes et indépendants l'un de l'autre. Échantillon E^1 : $(X_1^1, \cdots; X_n^1)$ de loi P_{θ_1} . Échantillon E_2 : $(X_1^2, \cdots, X_n^2)$ de loi P_{θ_2}. Le problème de test est $\begin{cases} H_0 : & \theta_1 = \theta_2 \\ Contre & \\ H_1 : & \theta \neq \theta_2 \end{cases}$ ou $\begin{cases} H_0 : & \theta_1 \leq \theta_2 \\ Contre & \\ H_1 : & \theta_1 > \theta_2 \end{cases}$

8.2 Comparaison de deux fréquences

8.2.1 Comparaison d'une fréquence expérimentale et d'une fréquence théorique (test de conformité)

Dans une population P, on étudie un caractère dichotomique prenant les modalités A et $\overline{A}$. Soit p la fréquence d'apparition du caractère (ou la proportion ou pourcentage) A dans la population ; f la fréquence d'apparition de A sur un échantillon de P.

Question : l'échantillon observé est-il représentatif de la population P, c'est-à-dire la différence observée entre p et f est-elle dûe aux fluctuations d'échantillonnage ?

La taille n de l'échantillon doit être telle que $n \geq 30$, $np > 5$ *et* $n(1-p) > 5$. Soit F la variable aléatoire représentant la fréquence d'apparition de A, F prend la valeur f sur l'échantillon. La statistique $U = \dfrac{F - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$ Sa valeur sur un échantillon est $u = \dfrac{f - p}{\sqrt{\frac{p(1-p)}{n}}}$.

Règles de décision pour un seuil α donné.

Test bilatéral

$\begin{cases} H_0 : & f = p \\ Contre & \\ H_1 : & f \neq p \end{cases}$ En notant $\mu_{1-\frac{\alpha}{2}}$ le fractile de $\mathcal{N}(0,\ 1)$, si $-\mu_{1-\frac{\alpha}{2}} \leq u \leq \mu_{1-\frac{\alpha}{2}}$, on rejette H_0 au seuil α.

Test unilatéral

$\begin{cases} H_0 : & f \geq p \\ Contre & \\ H_1 : & f < p \end{cases}$ Si $u < -\mu_{1-\alpha}$ on rejette H_0.

$$\begin{cases} H_0 : & f \leq p \\ Contre & \\ H_1 : & f > p \end{cases} \quad Si\ \ u > \mu_{1-\alpha}\ on\ rejette\ \ H_0$$

8.2.2 Comparaison de deux fréquences expérimentales (Test d'homogénéité)

Soient P_1 et P_2 deux populations (ou deux sous populations d'une même population P). Le caractère dichotomique est onservé sur les deux populations où la fréquence d'apparition de la modalité A est p_1 dans P_1 et p_2 dans P_2.

De P_1, on extrait un échantillon E_1 de taille n_1, sur lequel la fréquenced'apparition de A est f_1.
De même pour P_2, E_2 de taille n_2 de fréquence f_2.
Supposons qu'on a : $n_1 \geq 30,\ \ n_1p_1 \geq 5,\ \ n_1f_1 \geq 5,\ \ n_1(1-p_1) \geq 5\ \ et\ \ n_1(1-f_2) \geq 5$
$n_2 \geq 30,\ \ n_2p_2 \geq 5,\ \ n_2f_2 \geq 5,\ \ n_2(1-p_2) \geq 5,\ \ n_2(1-f_2) \geq 5$
Notons F_1 la variable prenant la valeur f_1 sur E_1, et F_2 la variable prenant f_2 sur E_2. Notons $\hat{p} = \dfrac{n_1f_1 + n_2f_2}{n_1 + n_2}$ la statistique $U = \dfrac{(F_1 - F_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \rightsquigarrow \mathcal{N}(0,\ 1)$

Sur E_1 et E_2 la statistique U prend la valeur $u = \dfrac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$

Règles de décision :

Test bilatéral

$\begin{cases} H_0 : & p_1 = p_2 \\ Contre & \\ H_1 : & p_1 \neq p_2 \end{cases}$ Sous H_0, $u = \dfrac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ Si $-\mu_{1-\frac{\alpha}{2}} \leq u \leq \mu_{1-\frac{\alpha}{2}}$, on ne rejette pas H_0.

Test unilatéral

$$\begin{cases} H_0 : & p_1 \geq p_2 \\ Contre & \\ H_1 : & p_1 < p_2 \end{cases} \quad ou \quad \begin{cases} H_0 : & p_1 \leq p_2 \\ Contre & \\ H_1 : & p_1 > p_2 \end{cases}$$

8.3 comparaison de moyennes

8.3.1 Comparaison d'une moyenne expérimentale et une moyenne théorique (Test de conformité)

Dans la population P, la variable X a pour espérance $E(X) = \mu$, et de variance $Var(X) = \sigma^2$. On extrait de P un échantillon de taille n, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ prend la valeur $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$, la variance estimée de X est $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ et prend la valeur $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2$

Problème :

La différence observée entre $\overline{x}_n$ et μ est-elle dûe aux fluctuations d'échantillonnage ? On a les règles de décision suivantes, au seuil de α donné.

La statistique $U = \dfrac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \rightsquigarrow T(n-1)$ (on suppose ici que $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$).

Test bilatéral

$$\begin{cases} H_0 : & \overline{X}_n = \mu \\ Contre & \\ H_1 : & \overline{X}_n \neq \mu \end{cases}$$

U prend la valeur $u = \dfrac{\overline{x}_n - \mu}{s_n/\sqrt{n}}$ sur l'échantillon. On rejette H_0 si l'on a : $-t_{1-\frac{\alpha}{2}}(n-1) \leq u \leq t_{1-\frac{\alpha}{2}}(n-1)$, où $t_{1-\frac{\alpha}{2}}(n-1)$ est le fractile de la loi de Student $T(n-1)$.

Test unilatéral

$$\begin{cases} H_0 : & \overline{X}_n \geq \mu \\ Contre & \\ H_1 : & \overline{X}_n < \mu \end{cases}$$

Si $u \geq -t_{1-\alpha}(n-1)$, on ne rejette pas H_0.

Remarque 8.3.1. *Pour n grand, on a :* $t_{1-\alpha}(n-1) \simeq \mu_{1-\alpha}$ *et* $t_{1-\frac{\alpha}{2}}(n-1) \simeq \mu_{1-\frac{\alpha}{2}}$.

8.3.2 Comparaison de deux moyennes expérimentales (Test d'homogénéité dans deux échantillons indépendants)

Population $P_1 : \mu_1, \sigma_1^2$; on extrait un échantillon E_1 : $\hat{\mu}_1 = \overline{X}_1$, *et* $\hat{\sigma}_1^2 = S_1^2$.

Population $P_2 : \mu_2, \sigma_2^2$; on extrait un échantillon E_2 : $\hat{\mu}_2$, $\hat{\sigma}_2^2 = S_2^2$. On veut vérifier si P_1 et P_2 ont la même moyenne. Pour cela, on considère la différence entre $\overline{X}_1$ et $\overline{X}_2$. La statistique

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0,\ 1)$$

Pour n_1 et n_2 grands, si σ_1^2 et σ_2^2 inconnues, on les estiment par $\hat{\sigma}_1^2 = S_1^2$ et $\hat{\sigma}_2^2 = S_2^2$. On note ainsi,

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad et \quad U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow \mathcal{N}(0,\ 1)$$

Cas des petits échantillons ($n_1 < 30$ ou $n_2 < 60$) : On fait l'hypothèse que les observations suivent la loi gaussienne.

Cas $\sigma_1 = \sigma_2 = \sigma$

Ka variable $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow T(n_1 + n_2 - 2)$, σ n'étant pas connu, on l'estime par

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

Sous H_0, $\mu_1 = \mu_2$, la variable T prende la valeur $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

Test bilatéral : $\begin{cases} H_0 : & \mu_1 = \mu_2 \\ Contre & \\ H_1 : & \mu_1 \neq \mu_2 \end{cases}$

On rejette H_0 si $-t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) \leq t \leq t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$

Test unilatéral : $\begin{cases} H_0 : & \mu_1 \geq \mu_2 \\ Contre & \\ H_1 : & \mu_1 < \mu_2 \end{cases}$ On rejette H_0 si $t \leq t_{1-\alpha}(n_1 + n_2 - 2)$.

$\begin{cases} H_0 : & \mu_1 \leq \mu_2 \\ Contre & \\ H_1 : & \mu_1 > \mu_2 \end{cases}$ On rejette H_0 si $t \leq t_{1-\alpha}(n_1 + n_2 - 2)$.

Cas $\sigma_1 \neq \sigma_2$

On note $S_d^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ Soit λ l'entier tel que $\frac{1}{\lambda} = \frac{1}{n_1-1}\left(\frac{S_1^2/n_1}{S_d^2}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2/n_2}{S_d^2}\right)^2$.

La variable $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_d^2}} \rightsquigarrow T(\lambda)$ ***On prendra λ l'entier le plus proche.***

Sous l'hypothèse H_0, la variable T prend la valeur : $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{S_d^2}}$ On calcule t sur les échantillons et on effectue les tests comme précédement.

Remarque 8.3.2. *Pour faire ces tests, il faut au préable tester l'égalité des variances des deux populations.*

8.3.3 Commparaison de deux moyennes expérimentales dans le cas d'échantillons appariés

Définition 8.3.1. *Les échantillons E_1 et E_2 sont appariés (associés par paires) si à chaque valeur X_i^1 de E_1 est associée une valeur X_i^2 de E_2.*

Exemple 8.3.1. ***(Santé)***

E_1 : groupe de malades avant application du traitement.

E_2 : groupe de même malades après traitement.

on calcule la variable $d_i = X_i^1 - X_i^2$ On a ainsi l'échantillon $(d_1, \cdots, d_n)$ qui a pour moyenne $\overline{d} = \frac{1}{n}\sum_{i=1}^n d_i$ et variance $S_d^2 = \frac{1}{n-1}\sum_{i=1}^n d_i^2 - \overline{d}^2$. La variable $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_d^2}{n}}}$ converge en loi vers la loi de Student $T(n-1)$. Le test consiste à comparer d à 0. $\begin{cases} H_0 : & \overline{d} = 0 \\ \textit{Contre} & \\ H_1 : & \overline{d} \neq 0 \end{cases}$

<u>**Conclusion**</u> **:** Si $n > 30$, on utilise la loi normale. Par contre, si $n \leq 30$ on utilise la loi de Student.

8.4 Comparaison de deux variances

Il s'agit ici des observations suivant des lois gaussiennes.

8.4.1 Variance théorique et variance expérimentale (Test de conformité)

La variable X définie sur la population P. $E(X) = \mu$ et $Var(X) = \sigma^2$. On constitue un échantillon E de P de taille n, de moyenne $\overline{X}$ et de variance S^2.

Question : La différence entre σ^2 et S^2 est elle dûe aux fluctuations d'échantillonnage ?
On a : $X \rightsquigarrow \mathcal{N}(\mu,\ \sigma^2)$, $Y^2 = (n-1)\frac{S^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{(n-1)}$ Sur l'échantillon Y^2 prend la valeur $y^2 = \frac{(n-1)}{\sigma^2}s^2$

test bilatéral

$$\begin{cases} H_0 : \quad S^2 = \sigma^2 \\ Contre \\ H_1 : \quad S^2 \neq \sigma^2 \end{cases}$$ au seuil α donné. On cherche a et b tels que

$P(a < Y^2 \leq b) = 1-\alpha \quad a = \mathcal{X}^2_{\frac{\alpha}{2}}(n-1) \ \ et \ \ b = \mathcal{X}^2_{1-\frac{\alpha}{2}}(n-1).$ $a \leq y^2 \leq b$, on ne rejette pas H_0

Pour $n > 30$, on utilise l'approximation normale $\sqrt{2Y^2} - \sqrt{2\nu - 1} \rightsquigarrow \mathcal{N}(0,\ 1)$, ν étant le nombre de degré de liberté du $\mathcal{X}^2$.

8.4.2 Comparaison de deux variances expérimentales

En reprenant les mêmes notations sur les échantillons E_1 et E_2. On veut tester

$$\begin{cases} H_0 : \quad \sigma_1^2 \leq \sigma_2^2 \\ Contre \\ H_1 : \quad \sigma_1^2 > \sigma_2^2 \end{cases}$$ On sait que $(n_1-1)\frac{S_1^2}{\sigma_1^2} \rightsquigarrow \mathcal{X}^2_{(n_1-1)}$ et $(n_2-1)\frac{S_2^2}{\sigma_2^2} \rightsquigarrow \mathcal{X}^2_{(n_2-1)}$ indépendamment.

$\dfrac{(n_1-1)\frac{S_1^2/\sigma_1^2}{n_1-1}}{(n_2-1)\frac{S_2^2/\sigma_2^2}{n_2-1}} = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \rightsquigarrow F(n_1-1,\ n_2-1)$ La variable $T = \frac{S_1^2}{S_2^2}$ est paramétrée par σ_2^2/σ_1^2.

Posons $\theta = \sigma_2^2/\sigma_1^2$ La loi de T est RVM croissant, donc il existe un test UPP_α de région critique : $W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ \frac{S_1^2}{S_2^2} \geq k \right\}$ pour tester les hypothèses $\begin{cases} H_0 : \quad \theta \leq 1 \\ Contre \\ H_1 : \quad \theta > 1 \end{cases}$ au seuil α.

Avec $P_{\theta=1}(W_\alpha) = \alpha$

$$W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ \frac{S_1^2}{S_2^2} \geq F_{1-\alpha}(n_1-1,\ n_2-1) \right\}$$

Pour tester les hypothèses

$$\begin{cases} H_0 : \quad \sigma_1^2 = \sigma_2^2 \\ Contre \\ H_1 : \quad \sigma_1^2 \neq \sigma_2^2 \end{cases} \qquad W_\alpha = \left\{ (x_1, \cdots, x_n) \ / \ \frac{S_1^2}{S_2^2} \leq F_{\alpha_1}(n_1-1,\ n_2-1) \ ou \ \frac{S_1^2}{S_2^2} \geq F_{1-\alpha_1}(n_1-1,\ n_2-1) \right\}$$

Avec $0 \leq \alpha_1 < \alpha$

Remarque 8.4.1. *On prendra le soin de mettre au numérateur la variance S_i^2 la plus grande.*

Chapitre 9

Tests de l'adéquation du modèle : Tests non paramétriques

9.1 Contexte

Dans les problèmes de test précédents on a fait l'hypothèse que les observations provenaient de famille spécifiée dans le modèle, or face aux obserations et ayant un modèle à leur proposer, il est nécessaire de soumettre ce modèle à une évaluation de son aptitude à rendre compte de la distribution du phénomèse étudié : il s'agit de l'adéquation du modèle des observations.
L'adéquation peut être juger d'un point de vue descriptif : QQ plot, PP gaussien, Droite de Henry. Elle peut être testée rigoureusement, dans ce cas si P est la vraie loi inconnue des observations et P_0 la loi supposée être suivie par les observations. On va tester :

$$\begin{cases} H_0 \ : \ P = P_0 \ \ (il \ y \ a \ ad\dot{-}quation) \\ Contre \\ H_1 \ : \ P \neq P_0 \ \ (il \ n'y \ a \ pas \ ad\dot{-}quation) \end{cases}$$

9.2 Le théorème de Pearson

Soit P_0 une loi discrète sur Ω fini. X une variable aléatoire définie sur $\Omega = \{a_1, \cdots, a_n\}$. X peut être qualitative ou ordinale. Soit P la vraie loi de X inconnue et $p_j = P(X = a_j)$ Soit $(X_1, \cdots, X_n)$ un échantillon de X de taille n, $\quad N_j = Card\{a_j\} \quad f_n^j = \dfrac{N_j}{n}$

Théorème 9.2.1.

$$n \sum_{j=1}^{k} \frac{(f_n^j - p_j)^2}{p_j} \ \ converge \ en \ loi \ vers \ \mathcal{X}^2_{(k-1)}$$

$\sum_{j=1}^k \dfrac{(f_n^j - p_j)^2}{p_j}$ *est la distance du* $\mathcal{X}^2$ *entre la distribution théorique* (p_j) *et la distribution empi-*

rique de X*, centrée sur* p_j*. Or* $\forall j$*,* f_n^j *converge presque s≊rement vers* p_j*, donc* $n\sum_{j=1}^{k}\frac{\left(f_n^j - p_j\right)^2}{p_j}$ *converge presque sûrement vers* 0.

9.3 Tests d'adéquation à une loi spécifiée

Soit P_0, une loi connue ; P la vraie loi suivie par X, inconnue. Le problème de test est :

$$\begin{cases} H_0 \,:\, P = P_0 \\ Contre \\ H_1 \,:\, P \neq P_0 \end{cases}$$

9.3.1 Test de KOLMOGOROV

Notons F la fonction de répartition de la variable aléatoire réelle X. F_0 la fonction de répartition de la loi P_0.

Théorème 9.3.1. *Si* F_0 *est la vraie loi des observations, alors, en notant* $K_n = \sup_x |F_n(x) - F_0(x)|$*, on a*

$$\sqrt{n}K_n \text{ converge en loi vers } K$$

Où K est la variable de la loi de KOLMOGOROV qui est fixe et tabulée. La loi de K n'a pas une expression explicite. Sous H_0*, on a :*

$$\lim_{n\to+\infty} P\left[\sqrt{n}K_n \leq t\right] = 1 - \sum_{k=1}^{+\infty}(-1)^k exp(-2k^2t^2)$$

La région critique du test de KOLMOGOROV est donné par $W_\alpha = \left\{\sqrt{n}K_n \geq k_{1-\alpha}\right\}$ *où* $k_{1-\alpha}$ *est le fractile d'ordre* $1-\alpha$ *de la loi de KOLMOGOROV.* $k_{1-\alpha}$ *est lue sur la table.*

Exemple 9.3.1. $n = 1000$, $\sup_x |F_n(x) - F_0(x)| = 0.047$, $\sqrt{n}K_n = 1.486$, $p - value = 0.024$*Le test obtenu est convergent car si* $F = F_0$*, alors* K_n *converge presque sûrement vers* $\sup_x |F(x) - F_0(x)| = \alpha \neq 0$ *et* $\sqrt{n}K_n$ *tend vers* $+\infty$

9.3.2 Test de KUIPER

Dans les conditions du test de KOLMOGOROV, en notant : $K_n^+ = \sup_x |F(x) - F_0(x)|$ *et* $K_n^- = \sup_x |F_0(x) - F_n(x)|$ On pose $V_n = K_n^+ + K_n^-$

KUIPER a montré que $\sqrt{n}V_n$ converge en loi vers Q, Q étant la variable d'une loi fixe tabulée. La région critiqueb du test est :

$$W_\alpha = \left\{\sqrt{n}V_n \geq q_{1-\alpha}\right\}$$

$q_{1-\alpha}$ étant le fractile d'ordre $1-\alpha$ de la loi de KUIPER.

Remarque 9.3.1. *En procédant par simulation, on montre que le test de KUIPER est plus puissant que le test de KOLMOGOROV.*

Autres tests : Anderson-Darling, Cramer-Von MISES. Deux échantillons indépendants issus d'une même population.

9.3.3 Test d'adéquation du $\mathcal{X}^2$

En utilisant la distance du $\mathcal{X}^2$, Si P_0 est la vraie loi de X, de probabilité p_0^i
$n\sum_{i=1}^{k}\frac{\left(f_n^i - p_0^i\right)^2}{p_0^i}$ converge en loi vers $\mathcal{X}^2_{(k-1)}$ La région critique du test est donnéepar

$$W_\alpha = \left\{ n\sum_{i=1}^{k}\frac{\left(f_n^i - p_0^i\right)^2}{p_0^i} \geq \mathcal{X}^2_{1-\alpha}(k-1) \right\}$$

Le test est auusi convergent.

Remarque 9.3.2. *L'utilisation de ce résultat asymptotique impose en pratique que les effectifs théoriques (np_i) et les effectifs empiriques (nf_n^i) soient supérieurs ou égaux à 5. D'où éventuellement un regroupement des modalités X_i de la variable X.*

9.4 Test du $\mathcal{X}^2$ d'adéquation à une famille de lois

on désire tester l'appartenance de la vraie loi inconnue P à une famille de lois, paramétrique plutôt que l'adéquation à une loi spécifique connue. On aura les hypothèses : $\begin{cases} H_0 \ : \ P \in \{P_\theta,\ \theta \in \Theta\} \\ Contre \\ H_1 \ : \ P \notin \{P_\theta,\ \theta \in \Theta\} \end{cases}$
P_θ doit être la loi la plus vraisemblable possible, ce qui suppose d'estimer θ par la méthode du maximum de vraisemblance. Puis on utilise la statistique de $\mathcal{X}^2$ pour tester l'adéquation de cette loi aux observations. Cependant les mêmes observations sont utilisées deux fois. Premièrement pour calculer de l'estimateur du maximum de vraisemblance, et deuxièmement pour calculer la valeur de la statistique du test.
Fisher montre que si $\theta \in \Theta \subset \mathbb{R}^p$, si $\hat{\theta}_n$ est l'EMV de θ, si $p_i(\hat{\theta}_n)$ est la probabilité de la loi du modèle correspondant à l'EMV $P_{\hat{\theta}_n}$, on a :

$$n\sum_{i=1}^{k}\frac{\left(f_n^i - p^i_(\hat{\theta}_n)\right)^2}{p^i(\hat{\theta}_n)} \doteq converge\ en\ loi\ vers\ \mathcal{X}^2_{(k-1)}$$

p est le nombre de paramètre à estimer. Le nombre de degré de liberté (ddl) du $\mathcal{X}^2$ est diminué de p. La région critique du test de niveau α s'en déduit :

$$W_\alpha = \left\{ n \sum_{i=1}^{k} \frac{\left(f_n^i - p_i(\hat{\theta}_n)\right)^2}{p_i(\hat{\theta}_n)} \geq \mathcal{X}^2_{1-\alpha}(k-p-1) \right\} \text{ ou } \left\{ \sum_{i=1}^{k} \frac{\left(n_i - p_i(\hat{\theta}_n)\right)^2}{np_i(\hat{\theta}_n)} \geq \mathcal{X}^2_{1-\alpha}(k-p-1) \right\}$$

avec $n_i = nf_n^i$

La même remarque sur les effectifs reste valable.

Le test de KOLMOGOROV est plus puissant que le test de $\mathcal{X}^2$

x_i	n_i	f_i	$p_i\ np_i$	F_n	F_0	$\lvert F_n(x) - F_0(x)\rvert$	$\frac{(n_i - np_i)^2}{np_i}$
x_1	n_1	f_1	p_1	$np_1\ F_1$	F_{01}		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	p_i	$np_i\ F_i$	F_{0i}		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Total	n	1	1	n			$\mathcal{X}^2$

9.5 Comparaison de l'échantillon ($l > 2$) : Test d'homogénéité

Sur la population P, un caractère regroupé en k groupes A_i. On dispose de l'échantillon ($l > 2$) $E_1, \cdots, E_l$ de la population P. $\forall i \in \{1, \cdots, k\}$, $\forall j \in \{1, \cdots, l\}$ on note O_{ij} =effectif observé de A_i dans l'échantillon j. L'effectif total $N = \sum_{j=1}^{l} \sum_{i=1}^{k} O_{ij}$

On veut tester l'hypothèse H_0 : les différences observées dans les groupes entre les différents échantillons (extraits d'une même population P) sont dues aux fluctuations d'échantillonnage.

Procédé :

- Calcul des effectifs théoriques c_{ij} sous H_0. On a $p_i = P(A_i) = \dfrac{\sum_{j=1}^{l} O_{ij}}{N} = \dfrac{N_{i.}}{N}$
  L'effectif calculé du groupe A_i dans l'échantillon E_j
  $C_{ij} = p_{i.}\left(\sum_{j=1}^{l} O_{ij}\right) = p_i \times N_{.j} = \dfrac{N_{i.} \times N_{.j}}{N}$

La statistique du test est

$$Y^2 \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(O_{ij} - C_{ij}\right)^2}{C_{ij}} \rightsquigarrow \mathcal{X}^2_{(k-1)(l-1)}$$

En pratique, il faut s'assurer que $C_{ij} \geq 5$; sinon, on regroupe des catégories A_i. La région critique du test est donnée par $W_\alpha = \left\{Y^2 \geq \mathcal{X}^2_{1-\alpha}((k-1)(l-1))\right\}$

9.6 Test d'indépendance entre deux caractères dans un tableau de contingence

Dans une population P, on mesure sur chaque individu les caractères A et B. $A = \left\{A_1; \cdots, A_k\right\}$ *et* $B = \left\{B_1; \cdots, B_l\right\}$ Le tableau fournit O_{ij} =effectifs de (A_i, B_j), l'effectif total est $N = \sum_{i=1}^{k} \sum_{j=1}^{l} O_{ij}$.
L'hypothèse à tester est H_0 : A et B sont deux caractères indépendants.
En utilisant les mêmes notations qu'en haut, on a : $N_{i.} = \sum_{j=1}^{l} O_{ij}$ *et* $N_{.j} = \sum_{i=1}^{k} O_{ij}$
sous H_0 (A et $B0$ indépendants), $P(A_i, B_j) = P(A_i) \times P(B_j) = \dfrac{N_{i.}}{N} \times \dfrac{N.j}{N}$ $\forall i, \forall j$.
Les effectifs théoriques $C_{ij} = N \times P(A_i, B_j) = \dfrac{N_{i.} \times N_{.j}}{N}$.

La statistique du test est $Y^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \dfrac{(O_{ij} - C_{ij})^2}{C_{ij}} \rightsquigarrow \mathcal{X}^2_{(k-1)(l-1)}$ appelé Khi-deux de contingence.
La région critique du test est $W_\alpha = \left\{Y^2 \geq \mathcal{X}^2_{1-\alpha}((k-1)(l-1))\right\}$

Remarque 9.6.1. *On doit s'assurer que le* $C_{ij} \geq 5$*; sinon regrouper des caractères dans A ou B.* $Y^2 = N\left[\sum_{i=1}^{k}\sum_{j=1}^{l} \dfrac{N_{ij}^2}{N_{i.} \times N_{.j}} - 1\right]$

9.7 Test du coefficient de corrélation

Le coefficient de corrélation entre deux variables X et Y est :

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E\left[(X - E(X))(Y - E(Y)\right]}{\sqrt{E\left[(X - E(X))^2\right] \times \left[(Y - E(Y))^2\right]}}$$

Sur l'échantillon (X_i, X_j), avec $i = 1 \cdots n$ et $j = 1 \cdots n$, l'estimateur de ρ est :

$$\hat{\rho} = r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}}$$

On peut donc tester l'absence de corrélation entre X et Y sur la base d'un échantillon. on teste alors les hypothèses suivantes : $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$

Si X et Y sont issues des lois normales, c'est-à-dire que X et Y sont deux caractères statistiques mesurés sur les mêmes individus.

$X \rightsquigarrow \mathcal{N}(m_x, \sigma_x^2)$, $Y \rightsquigarrow \mathcal{N}(m_y, \sigma_y^2)$ et $(X, Y) \rightsquigarrow \mathcal{N}(m, \sum)$ avec $\sum = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}$

STUDENT démontre que sous H_0 on a :

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \rightsquigarrow T(n-2)$$

La région critique du test pour un seuil α donné est :

$$W_\alpha = \{(x, y) \ / \ |r| > c\}$$

La fonction $r \longmapsto \dfrac{r}{\sqrt{1-r^2}}$ étant une fonction croissante, alors $|r| > c$ implique $\dfrac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > \dfrac{c\sqrt{n-2}}{\sqrt{1-c^2}}$ donc $\dfrac{c\sqrt{n-2}}{\sqrt{1-c^2}}$ est le fractile d'ordre $1-\dfrac{\alpha}{2}$ de la loi de Student $T(n-2)$:

$$\frac{c\sqrt{n-2}}{\sqrt{1-c^2}} = t_{1-\frac{\alpha}{2}}(n-2) \Longrightarrow c = \frac{t_{1-\frac{\alpha}{2}}(n-2)}{\sqrt{n-2+\left(t_{1-\frac{\alpha}{2}}(n-2)\right)^2}}$$

Remarque 9.7.1. .

1. *Pour ce test, il est courant d'utiliser la transformation de Fisher lorsque la taille de l'échantillon est grande i.e $n > 30$, la dite transformation de Fisher est :* $Z = \frac{1}{2}ln\left(\frac{1+r}{1-r}\right) = Argth(r)$ *La variable Z* ***suit la loi normale de paramètres*** $E(Z) = \frac{1}{2}ln\left(\frac{1+\rho}{1-\rho}\right)$ *et* $Var(Z) = \dfrac{1}{n-3}$

2. *Cette transformation de Fisher, en sens inverse, permet de calculer l'intervalle de confiance asymptotique pour ρ. On a*

$$\frac{1}{2}ln\left(\frac{1+r}{1-r}\right) - \frac{\mu_{1-\alpha/2}}{\sqrt{n-3}} \leq \frac{1}{2}ln\left(\frac{1+\rho}{1-\rho}\right) \leq \frac{1}{2}ln\left(\frac{1+r}{1-r}\right) + \frac{\mu_{1-\alpha/2}}{\sqrt{n-3}}$$

3. ***Si les variables X et Y ne sont pas gaussiennes****, les résultats précédents restent valables à condition que $n > 30$ Cependant, $r = 0$ n'implique pas que X et Y sont indépendantes.*

Avec la transformation de Fisher, on peut faire le test courant suivant :

$$\begin{cases} H_0 \,:\, \rho = \rho_0 \\ H_1 \,:\, \rho \neq \rho_0 \,,\; \rho_0 \, donn\dot{-} \end{cases}$$

La région critique du test est : $W_\alpha = \{|r| > c\}$ avec $P_{H_0}(W_\alpha) = \alpha$. On obtient

$$c = \frac{\pm t_{1-\alpha/2}(n-2)}{\sqrt{n-2+\left(t_{1-\alpha/2}(n-2)\right)^2}}$$

Chapitre 10

Quelques anciens sujets

Statistique & Probabilités - IFORD Voie B session 2025

Exercice 1

Un étudiant en Statistique veut organiser un jeu de hasard avec deux dés bien équilibrés, Moyennant une mise qu'il lui faut déterminer, un joueur devra lancer les dés : s'il obtient un double "6", on lui remettra quatre fois la mise ; si c'est un double (autre que des "6"), il recevra le double de sa mise initiale. A combien l'organisateur du jeu doit-il établir la mise s'il espère faire un profit net moyen de 2 dollar par joueur ?

Exercice 2

On a évalué que si le nom d'une personne est sur la liste d'attente d'un centre hospitalier pour faire du bénévolat sur demande, la probabilité d'être appelée d'un mois est de 0,3.
La liste d'un Hôpital est de 30 noms.

1. Quelle est la probabilité que plus du tiers des personnes de cette liste soient appelées à l'intérieur d'un mois ?
2. Sachant que 7 personnes ont déjà été appelées, que devient la probabilité que plus du tiers des personnes de cette liste soient appelées à l'intérieur d'un mois ?
3. Si la liste était de 60, combien peut-on espérer appeler de personnes à l'intérieur d'un mois ?

Exercice 3

Soient X et Y, deux variables aléatoires indépendantes de densités respectives f et g et de fonctions de répartition respectives F et G.

1. Donner les densités de $U = \sup\{X, Y\}$ et $V = \inf\{X, Y\}$.

2. Soient $X_1, ..., X_n$, n variables aléatoires indépendantes, de même loi, de densité :

$$f(x) = 1_{[0,1]}(x)$$

Écrire les densités de $U_n = \sup X_{i_{1\leqslant i\leqslant n}}$.

Exercice 4

Soit X une variable aléatoire réelle continue représentant la mesure d'un phénomène physique dans une unité donnée. On donne sa densité : $g(x) = \lambda e^{-\lambda x}$, $x > 0$, $\lambda > 0$. Or la donnée observée est en réalité $[X]$, où $[.]$ désigne la partie entière de X.

1. On pose $Y = [X]$. Donner la loi de Y, son espérance et sa variance.
2. Que dire lorsque la donnée observée est, cette fois, l'entier le plus proche ?

Exercice 5

La variable aléatoire X suit une loi de Pareto si sa densité est donnée par : $f(x) = \dfrac{\alpha\theta^\alpha}{x^{\alpha+1}} 1_{[\theta,+\infty[}(x)$ $(\alpha > 0, \theta > 0)$.

1. Calculer $E\left(X^k\right)$ en précisant les conditions d'existence.
2. Déterminer le mode et la médiane de X.

Exercice 6

Soit $(X_n)_{n\in\mathbb{N}^*}$, une suite de variables aléatoires indépendantes, de même loi, de densité :

$$f(x) = e^{-(x-\theta)} 1_{[\theta;+\infty[}(x), \quad \theta > 0$$

.

1. Donner la loi de $Y_n = \min X_{i_{1\leqslant i\leqslant n}}(x)$.
2. Montrer que Y_n converge en probabilité et en moyenne quadratique (L_2) vers θ.
3. Vers quelle loi converge la variable aléatoire $n(Y_n - \theta)$?

Exercice 7

Un échantillon de 6 États a donné les résultats suivants pour les variables X "consommation annuelle de cigarettes par tête" et Y "taux de mortalité due au cancer des poumons pour 100 000" (Fraumini, 1968).

État	X	Y
Delaware	3400	24
Indiana	2600	20
Iowa	2200	17
Montana	2400	19
New Jersey	2900	26
Washington	2100	20
Moyennes	2600	21

1. Calculer le coefficient de corrélation r de l'échantillon. Un tel coefficient de corrélation calculé à partir de données agrégées est appelé coefficient de corrélation "écologique".
2. Déterminer un intervalle de confiance à 95% pour le coefficient de corrélation ρ.
3. Tester, au seuil d'erreur de 5%, l'absence de liaison entre la consommation de cigarettes et le cancer du poumon.

Exercice 8

Soient X, Y, Z trois variables aléatoires réelles telles que :

— X suit une loi uniforme sur $]0,1[$ de densité : $f(x) = 1_{]0,1[}(x)$.

— la loi conditionnelle de Y pour $X = x$ fixé admet pour densité :

$$f_{Y|X=x}(y) = (y-x)e^{-(y-x)}1_{]x,+\infty[}(y)$$

.

— la loi conditionnelle de Z pour $(X,Y) = (x,y)$ fixé admet pour densité :

$$f_{Z|(X,Y)=(x,y)}(z) = (y-x)e^{-z(y-x)}1_{\mathbb{R}_+^*}(z)$$

.

1. Calculer la densité du triplet (X, Y, Z).
2. Calculer la densité de Z et celle de la loi conditionnelle de (X, Y) sachant $Z = z$.
3. Calculer : $E\left(\sqrt{Y-X}|Z=z\right)$ et $E\left(\sqrt{Y-X}\right)$.

   On donne $\Gamma\left(\frac{1}{2}\right) = \int_0^{+\infty} u^{-\frac{1}{2}}e^{-u}du = \sqrt{\pi}$.
4. On pose $U = Y - X$ et $V = Z(Y - X)$. Montrer que les trois variables X, U, V sont mutuellement indépendantes et exprimer les densités de U et de V.

Statistique & Probabilités - IFORD Voie A session 2025

Exercice 1

Une cabine pour skieurs est conçue pour une charge totale-limite de 10 000 livres. La capacité est limitée à 50 personne. On suppose que le poids moyen de tous les usagers de la cabine est de 190 livres et l'écart-type de 25 livres. Quelle est la probabilité que le poids total d'un groupe de 50 personnes choisies au hasard soit plus élevé que la charge limite de 10 000 livres ?

Exercice 2

On considère un échantillon suffisamment petit qui fera l'objet d'un suivi répété, afin d'estimer le nombre de repas précuisinés dans une grande ville. Le taux de réponse élevé signifie que l'échantillon est essentiellement aléatoire. On a enregistré, lors de l'enquête, le nombre de repas "prêts à consommer" pris par chaque individu au cours de la semaine précédente ; et en résumé l'on obtient : $\bar{X} = 0,82$, $s = 0,48$ et $n = 80$. Calculer l'intervalle de confiance à 95% pour la moyenne de la population totale de cette ville.

Exercice 3

On a tiré de façon indépendante un échantillon de 15 professeurs d'une grande université : 10 hommes et 5 femmes, dont les salaires annuels sont les suivants (en milliers de dollars) :

Hommes (X_1)		Femmes (X_2)
13	20	9
11	14	12
19	17	8
25	14	10
22	15	16
$\bar{X}_1 = 16$		$\bar{X}_2 = 11$

Ces moyennes échantillon donnent une estimation grossière des moyennes μ_1 et μ_2 des populations sous-jacentes. Peut-être est-il possible de s'en servir pour mettre un point final à une vieille querelle : un époux prétend qu'il n'y a pas de différence entre les salaires des hommes (μ_1) et ceux des femmes (μ_2). En d'autres termes, si on écrit la différence sous la forme $\Delta = \mu_1 - \mu_2$, il prétend que : $\Delta = 0$.

Son épouse, au contraire, prétend que la différence est de 7000 dollars : $\Delta = 7$.

Régler cette controverse, en construisant un intervalle de confiance à 95%.

Exercice 4

L'industrie pour laquelle vous travaillez vient d'acquérir une nouvelle machine-outil pour découper les tiges de métal nécessaires à la fabrication d'essieux. La notice du fabriquant assure que la longueur des tiges coupées suit une loi normale avec un écart-type de $0,15$ mm. Avant de lancer toute la production, vous désirez avoir une idée très précise sur la longueur moyenne des tiges qui seront coupées. Sachant que vous travaillez avec un niveau de confiance de 99%, déterminez la taille d'échantillon nécessaire pour cerner cette longueur avec une précision inférieure à $0,02$ mm.

Exercice 5

Les recherches en santé mentale montrent que le trouble de personnalité le plus fréquent est celui de la personnalité limite, atteignant en grande majorité des femmes (75%).

Sur 50 dossiers de cas de personnalité limite, quelle est la probabilité qu'il y ait au moins 40 dossiers de femmes ?

Exercice 6

Supposons que la distribution a priori d'une pente β de régression soit normale, avec une valeur centrale de $5,0$ et un écart-type de $0,50$. Supposons en plus qu'un échantillon de 8 observations (X, Y) ait été prélevé, et qu'il fournisse les statistiques suivantes :

$$\sum xy = 2100, \qquad \sum x^2 = 350, \qquad s \text{ résiduel} = 12,7$$

Supposons que σ est inconnu, de sorte que s remplacera σ quand il le faudra, et, en conséquence, $t_{0,025}$ remplacera $1,96$. Dans ce cas, calculer :

1. L'intervalle de confiance classique à 95%.
2. L'intervalle de confiance bayésien à 95%. Comment les comparer ?

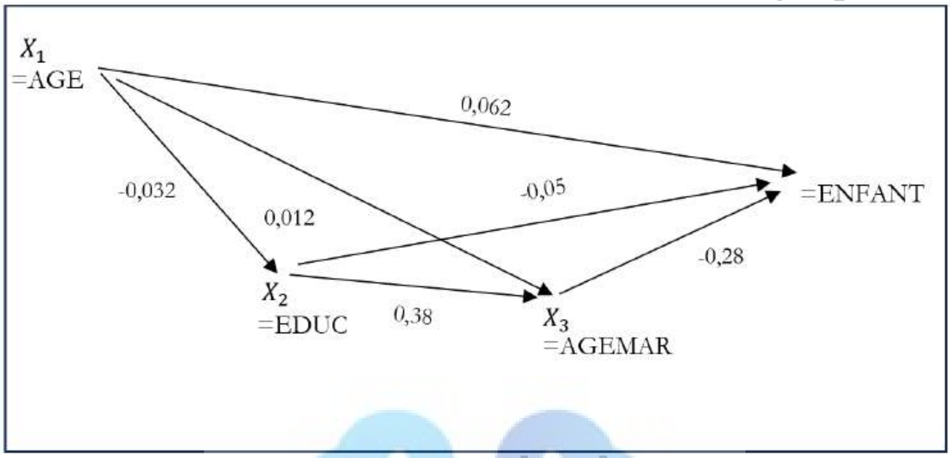
Exercice 7

Une étude plus complète de la fécondité (nombre d'enfants, ENFANT) de 4700 femmes dans un pays, a inclus un autre facteur : $X_3 = AGEMAR$ =l'âge de la femme quand elle s'est mariée. L'ensemble des variables à prendre en compte devient alors celui de la figure ci-dessous. On a effectué la régression de chaque facteur en fonction de toutes les variables situées à gauche ; les équations qui en résultent sont les suivantes :

— $Y = 0,062X_1 - 0,05X_2 - 0,28X_3$
— $X_3 = 0,012X_1 + 0,38X_2$
— $X_2 = -0,032X_1$

On a fait figurer les valeurs de ces coefficients sur la figure ci-dessous. Calculer l'effet total sur Y (nombre d'enfants) :

1. Des études X_2.
2. De l'âge X_1.
3. De l'âge au mariage X_3.



Exercice 8

Dans une école, on a mis sur pied un club des petits déjeuners car on estime à 15% la proportion des enfants de cette école qui ne déjeunent pas le matin. Afin d'étudier ce phénomène social, on forme un échantillon aléatoire constitué de 10 enfants.

1. Déterminer la loi de probabilité, l'espérance mathématique et la variance du nombre d'enfants de l'échantillon qui ne déjeunent pas le matin.
2. Quelle est la probabilité qu'il y ait moins de 4 enfants de l'échantillon qui ne déjeunent pas le matin ?
3. Quelle est la probabilité qu'au moins 7 enfants de l'échantillon déjeunent le matin ?
4. Quelle est la probabilité que 20% des enfants de l'échantillon ne déjeunent pas le matin ?

Statistique & Probabilités - IFORD Voie B session 2024

Exercice 1

Soient X et Y deux variables aléatoires indépendantes de loi $N(0,1)$. On définit la variable aléatoire $Z = \frac{X}{Y}$.

1. (a) Exprimer la densité de la loi de Z dite loi de Cauchy.
   (b) Comparer les positions respectives de la loi de Cauchy et la loi $N(0,1)$.

2. La loi de Z admet-elle des moments ?
3. Calculer la densité de la variable aléatoire $U = \theta + \sigma Z$ ($\theta \in \mathbb{R}$, $\sigma > 0$).

Exercice 2

Soit $\Omega = \{1, 2, ..., n\}$ ($n \geqslant 1$).

1. Définir une probabilité P sur Ω telle que : $i \in \Omega$, $P(\{i\}) = \frac{1}{n}$.
2. Trouver une condition nécessaire et suffisante sur n pour que : $\exists (A, B) \in F(\Omega) - \{\varphi, \Omega\}$, A et B indépendants.
3. On note :
   $\forall p \in \Omega$, $A_p = \{i \in \Omega / p$ divise $i\}$, $K = \{p,\ p$diviseur premier de $n\}$.
   $\varphi(n) = \text{Card}\{q \in \Omega / q$ est premier avec $n\}$.
   (a) Montrer que si $p_1, ..., p_k$ sont éléments de K, alors $A_{p_1}, ..., A_{p_k}$ sont indépendants.
   (b) Montrer que $\varphi(n) = n \prod_{p \in K} \left(1 - \frac{1}{p}\right)$.

Exercice 3

On a mis au point un test pour déceler une maladie qui provoque la débilité chez les enfants dans une proportion de 620 sur 100000. Les premiers essais sur la fiabilité du test ont été très encourageants : sur 500 enfants testés et qui avaient la maladie, seulement 5 ont eu un résultat erroné avec un test négatif, alors que sur 2000 autres enfants testés, et qui n'étaient pas atteints par la maladie, seulement 50 ont eu un résultat erroné, avec un test positif.
Pour voir dans quelle mesure ce test sera fiable pour tester à grande échelle l'ensemble de la population, calculer :

1. Si un enfant présente une réaction positive, à combien peut-on parier qu'il ait réellement la maladie ? Et avec quelle probabilité ?
2. Si un enfant présente une réaction négative au test, quel est alors le pari qu'il soit malade ? Et avec quelle probabilité ?
3. Si le test pour cette maladie ne comportait aucune erreur, quelles seraient les probabilités en 1) et en 2) ?

Exercice 4

Au défi étudiant, un des jeux consiste à lancer une pièce de 25 sur une table carrée de 1 m de côté, située à 3 m du lanceur. La planche est séparée en 100 carrés identiques. Un des carrés est rouge : s'il est touché par la pièce, 20 points sont attribués à l'équipe.
Considérant qu'il y a eu 400 lancers dans la soirée, quell est la probabilité qu'il y ait eu au moins

de 4 attributions de 20 points ? (Utilisez la loi binomiale, puis examinez ce que vous auriez obtenu en évaluant approximativement la probabilité désirée à l'aide d'une loi de Poisson).

Exercice 5

On considère n variables aléatoires indépendantes (X_i) de densité f_i. On définit les probabilités : $P_i = \int_{-\infty}^{X_i} f_i(t)dt$.

Montrer que $P = -\sum_{i=1}^{n} 2 \ln P_i$ suit une loi $\chi^2(2n)$.

Exercice 6

En 2004, un certain virus informatique a fait beaucoup de ravage durant les cinq jours qui ont suivi sa première apparition dans un pays. Dans une Université de ce pays, on prétend que 17, 9% des courriels étaient infectés. Pour convaincre ses membres de se procurer un bon anti-virus, l'association des chargés de cours a prélevé un échantillon de 50 courriels (sans remise) parmi les 10500 recus par ses membres durant ces 5 jours. Déterminer la probabilité qu'entre 7 et 12 de ces courriels étaient infectés.

Exercice 7

Lors d'une étude sur le reboisement dans un secteur donné, 100 arbres sont choisis aléatoirement. A l'aide des données ci-dessous, et sachant que le diamètre moyen obtenu est 260, 4 mm et d'écart type, 30, 88 mm, peut-on penser, avec un seuil fixé à 10%, que le diamètre suit une loi normale ?

Diamètres des arbres (en mm)	Nombres d'arbres
De 170 à moins de 190	1
De 190 à moins de 210	6
De 210 à moins de 230	8
De 230 à moins de 250	16
De 250 à moins de 270	34
De 270 à moins de 290	22
De 290 à moins de 310	7
De 310 à moins de 330	4
De 330 à moins de 350	2
Total	100

Exercice 8

Afin de comparer trois variétés de pommes de terre, on a fait l'expérience suivante : chaque variété fut plantée au hasard sur trois parcelles de terre de taille égale, mais ayant chacune trois types de sol différents. On a observé les rendements suivants, en boisseaux par parcelle :

Sol	Variété de pomme de terre		
	A	B	C
Sable	21	20	16
Argile	16	18	11
Terreau	23	31	24

1. Construire le tableau ANOVA.
2. Calculer les intervalles de confiance à 95% pour les différences entre les trois variétés.
3. Le botaniste qui a développé la variété B a remarqué qu'il avait travaillé dix ans pour découvrir un produit qui pousse bien sur du terreau. Si on observe les données, peut-on dire qu'il a réussi ? A la lumière de cette information, que pourrait-on dire quant à l'analyse effectuée en 1) et en 2) ?

Statistique & Probabilités - IFORD Voie A session 2024

Exercice 1

On lance deux dés équilibrés de couleur différente. On considère les évènements suivants :

Q : le premier dé, le dé vert, donne 4.
T : le second dé, le dé rouge, donne 3.
A : la somme des résultats donne 7.

Que pouvez-vous affirmer au sujet de l'indépendance de A et Q ? A et T ? et $Q \cap T$?

Exercice 2

Parmi les 530 étudiants embauchés pendant la période des vacances par le Service des Loisirs de la ville, 13% sont âgés de plus de 21 ans. Quelle est la probabilité que sur un échantillon de 100 dossiers de ces employés choisis au hasard et avec remise, au moins 20% soient des employés de plus de 21 ans ?

Exercice 3

Jerry et Poupou se sont intéressés aux résultats scolaires de 110 étudiants et à leurs habitudes concernant le tabac. Les résultats montrent que ces étudiants ont des résultats scolaires excellents, satisfaisants et médiocres dans les proportions de 20%, 60% et 20%. De plus, ils ont repertorié que 30 des 110 étudiants fumaient beaucoup, 30 autres modéraient, et les autres, pas du tout.

Jerry et Pougou sont sûrs qu'il y a indépendance entre les résultats scolaires et la consommation de tabac. Sans plus tarder, ils construisent un tableau de contingence illustrant la répartition des étudiants selon ces deux aspects. Qu'obtiennent-ils ?

Exercice 4

Un facteur doit distribuer n lettres dans n boites aux lettres dont on a retiré les plaques, les rendant indiscernables. Comme les n lettres portent n noms différents, le facteur suppose qu'il y a une lettre et une seule à distribuer dans chaque boite et il dispose donc les lettres au hasard. Calculer la probabilité qu'une lettre qu moins soit distribuée à son destinataire réel. Calculer sa limite quand n croit.

Exercice 5

Le comité de conservation des rivières d'une région examine la concentration de zinc à partir de 12 échantillons prélevés sans remise dans les rivières de la région afin de savoir si les résultats ont été modifiés depuis l'introduction des nouvelles lois. Ils obtiennent une moyenne de $2,6$ grammes par millilitre. En supposant que l'écart type de la concentration de zinc de toutes leurs rivières est le même que celui qui avait été trouvé l'an dernier, à savoir $0,3g/ml$, déterminer un intervalle de confiance pour la concentration moyenne avec un niveau de confiance de 90%.

Exercice 6

Une boite contient 12 billets numérotés de 1 à 12. Sarah en tire 3, sans remise. Soit X la variable aléatoire représentant le plus élevé des trois numéros tirés.

1. Donner la distribution de probabilité de X et sa représentation graphique.
2. Quelle est la probabilité que le numéro tiré le plus élevé soit supérieur à 7 ?

Exercice 7

Le temps requis par Chloé pour se rendre au travail suit une loi Normale de moyenne 47 et d'écart-type 8 minutes.

1. Quelle est la probabilité qu'elle prenne plus d'une heure un certain matin pour se rendre au travail ?
2. Si on considère que le temps requis un certain matin pour se rendre au travail n'influence pas le temps requis les autres matins, quelle est la probabilité que sur 60 matins, Chloé prenne plus d'une heure pour se rendre au travail plus de 5 fois ?

Exercice 8

Un jeu de rôles est organisé. Chaque participant doit choisir au hasard une profession (dentiste, notaire ou médecin), un moment clé de sa vie professionnelle (actif ou retraité) et un des deux traits de caractère suivants : pessimiste ou optimiste. on note les choix de chacun. (Aucun choix n'a plus de notoriété qu'un autre ...)

1. Illustrer les possibilités de rôles par un diagramme en arbre.

   On considère que tous les rôles sont équiprobables. Un rôle est choisi au hasard.
2. Quelle est la probabilité de tout rôle élémentaire qui peut être choisi ?
3. Quelle est la probabilité que le rôle choisi soit celui d'un retraité optimiste ?
4. Quelle est la probabilité que le rôle choisi soit celui d'un médecin ou d'un optimiste ?

   On suppose maintent que tous les rôles ne sont pas équiprobables et qu'on a :

$P(DAP) = 0,025$	$P(NAO) = 0,1$	$P(MRP) = 0,1$
$P(NAP) = 0,05$	$P(MAO) = 0,2$	$P(DRO) = 0,025$
$P(MAP) = 0,1$	$P(DRP) = 0,025$	$P(NRO) = 0,1$
$P(DAO) = 0,025$	$P(NRP) = 0,05$	$P(MRO) = 0,2.$

5. Que devient la probabilité que le rôle choisi soit celui d'un retraité optimiste ?
6. Que devient la probabilité que le rôle choisi soit celui d'un médecin ou d'un optimiste ?

Statistique & Probabilités - IFORD Voie B session 2023

Exercice 1

Soit l'expérience consistant à effectuer 4 fois le tir avec remise d'une boule à partir d'un sac qui en contient 130 : 10 blanches, 20 noires et 100 vertes. On s'intéresse aux quantités possibles de boules blanches tirées et à la probabilité que cela se produise.

1. Déterminer l'épreuve de *Bernoulli* sous-jacente à cette expérience.
2. Si B représente le nombre de boules tirées, déterminer sa loi de probabilité.

Exercice 2

Le nombre de rhumes attrapés par un individu en l'espace d'un an est une variable aléatoire de Poisson de paramètre $\lambda = 5$. Un vaccin préventif est mis sur le marché par une compagnie pharmaceutique. Lors de sa présentation au congrès des pharmaciens, la compagnie ne promettait pas "absence totale de rhumes durant l'année pour les vaccinés", mais affirmait plutôt que pour 75% des vaccinés, le paramètre sera abaissé λ à 3.

Un individu reçoit ce vaccin. Étant donné qu'il a attrapé exactement deux rhumes durant l'année, quelle est la probabilité que le vaccin ait eu l'effet escompté sur lui ?

Exercice 3

Trois hommes font un travail similaire d'empaquetage. Le nombre de boites empaquetées par chacun, au cours de trois heures déterminées, est donné dans le tableau suivant :

Heure	Homme		
	A	B	C
11 - 12	24	19	20
13 - 14	23	17	14
16 - 17	25	21	17

Calculer le tableau ANOVA, en y incluant la probabilité critique pour les deux hypothèses nulles.

Exercice 4

Un échantillon aléatoire de 4 notes est prélevé dans un amphithéâtre : 64, 66, 89 et 77. Dans un autre amphithéâtre, on tire de façon indépendante un échantillon aléatoire de 3 notes. : 56, 71 et 53. Calculer l'intervalle de confiance pour la différence entre les deux moyennes des amphis : $\mu_1 - \mu_2$.

Exercice 5

Un échantillon aléatoire de 7 femmes montrait une pente positive du taux de cholestérol dans le sang, selon l'age. Si la pente des moindres carrés était de 0,024, avec un écart-type de 0,010, quelle serait la pente bayésienne réduite ?

Exercice 6

Dans une cafétéria, du lundi au vendredi inclusivement, le distributeur de lait fonctionne environ 800 fois par semaine. La quantité de liquide versé à chaque remplissage suit une loi normale de moyenne225 ml et d'écart-type 8 ml.

1. Quelle est la probabilité que la moyenne de liquide contenue dans 20 verres remplis au hasard, durant une semaine donnée, soit de plus de 228 ml ?
2. Quelle est la probabilité qu'un verre déborde si sa capacité est de 235 ml ?
3. Déterminer un intervalle centré à la moyenne dans lequel on retrouvera 75% des volumes moyens de liquide contenu dans des échantillons de 20 verres remplis au hasard, durant la même semaine.
4. Le responsable de la cafétéria a reçu de nouveaux verres, lesquels ne peuvent contenir plus de 220 ml. Il sait que le distributeur doit subir un ajustement s'il ne veut pas qu'une majorité de verres débordent lors du remplissage. Si le le technicien n'ajuste pas la dispersion, à combien doit-il ajuster la moyenne afin qu'au plus 10% d'un échantillon de 25 de ces nouveaux verres, remplis la même semaine, ne débordent lors du remplissage ?

Exercice 7

Soit $X = (X_i)_{\text{i=1 à n}}$, un vecteur aléatoire gaussien de dimension n, de loi $N(O, I_n)$.
Soit $\theta = (\theta_i)_{\text{i=1 à n}}$ un vecteur de $\mathbb{R}^n$. On définit la variable aléatoire :

$$Y = \|X + \theta\|^2 = \sum_{i=1}^{n}(X_i + \theta_i)^2.$$

Y suit une loi du khi-deux décentrée à n ddl et de paramètre d'excentricité (ou de décentrage) : $\lambda = \sum_{i=1}^{n} \theta_i^2 = \|\theta\|^2$. On la note : $\chi^2(n, \lambda)$.

1. Montrer que la loi de Y ne dépend que de la norme de θ et non de θ lui-même.
2. Calculer son espérance et sa variance.
3. Si $X = (X_i)_{\text{i=1 à n}}$ sont n variables aléatoires indépendantes de lois respectives $N(m_i, \sigma_i)$. On dit que $U = \sum X_i^2$ suit une loi du khi-deux décentrée généralisée. Calculer : $E(X)$ et $V(X)$.

Exercice 8

Des tests effectués sur une surface particulièrement rugueuse afin de mesurer l'endurance des pneus de camions fabriqués selon un nouveau procédé. Or, sur cette surface, il s'avère que des crevaisons se produisent pour 25% de ces pneus. Pour le prochain essai, 15 de ces nouveaux pneus seront mis à l'épreuve.

1. Quelle est la probabilité que moins de 4 crevaisons auront lieu ?
2. Déterminer μ et σ, ainsi que la probabilité annoncée par l'inégalité de Chebyshev au sujet de l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$, et interpréter cette probabilité dans le contexte.
3. Si X représente le nombre de pneus crevés, déterminer $\mathbb{P}(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma)$ à l'aide du modèle suivi par X.

   Deux lots de pneus sont formés : un contenant des pneus fabriqués selon le nouveau procédé, l'autre, des pneus fabriqués selon l'ancien. En examinant seulement les rainures des pneus, on sait que 60% des camionneurs affirmeraient préférer rouler avec des pneus provenant du lot A. Or, un autre test est effectué : un camionneur, ayant choisi à l'aveugle 10 pneus provenant d'un même lot, fera rouler son camion sur la surface rugueuse.
4. Sachant que le lot A contient des pneus de l'ancien procédé et que 30% de ces anciens pneus ne passaient pas le test de la surface rugueuse, calculer la probabilité que 2 pneus subissent une crevaison lors du test.

Statistique & Probabilités - IFORD Voie A session 2023

Exercice 1

Vous vous promenez le long d'une voie ferrée, en pleine rêverie. Vous prenez soudain conscience d'un grand bruit, droit derrière vous. En quoi consiste l'erreur de première espèce ? L'erreur de deuxième espèce. Laquelle est la plus grave ? Que faites-vous ?

Exercice 2

Soient X et Y deux variables aléatoires suivant respectivement les lois $\chi^2(p)$ et $\chi^2(q)$ $(p > q)$. Soit Z une variable aléatoire indépendante de Y, telle que $X = Y + Z$.

Montrer que Z suit une loi $\chi^2(p - q)$.

Exercice 3

Une chaine de fabrication produit 40 000 fours dont 32 000 (80%) sont bons, ne nécessitant donc aucune modification. Le service de contrôle-qualité qui ne connait pas ces chiffres, prélève un échantillon aléatoire de 10 fours pour estimer la qualité de l'ensemble de la fabrication.

Quelle est la probabilité que dans cet échantillon, on obtienne 5 fours défectueux et 5 bons ?

Exercice 4

Les deux variables aléatoires X et Y ont la distribution conjointe $p(x, y)$ suivante :

X	Y			
	10	20	30	40
20	0,04	0,08	0,08	0,05
40	0,12	0,24	0,24	0,15

Sont-elles indépendantes ?

Exercice 5

On suppose que les étudiants d'un cours de statistique ont des notes normalement distribuées autour d'une moyenne de 72 avec un écart-type de 9.

1. Calculer la probabilité qu'un seul étudiant choisi au hasard ait une note supérieure à 80.
2. Calculer la probabilité qu'un échantillon de 10 étudiants ait une note moyenne supérieure à 80.
3. Si la population n'était pas normale quelle aurait été la réponse pour la question 2) ?

Exercice 6

Supposer que pour un échantillon aléatoire de 4 familles on observe les revenus annuels et l'épargne correspondante suivante (en milliers de dollars) :

Famille	Revenu X	Épargne S
A	22	2,0
B	18	2,0
C	17	1,6
D	27	3,2

1. Estimer la droite de régression $X = \alpha + \beta X$.
2. Construire l'intervalle de confiance à 95% de la pente β.
3. Faire figurer sur un graphique les 4 points et la droite ajustée ; indiquer alors le ... possible, les diverses pentes acceptables, découlant du calcul de l'intervalle de confiance de la question 2).

Exercice 7

Dans une grande université, on a constitué, en 1969, des échantillons indépendants de 5 professeurs hommes et 5 professeurs femmes, et relevé les traitements annuels ci-dessous :

Femmes	Hommes
9	19
12	9
8	12
10	11
16	22

1. Calculer un intervalle de confiance pour la différence du traitement moyen entre les hommes et les femmes.

2. Dans quelle mesure une telle différence moyenne indique-t-elle une discrimination envers les femmes ?

Exercice 8

En 1954 une expérience fut menée sur une grande échelle pour tester l'efficacité d'un nouveau vaccin contre la polio. Parmi les 740 000 enfants choisis dans les classes primaires à travers le pays A, 400 000 étaient volontaires. La moitié d'entre eux furent choisis au hasard pour recevoir une injection du vaccin, la moitié restante, une injection placebo d'eau salée. Les résultats furent les suivants :

Groupe	Nombre d'enfants	Nombre de cas de polio
Vaccinés	200 000	57
Placebo (témoins)	200 000	142
Non volontaires	340 000	157

1. Pour chacun des trois groupes, calculer le taux de polio (nombre de cas pour 100 000).

2. Estimer la réduction du taux de polio obtenue grâce à la vaccination, et donner l'intervalle de confiance à 95%.

3. Supposer que tous les volontaires ont été vaccinés, les non-volontaires constituant le groupe témoin.

   (a) Avant d'analyser les données, critiquer cette procédure.

   (b) Quelles sortes d'informations obtiendrait-on ? Auraient-elles apporté la réponse correcte à la question 2) ?

Statistique & Probabilités - IFORD Voie B session 2022

Exercice 1

L'entraineur d'une équipe de football affirme que cette année, la stratégie de jeu qu'il préconise en 2^e demie de jeu est meilleure qu'avant. Désireux de vérifier cette "bonne nouvelle", son assistant examine les statistiques de matchs des années antérieures, lesquelles indiquent que la moyenne de points accumulés en 2^e demie est de 35. Avec un seuil de signification de 10%, et en supposant que le nombre de points accumulés suit une loi normale, peut-on affirmer que l'entraineur a raison de vanter sa nouvelle technique si un échantillon de 25 matchs de cette année révèle une moyenne de 38,1 points gagnés en 2^e demie, avec un écart-type de 6,6.

Exercice 2

L'Institut National de la Statistique d'un pays X rapporte qu'en 2008, la proportion de naissances issues de parents non mariés atteignait 59,2%. Le Service de la planification des naissances de IFORDIA croit que cette proportion était plus petite dans sa région. En utilisant un seuil de 5%, peut-on confirmer cette croyance si un échantillon de 55 naissances choisies parmi toutes les naissances ayant lieu à IFORDIA en 2008 révèle que 52,2% d'entre elles sont issues de parents non mariés ?

Exercice 3

L'an dernier, sur un échantillon aléatoire de 254 personnes ayant fait une réservation dans un restaurant, 198 d'entre elles ont demandé une table dans la section non-fumeur. Cette année, sur un échantillon de 322 personnes, il y en a eu 265. Peut-on conclure avec un seuil de 10% que la proportion des demandes pour la section non-fumeur a significativement augmenté ?

Exercice 4

Des composantes électroniques peuvent être assemblées de deux façons différentes pour produire, semble-t-il, des mêmes fonctions. Or, ces montages n'étant pas infaillibles, on a répertorié quatre types de mauvais fonctionnement. En examinant le tableau de fréquences ci-dessus, peut-on conclure à l'indépendance entre les types de mauvais fonctionnement et les deux techniques d'assemblage et ce, avec un seuil de signification de 5% ?

Assemblage	Mauvais fonctionnement			
	Type 1	Type 2	Type 3	Type 4
Assemblage 1	44	23	18	10
Assemblage 2	16	5	6	13

Exercice 5

Soient X et Y deux variables aléatoires indépendantes de loi $N(0,1)$.
On définit la variable aléatoire $Z = \frac{X}{Y}$.

1. Exprimer la densité de la loi de Z dite loi de Cauchy. Comparer les positions respectives de la loi de Cauchy et de la loi $N(0,1)$.
2. La loi de Z admet-elle des moments ?
3. Calculer la densité de la variable aléatoire $U = \theta + \sigma Z$ $(\theta \in \mathbb{R}, \sigma > 0)$.

Exercice 6

Soit $f(x) = ke^{-\frac{|x|}{\theta}}$, θ étant un paramètre réel strictement positif.

1. Déterminer k de façon que f soit la densité de probabilité d'une variable aléatoire X.
2. On pose $Y = \frac{X}{\theta}$. Déterminer la loi de Y, appelée loi de Laplace normalisée.
3. Calculer $E(Y^r)$, $\forall r \in \mathbb{N}^*$. En déduire $E(Y)$ et $V(Y)$.

Exercice 7

Par une expérience aléatoire contrôlée, Schunk et Lehman ont défini en 1954 la relation entre la mortalité infantile et une thérapeutique par l'oxygène. Tous les enfants dont le poids à la naissance était compris entre 1000 et 1850 g et qui étaient nés depuis moins de 12 heures avant leur admission en pouponnière, furent soumis, après tirage au hasard, à des traitements soit à forte, soit à faible concentration d'oxygène. Les résultats furent les suivants :

Concentration en oxygène	Survie (ou non) après trois mois		
	Décédés	Survivants	Total
Forte	9	36	45
Faible	12	28	40
Total	21	64	85

1. Calculer l'intervalle de confiance approprié pour estimer l'effet d'une concentration d'oxygène sur la mortalité infantile.
2. Calculer la probabilité critique pour H_0 (pas d'effet).
3. Donner une interprétation littéraire simple des réponses obtenues en 1) et en 2).

Statistique & Probabilités - IFORD Voie A session 2022

Exercice 1

Un groupe de 21 adultes est formé de 8 anglophones (5 femmes et 3 hommes), les autres étant francophones. Sachant qu'il y'a 2 fois plus d'hommes que de femmes, quelle est la probabilité qu'une personne choisie au hasard...

1. Soit anglophone, sachant que c'est une femme ?
2. Soit une femme, sachant que c'est une personne anglophone ?

Exercice 2

L'autobus passe au coin de la rue à 7h00, 7h15, 7h30, 7h45 et 8h00. Vous devez prendre cet autobus, mais vous ne connaissez pas son horaire. Vous décidez de vous pointer au coin de la rue au hasard, entre 7h et 7h30.

Quelle est la probabilité que votre attente soit de moins de 5 minutes ?

Exercice 3

Votre banque modifie son système de NIP (Numéro d'identification Personnelle). Dorénavant, ces codes seront composés de 2 lettres suivies de 3 chiffres. Combien de codes sont possibles avec ce nouveau système si les lettres et les chiffres doivent être tous différents ?

Exercice 4

Soient X et Y deux variables aléatoires discrètes avec les distributions de probabilité p_X et p_Y et leur distribution conjointe p.

p Y X	3	4	p_X Total
-1	0,1	0,1	0,2
0	0,1	0	0,1
1	0,3	0,2	0,5
2	0,1	0,1	0,2
p_Y Total	0,6	0,4	1

1. Déterminer $E(x)$ et $E(Y)$.
2. Utilisant si possible $E(X)$ et $E(Y)$. déterminer $E(8X + 230)$, $E(X^2)$, $E(X + Y)$ et $E(X, Y)$.

Exercice 5

Lors d'un contrôle de qualité concernant le fonctionnement de souris d'ordinateurs, 113 souris sont déclarées non fonctionnelles sur un échantillon aléatoire de 1200 souris. Estimer, pour l'ensemble des souris fabriquées par la compagnie, la proportion de souris non fonctionnelles avec un niveau de confiance à 95%.

Exercice 6

La distribution des revenus de la population des hommes d'un pays était approximativement la suivante en 1975.

Revenu annuel (en milliers de dollars US)	Proportion d'hommes ayant pour revenu (en %)
0 - 5	36
5 - 10	23
10 - 15	20
15 - 20	11
20 - 25	5
25 - 30	3
30 - 35	2

Exercice 7

Dans une clinique, on cherche à mettre au point un test peu couteux et fiable, pour dépister le diabète. Les résultats obtenus grâce au test sont les suivants : 95% des individus présentant un résultat positif au test (i.e diagnostiqués comme ayant le diabète) sont effectivement diabétiques. 98% des individus n'ayant pas le diabète présentent un test négatif (les 2% restant sont diagnostiqués à tort comme diabétiques). La clinique dessert une population d'environ 10000 patients et le taux de diabète est de 1% environ. Le directeur s'intéresse à 3 résultats :

1. Combien y-a-t-il environ de patients diabétiques qui ne soient pas dépistés comme tels (test négatif) ?
2. Combien environ y-a-t-il de patients pour lesquels le diagnostique est positif (nécessitant donc une surveillance particulière) ?
3. Quelle est la proportion des patients de la question 2 qui sont actuellement diabétiques ?
4. Le directeur ne peut croire à votre réponse de la question 3, expliquez aussi simplement que possible pourquoi cette proposition est aussi faible.

Statistique & Probabilités - IFORD Voie B session 2021

Exercice 1

Une maladie M affecte une personne sur 1000 dans une population donnée. On dispose d'un test sanguin qui détecte M avec une fiabilité de 99% lorsque cette maladie est effectivement présente. Cependant, on obtient aussi un résultat faussement positif pour 0,2% des personnes saines testées.

Quelle est la probabilité qu'une personne soit réellement malade lorsque son test est positif ?

Exercice 2

Soient X et Y deux variables aléatoires indépendantes de loi $N(0, 1)$.
On définit la variable aléatoire $Z = \dfrac{X}{Y}$.

1. Exprimer la densité de la loi de Z dite loi de Cauchy. Comparer les positions respectives de la loi de Cauchy et de la loi $N(0, 1)$.
2. La loi de Z admet-elle des moments ?
3. Calculer la densité de la variable aléatoire $U = \theta + \sigma Z$ ($\theta \in \mathbb{R}$, $\sigma > 0$).

Exercice 3

Afin d'augmenter le nombre de personnes transportées, une compagnie aérienne vend plus de billets qu'elle n'a de places en pariant sur les absences de certaines de ses passagers. On suppose qu'avec une probabilité p, le passager ne se présente pas à l'embarquement. Notons n le nombre de places sur un vol et l le nombre de billets vendus pour ce même vol. La compagnie rembourse le billet si une personne présente à l'embarquement ne peut voyager à cause du surbooking.

On souhaite calculer l afin que la compagnie ne doit rembourser les passagers sue dans les 1% des cas.

Considérons $X_i = \begin{cases} 1 \text{ si le passager se présente à l'embarquement} \\ 0 \text{ si le passager ne se présente pas à l'embarquement} \end{cases}$

On supposera les variables aléatoires X_i indépendantes.

1. Donner la loi de X_i puis celle de $S_i = \sum_{i=1}^{l} X_i$.
2. Donner une expression littérale (exacte), en fonction de p, n et l, de la probabilité que la compagnie soit obligée de rembourser.
3. Par application du théorème central limite, donner une approximation de cette quantité en fonction de n, l, p, $q = 1 - p$ et de la fonction F de répartition de la loi Normale centrée réduite.

4. En déduire l pour $n = 100$ et $p = 0,2$.

Exercice 4

Soit X une variable aléatoire réelle représentant la mesure d'un phénomène physique dans une unité donnée. On donne sa densité $g(x) = \lambda e^{-\lambda x}$, $x > 0$, $\lambda > 0$. Or, la densité observée est en réalité $[X]$ où $[.]$ désigne la partie entière de X.

1. On pose $Y = [X]$. Donner la loi de Y, son espérance et sa variance.
2. Que dire lorsque la donnée observée est, cette fois, l'entier le plus proche ?

Exercice 5

La vitesse des véhicules sur une autoroute provinciale est en moyenne de 112km/h et de variance 225 $\left(\frac{km}{h}\right)^2$. On considère que la loi normale est un bon modèle pour décrire la distribution de cette variable.

1. Théoriquement, la limite maximale de vitesse sur les autoroutes est de 100km/h. Calculer la proportion de véhicules qui roulent à une vitesse supérieure à la limite permise selon le modèle probabiliste utilisé.
2. A quelle vitesse un automobiliste devrait-il rouler sur une autoroute pour que 40% des véhicules roulent moins vite que lui ?
3. Une voie ferrée longe une autoroute et un train y circule à 105 km/h. Quelle est la probabilité que la vitesse du train et celle d'un véhicule choisi au hasard (et roulant dans la même direction que le train), diffèrent de plus de 10km/h ?
4. La distance de freinage (en m) est directement proportionnelle à la vitesse du véhicule (en m/s) au carré. La relation entre ces deux quantités est exprimée par la relation $D = -V^2/2a$, où a est une constante négative représentant la décélérations due à la force de freinage.

   Considérons ici que $a = -6(m/s)^2$. Quelle est la probabilité qu'il faille plus de 100 mètres à une voiture pour s'arrêter sur une autoroute provinciale ?

Exercice 6

Soient X et Y deux variables aléatoires suivant respectivement les lois $chi^2(p)$ et $\chi^2(q)$ $(p > q)$.

Soit Z une variable aléatoire indépendante de Y, telle que $X = Y + Z$.

Montrer que Z suit une loi $\chi^2(p - q)$.

Exercice 7

Soient $n \in \mathbb{N}^*$ et $\theta > 0$. On rappelle qu'une variable aléatoire réelle Y suit la loi $\gamma(n, \theta)$ si elle admet la densité : $f_y(t) = \dfrac{\theta^n}{(n-1)!} t^{n-1} \mathrm{e}^{-\theta t}$ ($t > 0$, on pose $0! = 1$).

1. Déterminer $\displaystyle\int_0^{\infty} t^{n-1} \mathrm{e}^{-\theta t} dt$.
2. Soient Y_1 et Y_2 deux variables aléatoires indépendantes. On suppose que Y_1 suit une loi $\gamma(n, \theta)$ et Y_2 suit une loi $\gamma(1, \theta)$ (loi exponentielle).
   (a) Déterminer la loi de $Y_1 + Y_2$.
   (b) En déduire la loi de la somme de N variables indépendantes et identiquement distribuées de loi $\gamma(1, \theta)$.
3. Soit X une variable aléatoire réelle de densité $f_x(t) = \theta t^{\theta - 1}$ ($0 < t < 1$). Calculer $E(X)$ et $V(X)$.
4. Déterminer la loi de $Z = \log\left(\dfrac{1}{X}\right) = -\log(X)$. Calculer $E(Z)$.
5. On observe un n-échantillon $X_1, X_2, \ldots, X_n$ de variables indépendantes de même loi que X. Trouver l'estimateur du maximum de vraisemblance W_n de θ.
6. Calculer le biais $E(W_n) - \theta$ de l'estimateur.

Statistique & Probabilités - IFORD B 2019 - Cameroun

Exercice 1

Dans une population, la probabilité de naissance d'un garçon est de 0,52. On sait d'autre part que 2% des filles et 1% des garçons présentent à la naissance une luxation congénitale de la hanche. On considère les évènements suivants :

G : " naissance d'un garçon "
L : "naissance d'une fille "
L : " le nouveau-né souffre d'une luxation de la hanche.

1. Déterminer les probabilités de " L sachant G" (notation L/G) et de "L sachant F" (notation L/F).
2. Calculer les probabilités des évènements " G et L" et " F et L ". En déduire la probabilité de L.

Exercice 2

On considère une cible circulaire de rayon r conçue pour un jeu de fléchettes, sur laquelle des cercles de rayon $\dfrac{r}{4}$, $\dfrac{r}{2}$ et $\dfrac{3r}{4}$ ont été dessinés, délimitant ainsi 4 régions différentes.

Considérant que les fléchettes tirées sur cette cible le seront de façon aléatoire, calculer la probabilité d'atteindre chaque région.

Exercice 3

La durée de fonctionnement d'un ordinateur avant sa première panne est une variable aléatoire continue dont la fonction de densité est donnée par :

$$f(x) = \begin{cases} ce^{-\frac{x}{100}} \text{ si } x \geqslant 0 \\ 0 \text{ si } x < 0 \end{cases}$$

1. Quelle est la probabilité que l'ordinateur fonctionne moins de 100 heures avant sa première panne ?
2. Quelle est la probabilité que cette durée de fonctionnement soit comprise entre 50 et 150 heures ?
3. Trouver la fonction de répartition associée à cette variable.
4. Représenter la question 2) en utilisant cette fois la fonction de répartition.

Exercice 4

Des vis et des rondelles de métal sont fabriquées pour s'imbriquer ensemble. Le diamètre du trou de la rondelle suit une loi normale d'écart-type 0,07 mm, alors que le diamètre de la vis suit une loi normale d'écart-type 0,05 mm. On prend une vis et une rondelle au hasard. Quelle est la probabilité que les deux ne puissent pas s'imbriquer ?

1. ... si le diamètre moyen de la vis et celui de la rondelle sont tous deux égaux à 20 mm ?
2. ... si 19,95 mm et 20,05 mm sont les diamètres moyens respectifs de la vis et de la rondelle ?

Exercice 5

Lorsque S succès se produisent dans n épreuves, la proportion de l'échantillon $P = \dfrac{S}{n}$ est habituellement utilisée comme un estimateur de la probabilité du succès π. Mais, parfois, il y a de bonnes raisons d'employer l'estimateur $P^* = \dfrac{S+1}{n+2}$. P^* peut être écrit comme une combinaison linéaire de l'estimateur habituel P : $P^* = \dfrac{nP+1}{n+2} = \left(\dfrac{n}{n+2}\right) P + \left(\dfrac{1}{n+2}\right)$.

1. Quelle est l'erreur quadratique moyenne (ERQM) de P ? Est-il consistant ?
2. Quelle est l'ERQM de P^* ? Est-il consistant ?
3. Pour décider du meilleur estimateur entre P et P^*, a-t-on besoin de la consistance ? De quel critère aurait-on besoin ?

4. Dresser le tableau de l'efficacité de P par rapport à P^*, par exemple lorsque $n = 10$ et $\pi = 0; 0,1; 0,2; ...; 0,9; 1,0$;
5. Indiquer certaines circonstances possibles où l'on pourrait préférer utiliser P^* au lieu de P pour estimer π.

Exercice 6

Dans quelle mesure l'alcool affecte-t-il sérieusement le développement cérébral prénatal ? Pour étudier cette question (Jones et al., 1974), on a retenu six femmes qui ont été des alcooliques chroniques durant la grossesse. On a mesuré le QI de leurs enfants, 7 ans après la naissance des ces derniers. Cet échantillon de taille 6 a donné une moyenne de 78 et $\sum(X - \bar{X})^2 = 1805$.

On a constitué un groupe témoin de 46 femmes représentant en gros les mêmes caractéristiques (mêmes moyennes d'age, éducation, statut conjugal etc.) - n'ayant eu aucune tendance à l'alcoolisme pendant leur grossesse. Leurs 46 enfants ont des résultats de QI qui donnent une moyenne de 99 et $\sum(X - \bar{X})^2 = 11520$.

1. Si cette étude avait réalisée dans le cadre d'une expérience aléatoire, quel serait l'intervalle de confiance pour la différence de QI due à l'alcoolisme prénatal ?
2. Puisqu'il s'agit en fait d'une étude empirique, comment devrait-on modifier l'affirmation de 1) ?

Exercice 7

Dans un échantillon issu d'un lot de radios, 6 étaient défectueuses et 14 étaient bonnes.

1. Calculez l'intervalle de confiance (IC classique) à 95% pour la proportion des radios défectueuses dans le lot.
2. Supposons que le relevé des lots passés ait donné relativement peu de défectueux, et qu'il pourrait être approximé par la distribution à priori.

   $P(\pi) \propto \pi^1(1 - \pi)^9$.

   Calculez l'IC bayésien à 95%. Quelle est la différence avec l'IC classique ?

Statistique & Probabilités - IFORD Voie A session 2019

Exercice 1

Dans le collège X, en Sciences de la nature, le programme est composé à parts égales de filles et de garçons et 60% des élèves veulent se diriger dans le domaine de la santé. De plus, on sait que la probabilité qu'un élève soit un garçon voulant aller dans un domaine autre que celui de la santé est de 0,35.

a) Symboliser les évènements mentionnés et relever les probabilités explicites de cette situation.

Un élève est choisi au hasard. Déterminer la probabilité des évènements suivants :

b) L'élève ne se dirige pas dans le domaine de la santé.

c) L'élève est un garçon se dirigeant dans le domaine de la santé.

d) L'élève est un garçon ou un élève se dirigeant dans le domaine de la santé.

e) L'élève est une fille et elle ne se dirige pas dans le domaine de la santé.

Exercice 2

Vous vous réveillez la nuit avec un mal de tête et vous avalez un cachet provenant d'une bouteille sur votre table de nuit. Peu de temps après, vous vous réveillez avec des maux d'estomac et vous vous rendez compte que sur votre table de nuit, il y a 3 flacons : 2 contiennent un médicament contre le mal de tête, et le troisième contient un anti-inflammatoire qui cause effectivement des maux d'estomac dans 80% des cas. Cependant, votre médecin vous avait averti que les cachets pour le mal de tête pouvaient aussi provoquer des maux d'estomac dans 5% des cas.

Quelle est la probabilité que vous vous soyez trompé de médicament au milieu de la nuit sachant qu'en ce moment, votre estomac vous fait souffrir ?

Exercice 3

Une variable aléatoire peut prendre les valeurs 1, 2, 3 et 4. Sachant que : $P(X = 1) = \frac{1}{4}$, $P(X \leqslant 2) = \frac{3}{4}$ et $P(X \leqslant 3) = \frac{7}{8}$.

1. Déterminer la loi de probabilité de cette variable.
2. Trouver son espérance mathématique et sa variance.

Exercice 4

On tire au hasard 3 boules d'une urne en contenant 3 rouges, 4 blanches et 5 bleues. Soit R et B les variables désignant respectivement le nombre de boules rouges et le nombre de boules blanches tirées.

a) Trouver la distribution de la loi de probabilité P_R de la variable R.

b) Trouver la distribution de la loi de probabilité P_B de la variable B.

c) Calculer les probabilités $P(R = r_i \text{ et } B = b_j)$ pour toutes les valeurs $r_i \in CH(R)$ et $b_j \in CH(B)$ et noter les résultats en utilisant la notation $p(r_i, b_j) = p(R = r_i \text{ et } B = b_j)$.

d) Créer un tableau à double entrée avec en tête de lignes, les valeurs de $CH(R)$ et en tête colonnes, les valeurs de $CH(B)$. A l'intersection des lignes et des colonnes, inscrire les probabilités $p(r_i, b_j) = p(R = r_i$ et $B = b_j)$ correspondantes. Ajouter une colonne et une ligne "Total". Que remarque-t-on ?

Exercice 5

Une étude nationale sur les ménages a montré que le taux d'écoute d'une certaine émission télévisée était de 30% pour les femmes et de 50% pour les hommes. Elle a montré par ailleurs que si la femme suivait cette émission, la proportion des maris la suivant en même temps passait à 60%. On tire un couple au hasard. Quelle est la probabilité que :

a) Les 2 conjoints suivant l'émission ?

b) Au moins l'un des 2 suive l'émission ?

c) Aucun des 2 ne suive l'émission ?

d) Si le mari la suit, la femme la suive aussi ?

e) Si le mari ne la suit pas, la femme la suive ?

Exercice 6

Pour un lot de 380 arbres de taille suffisante pour qu'ils soient vendus, on a mesuré, en pieds, la diamètre du tronc à mi-hauteur (D) et la hauteur utilisable (H) de chaque arbre. Les 380 couples de mesures ont été classés dans le tableau suivant (les effectifs ont été convertis en fréquences relatives).

D	H	
	20	25
1,0	0,16	0,09
1,25	0,15	0,30
1,50	0,03	0,17
1,75	0,00	0,10

Comme on peut considérer qu'un tronc est à peu près cylindrique, le volume V de bois correspondant à un tronc est $V = 0,4D^2H$.

a) Calculer le volume moyen de bois par arbre, $E(V)$.

b) L'écart-type de V.

c) Quel est le volume global du lot de 380 arbres ?

d) Calculer le diamètre moyen $E(D)$ et la hauteur moyenne $E(H)$.

Est-il exact que $E(V) = 0,4\,[E(D)]^2\,[E(H)]^2$?

Statistique & Probabilités - IFORD A 2019 - Cameroun

Exercice 1

Un employé, responsable du contrôle de la qualité des lampes électriques, doit tester avec un seuil de signification de 5 %, la durée moyenne théorique de 2500 heures d'une certaine marque de lampes de 60 watts. il ait que la durée de vie de ces lampes suit une loi normale d'écart type 55 heures. La moyenne obtenue pour un échantillon de 20 lampes est de 2479 heures. Que décidera-t-il ?

Exercice 2

Newton fait application pour le recrutement de clients par téléphone : il doit vendre de l'assurance solde pour les détenteurs de carte de crédit d'une grande chaine de magasin. La probabilité qu'une personne rejointe par le téléphone accepte de s'assurer est de 0,22.

1. Quelle est la probabilité que le premier acheteur soit la personne rejointe au $4^{\text{ième}}$ appel ?
2. Quelle est la probabilité que Newton doive faire au moins 4 appels avant d'avoir le premier acheteur ?
3. Sachant que ce matin, Newton n'a pas eu de succès avec ses 2 premiers appels, quelle est la probabilité qu'il n'ait pas d'acheteurs pour les 5 premiers appels d'aujourd'hui ?
4. Pensant à tous les employés comme Newton, combien doivent-ils faire d'appels en moyenne pour avoir un premier acheteur ?

Exercice 3

Un chien dépisteur de drogue a été entrainé selon une nouvelle méthode. A l'aéroport d'une ville A, on rapporte qu'il détecte en moyenne 1,7 cas de passation de drogue par semaine et que dans la vile B sa moyenne hebdomadaire passe à 2,3 cas.

1. Quelle est la probabilité que ce chien détecte plus de 5 cas en 2 semaines dans la ville A ?
2. On raconte dans le milieu que l'an dernier, à un des 2 aéroports, ce chien à déjà détecté 5 cas en une semaine. Sachant cela, quelle est la probabilité que ce soit dans la ville A si le chien a été utilisé 2 semaines sur 3 dans la ville B ?

Exercice 4

On se demande s'il existe une association entre les notes en statistique et en français des élèves d'une classe de terminales. Voir tableau ci-après :

Individus	1	2	3	4	5	6	7	8	9	10
Note en statistiques (X)	12	7	13	15	8	11	9	17	15	6
Note en français (Y)	15	18	5	12	3	16	15	13	8	14

1. Définir la population étudiée.
2. Quel test doit-on utiliser pour répondre à cette préoccupation ?
3. Calculer le coefficient de corrélation linéaire entre les deux séries X et Y.

Exercice 5

1. Esquisser le graphe de la fonction suivante et montrer que c'est une fonction de densité $f(x)=$

$$\begin{cases} \frac{x}{4} \ \ si \ \ 0 < x < 1 \\ \frac{5x}{4} - 1 \ \ si \ \ 1 \leq x < 2 \\ 0 \quad ailleurs \end{cases} \tag{10.1}$$

2. Soit X une variable aléatoire qui obéit à cette fonction de densité.
3. Déterminer sa fonction de répartition.
4. Calculer les probabilités suivantes : $P(X < 0,5)$, $P(X \geq 1,5)$, $P(0,5 < X < 1,5)$.
5. Calculer son espérance et sa variance.

Exercice 6

Le samedi midi, comme il y'a peu d'affluence à la cafétéria du collège, on n'y offre que deux options : le repas complet ou un gouter. Samedi dernier, 45 personnes, dont 32 locataires de la résidence, se sont présentées à cette cafétéria. Il y'a eu 30 repas complets vendus et 75 % des locataires de la résidence en ont acheté un. Afin de bien répondre à la demande, la responsable de la cuisine offre toujours un tableau complet de ventes pour chaque période de repas, tenant compte de la nature des clients (locataires ou non) et du type de repas acheté.

1. Qu'obtiendra-t-elle comme tableau de contingences pour ce samedi midi ?

   On choisit une facture au hasard

2. Quelle est la probabilité que ce soit la facture d'un client occasionnel ?
3. Quelle est la probabilité que ce soit la facture d'un locataire ayant choisi un gouter ?

4. Si on sait que la facture choisie est celle d'un locataire, quelle est la probabilité que ce soit celle d'un gouter ?
5. Si on sait que la facture choisie est celle d'un gouter, quelle est la probabilité que ce soit celle d'un locataire ?

Statistique & Probabilités - IFORD Voie B session 2018

Exercice 1

Dans une menuiserie, on propose aux apprentis 4 activités pour la journée. Les 15 apprentis (ne se connaissent pas) se répartissent aléatoirement entre les différentes activités.

1. Combien y a-t-il de répartitions possibles de ces apprentis ?
2. Combien y a-t-il de configurations pour lesquelles ?
   (a) Exactement trois activités ne soient pas désirées ?
   (b) Deux activités exactement ne soient pas désirées ?
   (c) Une activité exactement ne soit pas désirée ?
3. Déduire des questions précédentes le nombre de configurations pour lesquelles chaque activité a été désirée par au moins un apprenti.

Exercice 2

Dans un tiroir, se trouve 10 paires de gants (mains gauches et droites discernables) ; tous les gants sont de couleurs différentes. On prend simultanément 4 gants au hasard dans le tiroir.

1. Quelle est la probabilité d'avoir 2 paires de gants ?
2. Quelle est la probabilité d'avoir au moins une paire de gants ?
3. Quelle est la probabilité d'avoir exactement une paire de gants ?

Exercice 3

1. Une secrétaire effectue n appels vers n correspondants distincts (n entier, $n \geqslant 2$). On admet que ces n appels constituent n expériences indépendantes et que pour chaque appel, la probabilité d'obtenir le correspondant est p et que la probabilité de ne pas l'obtenir est $q = 1 - p$.

   On note X le nombre de clients obtenus.

   (a) Quelle est la loi de X ?
   (b) Donner l'espérance et la variance de X.

2. Après ces n recherches, la secrétaire demande une deuxième fois, et dans les mêmes conditions, chacun des $n-X$ correspondants qu'elle n'a pas obtenus la première fois. Soit Y la variable aléatoire égale au nombre de correspondants obtenus lors de la deuxième série d'appels et soit $Z = X + Y$ le nombre total de correspondants obtenus.

   (a) Quelles sont les valeurs prises par Z ?

   (b) Calculer les probabilités $p_0 = P(Z = 0)$ et $p_1 = P(Z = 1)$.

   Montrer que $p_1 = npq^{2n-2}(1 + q)$.

   (c) Calculer la probabilité conditionnelle : $P_{(X=k)}(Y = h)$ pour $k \in \{0, ..., n\}$ et $h \in \{0, n - k\}$.

   (d) Démontrer que pour tout $s \in \{0, ..., n\}$, $P(Z = s) = \sum_{k=0}^{s} P(X = k \cap Y = s - k)$.

   (e) Calculer $P(Z = s)$ pour $s \in \{0, ..., n\}$ et montrer que Z suit une loi binomiale de paramètres n et $p(1 + q)$.

   On pourra montrer que $C_n^k C_{n-k}^{s-k} = C_n^s C_s^k$.

Exercice 4

Le tableau ci-après donne les résultats de l'enquête "Internet : accès et utilisation" réalisée auprès de 1000 individus dans un pays relatif à deux variables à savoir : "le nombre moyen d'heures de connexion par mois" et "l'ancienneté en mois dans l'utilisation d'internet".

Heures connexion : X	Ancienneté : Y						Total
	< 3	$[3;6[$	$[6;12[$	$[12;24[$	$[24;36[$	$\geqslant 36$	
< 2	42	3	75	46	9	42	
$[2;5[$	1	5	9	4	5	14	
$[5;10[$	32	6	87	63	10	76	
$[10;20[$	9	8	86	42	10	40	
$\geqslant 20$	31	15	100	59	14	57	
Total							

1. (a) Compléter le tableau et donner lui un titre.

   (b) Comment appelle-t-on ce genre de tableau ?

   (c) Quelle est la nature des variables étudiées ?

2. (a) Donnez le tableau de la distribution conditionnelle (effectifs absolus et fréquences en pourcentage) de la variable Y sachant que la variable $X \geqslant 10$.

(b) Donner un titre à ce tableau.

3. Parmi les individus dont l'ancienneté est inférieure à 6 mois, quelle est la proportion de ceux dont le nombre moyen d'heures de connexion par mois est au moins égal à 10h ?
4. On suppose que le nombre moyen d'heures de connexion par mois est au maximum égal à 40h.
   (a) Donner les fréquences cumulées de la distribution marginale de la variable X et tracez le graphe de cette distribution.
   (b) Déduire les valeurs approchées (sans faire de calculs numériques) des quartiles de cette distribution.
   (c) Calculer la moyenne marginale de X et sa variance.
5. Au seuil de 5%, peut-on dire que X et Y sont indépendantes ?

Exercice 5

La limite légale du taux de présence X d'un polluant contenu dans les déchets d'usine est 6mg/kg. On effectue un dosage sur 12 prélèvements de 1kg, pour lesquels on observe les taux x_i, $i = 1, ..., 12$ tels que :

$$\sum_{i=1}^{12} x_i = 84 \text{ et } \sum_{i=1}^{12} x_i^2 = 1413$$

On admet que X suit une loi normale $\mathcal{N}(m, \sigma)$.

1. Résoudre avec un risque de première espèce fixé $\alpha = 0,025$ le problème de test suivant :

$$\begin{cases} \mathrm{H_0} : \mathrm{m} = 6 \\ \mathrm{H_1} : \mathrm{m} = \mathrm{m_1} \quad (\mathrm{m_1} > 6) \end{cases}$$

2. Proposer une région critique pour le test suivant avec $\alpha = 0,025$:

$$\begin{cases} \mathrm{H_0} : \mathrm{m} \leqslant 6 \\ \mathrm{H_1} : \mathrm{m} > 6 \end{cases}$$

   L'usine se conforme-t-elle à la législation ?
3. Peut-on calculer la puissance de ce dernier test ?

Exercice 6

L'effectif de la population d'un pays du 1er Janvier de l'année se présente comme indiqué dans le tableau ci-après :

Effectif de la population au 1er Janvier des années 1988 à 1991.

Dates	Effectifs Population
1/1/1998	100 000 000
1/1/1989	102 000 000
1/1/1990	104 346 000
1/1/1991	105 598 152

1. Calculer le multiplicateur de cette population aux dates indiquées ;
2. A l'aide d'une moyenne géométrique, déduire le multiplicateur moyen ;
3. Donner le taux d'accroissement (en %) moyen annuel de cette population ;
4. En supposant que ce taux d'accroissement reste constant, quel est le temps de doublement de cette population ?

Statistique & Probabilités - IFORD Voie A session 2018

Exercice 1

Une école de formation professionnelle de niveau Master compte 100 étudiants dont 60 en Master 1 et 40 en Master 2. On veut en choisir quatre afin de préparer la fête de fin d'année ; il faut un président, un vice-président, un trésorier et un secrétaire. Combien y a-t-il de groupes de quatre étudiants :

1. Au total ?
2. Avec deux étudiants de Master 1 et deux étudiants de Master 2 ?
3. Avec un président et un vice-président qui ne sont pas de la même classe ?
4. Avec un seul étudiant de Master 2 et qui serait président ?

Exercice 2

Parmi 12 personnes, 9 fument des cigarettes, 3 fument la pipe, 2 fument des cigarettes et la pipe.

1. Combien y a-t-il de non fumeurs ?
2. On choisit au hasard 3 personnes. Calculer les probabilités des évènements suivants :
   (a) A : "Les 3 personnes choisies fument la pipe" ;
   (b) B : "Aucune des 3 personnes choisies ne fume" ;
   (c) C : "Au moins 2 des 3 personnes choisies fument la pipe".

Exercice 3

Une urne contient six boules numérotées de 1 à 6. On en tire trois simultanément. Soit X la variable aléatoire égale au plus petit des nombres obtenus.

1. Définir la loi de X.
2. Calculer $E(X)$, $V(X)$ et $\sigma(X.)$

Exercice 4

Dans une école d'ingénierie, le temps nécessaire aux étudiants pour terminer une épreuve d'examen est une variable normale, de moyenne 90 minutes et d'écart-type 15 minutes.

1. Quelle est la proportion des étudiants qui terminent l'épreuve en moins de 2 heures ?
2. Quelle devrait être la durée de l'épreuve si l'on souhaite que 90% des étudiants puissent la terminer ?

Exercice 5

Le tableau ci-après donne les résultats de l'enquête "Internet : accès et utilisation" réalisée auprès de 1000 individus dans un pays relatif à deux variables à savoir : "le nombre moyen d'heures de connexion par mois" et "l'ancienneté en mois dans l'utilisation d'internet".

Heures connexion : X	Ancienneté : Y						Total
	< 3	$[3;6[$	$[6;12[$	$[12;24[$	$[24;36[$	$\geqslant 36$	
< 2	42	3	75	46	9	42	
$[2;5[$	1	5	9	4	5	14	
$[5;10[$	32	6	87	63	10	76	
$[10;20[$	9	8	86	42	10	40	
$\geqslant 20$	31	15	100	59	14	57	
Total							

1. (a) Compléter le tableau et donner lui un titre.
   (b) Comment appelle-t-on ce genre de tableau ?
   (c) Quelle est la nature des variables étudiées ?
2. (a) Donnez le tableau de la distribution conditionnelle (effectifs absolus et fréquences en pourcentage) de la variable Y sachant que la variable $X \geqslant 10$.
   (b) Donner un titre à ce tableau.

3. Parmi les individus dont l'ancienneté est inférieure à 6 mois, quelle est la proportion de ceux dont le nombre moyen d'heures de connexion par mois est au moins égal à 10h ?

4. On suppose que le nombre moyen d'heures de connexion par mois est au maximum égal à 40h.

   (a) Donner les fréquences cumulées de la distribution marginale de la variable X et tracez le graphe de cette distribution.

   (b) Déduire les valeurs approchées (sans faire de calculs numériques) des quartiles de cette distribution.

   (c) Calculer la moyenne marginale de X et sa variance.

Exercice 6

Le tableau ci-après donne la répartition (%) des salariés et non-salariés par sexe pour les actifs de 15 ans ou plus ayant un emploi et vivant dans le pays Ifordia.

	Hommes	Femmes
Non salariés	**13,4**	**7,3**
Salariés	**86,6**	**92,7**
Intérimaires	2,8	1,4
Apprentis	1,7	0,9
Contrats à durée déterminée	6,0	10,8
Contrats à durée indéterminée	76,1	79,6
	100	100
Total des emplois (milliers)	13 670	12 243

1. Les femmes sont-elles plus souvent salariées que les hommes ? Justifier votre réponse.

2. La répartition entre les statuts "salarié" et "non-salarié" est-elle indépendante du sexe ? Justifier votre réponse.

3. Pour l'ensemble des hommes et des femmes, quel est le pourcentage des non-salariés ?

4. Pour l'ensemble des hommes et des femmes, quel est le pourcentage des salariés ?

5. Pour l'ensemble des hommes et des femmes, quel est le pourcentage des apprentis ?

Statistique & Probabilités - IFORD Voie B session 2017

Exercice 1

Le pays Ifordia compte 770 000 habitants dont 60% sont d'origine asiatique, 39% sont noirs et 1% blancs. On tire un échantillon de 7 personnes au hasard.

1. Quelle est la probabilité d'obtenir une majorité d'asiatique ?
2. Quelle est la probabilité de n'obtenir aucun blanc ?
3. Quel est le nombre espéré d'asiatique dans l'échantillon ?

Exercice 2

Soit x une variable aléatoire continue et y sa densité de probabilité définie par :

$$y = \frac{1}{4} - \frac{x}{32}$$

1. Représenter graphiquement les variations de y dans son intervalle.
2. Donner la fonction de répartition de la variable x étudiée.
3. Vérifier par le calcul et par la représentation graphique de la question 1) que :
$$\int_0^8 \left(\frac{1}{4} - \frac{x}{32}\right) dx = 1.$$
4. Représenter graphiquement les variations de la fonction de répartition dans l'intervalle $[0; 8[$.
5. — Calculer : $\int_0^4 \left(\frac{1}{4} - \frac{x}{32}\right) dx = 1.$
   — Vérifier l'exactitude du résultat obtenu au moyen des représentations graphiques des questions 1 et 4.
6. — Calculer $\mathrm{E}(x)$.
   — Vérifier que : $\int_0^8 (x - \mathrm{E}(x)) \left(\frac{1}{4} - \frac{x}{32}\right) dx = 0.$
7. Calculer $V(x)$ et σ_x.

Exercice 3

Soient deux variables statistiques X et Y mesurées au cours des 10 années successives :

Année	1	2	3	4	5	6	7	8	9	10
X	1	3	8	11	20	24	28	32	40	48
Y	15	15	17	18	21	22	25	27	32	39

1. — Calculer le coefficient de corrélation linéaire entre les variables X et Y.
   — Calculer les coefficients de la droite des moindres carrés de Y en X.
   — Représenter le nuage des 10 points (X,Y) et tracer la droite des moindres carrés.
   — Quelle est la part de la variance de Y expliquée par ce modèle linéaire ?
2. — Calculer le coefficient de corrélation linéaire entre X et $Z = \ln(Y)$.
   — En déduire une relation approximative entre X et Y pour les 10 observations.
   — Représenter le nuage des 10 points (X,Z) et tracer la relation trouvée.
   — Quelle est la part de la variance de Y expliquée par cette relation ?
   — Comparer cette relation avec le modèle linéaire de la question 1.

Exercice 4

Le stationnement des voitures dans une certaine ville ne peut se faire règlementairement dans des parkings payants. Un automobiliste en infraction a une probabilité égale à p de se voir infliger une contravention.

Si un automobiliste se met systématiquement en stationnement interdit 30 fois dans un mois, quelle est la probabilité qu'il ait 15 contraventions au plus si $p = 0,4$? $p = 0,7$?

Préciser l'hypothèse nécessaire à la résolution de cet exercice.

Exercice 5

Une usine fabrique des imprimantes laser dont la durée de vie (exprimée en millions de pages) est une variable aléatoire suivant une loi $\mathcal{N}(2; 0,25)$.

1. Calculer la probabilité que la durée de vie d'une imprimante soit supérieure à 2,2 millions de pages.
2. On tire au hasard, avec remise, 200 imprimantes dans la production.
   — Calculer la probabilité que les deux premières imprimantes tirées aient une durée de vie supérieure à 2,2 millions de pages.
   — Donner, en justifiant, la loi de la variable aléatoire Z égale au nombre d'imprimantes tirées dont la durée de vie est supérieure à 2,2 millions de pages.
   — Donner, en justifiant, une loi approchée de la loi de Z.
   — Calculer, en utilisant la loi approchée, $p(Z > 50)$ et $p(50 \leqslant Z < 56)$.

Exercice 6

Dans le pays Ifordia, un sondage sur l'application des règles de sécurité sur la certitude des projets d'installations nucléaires a donné les résultats suivants : sur les 420 personnes qui ont répondu et qui avaient entre 18 et 30 ans, 24% ont répondu "Oui" ; parmi les 510 personnes qui avaient entre 30 et 50 ans, 34% ont répondu "Oui".

1. Calculer la probabilité critique pour H_0 : pas de différence selon l'age.

2. Au seuil de 5%, peut-on rejeter H_0 ?

Exercice 7

Les résultats d'une enquête auprès d'une population africaine ont montré que 40% des individus ne sont jamais allés en Europe et que 55% des individus n'ont jamais pris l'avion, mais que 25% sont déjà allés en Europe et ont déjà pris l'avion.

On note A, l'évènement : "Être allé en Europe" et B, l'évènement : "Avoir pris l'avion".

1. Quelle est la probabilité qu'un individu tiré au hasard dans cette population
   — Soit déjà allé en Europe ou ait déjà pris l'avion ?
   — Ne soit jamais allé en Europe et n'ait jamais pris l'avion ?

2. Quelle est la probabilité qu'un individu tiré au hasard dans cette population soit déjà allé en Europe sachant qu'il a déjà pris l'avion ?

3. Les évènements A et B sont-ils indépendants ? incompatibles ?

Statistique & Probabilités - IFORD Voie A session 2017

Exercice 1

Dans une ville du pays IFORDIA, on a procédé à un dénombrement de la population que l'on a classée selon le sexe et l'age. Le résultat de ce dénombrement figure dans le tableau ci-après :

Le groupe d'ages « 0 » est composé des enfants âgés de moins d'un an.

Le groupe d'age « 1-4 ans» concerne les enfants d'un an et plus qui n'ont pas encore fêté leurs anniversaires.

Le groupe d'ages « 75 ans et plus» concerne les personnes âgées d'au moins 75 ans. Pour ce groupe d'ages, on admettra qu'aucun individu de la population n'atteint l'age de 90 ans.

Les autres groupes d'ages notés « i-i+4» ou i est un nombre entier multiple de 5.Chaque groupe « i-i+4» concerne les personnes dont l'age est supérieur ou égal à i ans. qui n'ont pas encore l'age de i+5 ans.

Répartition de la population par sexe et par groupes d'ages

Sexe Groupe d'ages	Masculin	Féminin	Total
0 ans	350	323	673
1 - 4 ans	1182	1182	2364
5 - 9 ans	1354	1321	2675
10 - 14 ans	1377	1210	2587
15 - 19 ans	1206	1061	2267
20 - 24 ans	783	751	1534
25 - 29 ans	560	358	1098
30 - 34 ans	422	490	972
35 - 39 ans	393	550	973
40 - 44 ans	368	396	764
45 - 49 ans	313	366	679
50 - 54 ans	257	239	496
55 - 59 ans	197	172	369
60 - 64 ans	117	103	220
65 - 69 ans	40	59	99
70 - 74 ans	24	46	70
75 ans et +	21	57	78
Total	**8964**	**8864**	**17828**

1. Dans la population totale, déterminer :
   — La population des personnes âgées de moins de 20 ans ;
   — La proportion des personnes âgées entre 15 et 49 ans.
2. Construire l'histogramme de la distribution de la population totale ;
   — Calculer l'age moyen de toute la population ;
3. On appelle « rapport de masculinité», noté RM par la suite,,le rapport de l'effectif de la population masculine par l'effectif de la population féminine, soit :
   $RM = \dfrac{\text{effectif de la population masculine}}{\text{effectif de la population feminine}}.$
   — Calculer le rapport de masculinité pour l'ensemble de la population.

   (a) — Calculer pour chaque groupe d'age, le rapport de masculinité. Ces rapports seront calculés avec 2 décimales.
   — Présenter sous forme d'un tableau statistique en les regroupant par classe d'amplitude 0,10 ; la première classe étant borné inférieurement par 0,30.

Exercice 2

Soit X une variable aléatoire statistique définie sur 1000 individus. La fonction de fréquence cumulée de x est définie ainsi :

$$\begin{cases} F(x) = 0 \; si \; X < -3 \\ F(x) = 0,10 \; si - 3 \leq X < 2 \\ F(X) = 0,45 \; si \; 2 \leq X < 3 \\ F(X) = 0,85 \; si \; 5 \leq X < 10 \\ F(X) = 1 \; si \; 10 \leq X \end{cases} \tag{10.2}$$

1. Combien de valeurs différentes X prend-t-elle ?
2. Pour combien d'individus X prend-t-elle la valeur 2 ?
3. Quelle est la moyenne μ de X ?
4. Quelle est la variance de X ?

Exercice 3

Montrer que la variance d'une distribution binomiale de paramètre n et p est au plus égale à $\frac{n}{4}$

Exercice 4

On se trouve en face de huit points A, B, C, D, E, F, G, H disposés de façons à ce que jamais on ne trouve plus de deux de ces points en ligne droite.

1. Combien de segment peut-on construire avec ces huits points ?
2. Combien de triangle peut-on construire avec ces huit points ?

Exercice 5

Le tableau suivant récapitule les notes sur 20 de 10 élèves d'une école de formation professionnelle en science de la population au cours de l'année 2016 dans deux disciplines X (économie) et Y (fécondité)

Élèves	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
X	8	11	12	16	10	11	18	12	13	6
Y	1	7	10	12	3	4	14	8	9	17

1. Calculer le coefficient de corrélation linéaire entre les variables X et Y.

2. Calculer les coefficients de la droite des moindres carrés de Y en X.
3. Représenter le nuage des 10 points (X, Y) et tracer la droite des moindres carrés.
4. Quelle est la part de la variance de Y non expliquée par la relation linéaire ?
5. Reprendre les questions 1 à 4 en excluant $e_1 0$.
6. Comparer les deux modèles linéaires.

Exercice 6

Les résultats d'une enquête auprès d'une population africaines ont montré que 40% des individus ne sont jamais allés en Europe et que 55% des individus n'ont jamais pris l'avion, mais que 25% sont déjà allés en Europe et ont déjà pris l'avion. On note A, l'événement « être allé en Europe» et B, l'événement « avoir pris l'avion»

1. Quelle est la probabilité qu'un individu tiré au hasard dans cette population :
   — soit déjà allé en Europe ou ait déjà pris l'avion ?
   — ne soit jamais allé en Europe et n'ait jamais pris l'avion ?
2. Quelle est la probabilité qu'un individu tiré au hasard dans cette population soit déjà allé en Europe sachant qu'il a déjà pris l'avion ?

Exercice 7

Pour déterminer l'age moyen de ses clients, une grande entreprise de confection pour homme prélève un échantillon aléatoire de 50 clients et trouve $\bar{X} = 36$. Si on connait $\sigma = 12$, trouver un intervalle de confiance à 95% pour l'age moyen de l'ensemble des clients.

Statistique & Probabilités - IFORD Voie A session 2016

Exercice 1

Dans une ville moyenne du pays IFORDIA, il y a quatre pharmacies.

1. On demande à chaque pharmacien quel jour hebdomadaire de fermeture lui convient . De combien de façons différentes peuvent à priori se présenter les choix des quatre pharmaciens.
2. Certains pharmaciens ayant choisi le même jour de fermeture, ce qui ne peut être accepté, on réunit à nouveau les quatre pharmaciens en leur demandant de modifier éventuellement leur choix, les quatre jours de la semaine devant être différents.
3. On s'aperçoit que aucun pharmacien n'entend fermer le lundi, ni le mardi. Quels est alors le nombre de choix différents ?

Exercice 2

Le tableau statistique suivant nous vous est présenté :

Variables x_i	Effectifs n_i
2	1
3	n_2
6	n_3
9	2
14	n_5
Total	10

Sachant que la moyenne arithmétique de cette distribution est égale à 7,5 et que sa variance est égale à 14,85 ; calculer les effectifs manquants n_2, n_3, n_5.

Exercice 3

Le prénom Monique s'écrit à l'aide de 7 lettres différentes. On prépare 7 morceau de papier portant chacun l'une de ces lettres. On tire au hasard de ces papiers (un papier tiré n'est pas remis en jeu). Quelle est :

1. La probabilité d'avoir tiré les lettres i et u ?
2. La probabilité d'avoir tiré 2 voyelles exactement ?
3. La probabilité d'avoir tiré au moins une voyelle ?
4. La probabilité d'avoir tiré au moins une consonne ?

Exercice 4

Au cours d'un sondage sur les élections présidentielles dans un pays donné, 24 personnes sur 100 déclarent vouloir voter pour le candidat du parti socialiste.

1. Déterminer un intervalle de confiance au niveau 0,95 pour la proportion d'électeurs votant pour ce candidat.
2. Déterminer la taille de l'échantillon qu'il aurait fallu choisir pour obtenir un intervalle de confiance d'amplitude 2×10.

Exercice 5

Dans une classe de 25 élèves d'une école de formation professionnelle en démographie, on a relevé les notes obtenues en analyse de la fécondité X et en analyse de la mortalité Y. On obtient le tableau suivant :

$X\|Y$	4	8	10	12	14	Total
6	1	2	0	0	0	
8	3	2	0	4		9
10		2	0	1	1	
14	0	0		1	0	
16	0	0	1		1	4
Total	4		5		2	

1. Compléter ce tableau et donner lui un titre. Comment appelle-t-on ce genre de tableau ?
2. Déterminer les moyennes et les variances marginales des variables X et Y.
3. Déterminer les moyennes et les variances conditionnelles de X selon Y.
4. Vérifier la formule $V(X) = V_1\bar{(}X) + V(\bar{X}_j)$

Exercice 6

Dans un pays sahélien, la probabilité qu'une personne donnée contracte la méningite en un an est 0,4. La probabilité pour que cette personne soit atteinte d'une maladie G, autre que la méningite, pendant la même période est 0,2. On suppose que contracter la méningite et la maladie G sont deux évènements indépendants.

Quelle est la probabilité pour que cette personne contracte au moins une de ces deux maladies en un an ?

Statistique & Probabilités - IFORD Voie B session 2015

Exercice 1

Une étude sur le chiffre d'affaires d'une population de 300 PME a donné les résultats (en milliers de Francs CFA) :

Minimum	3 500
Moyenne	4 900
Écart-type	650
Étendue ou écart inter-quartile	1 100
Médiane	4 600
Premier quartile	4 100
Étendu	5 000

1. Définir la population, l'unité statistique, le caractère étudié et sa nature.
2. Donnez la valeur du troisième quartile et donnez sa signification ainsi que celles des deux autres quartile.
3. Que signifie le fait que la moyenne soit supérieure à la médiane ?
4. Représenter l'histogramme de cette distribution avec quatre classes d'effectifs égaux.

Exercice 2

Le tableau suivant donne le revenu annuel moyen des ménages, en euros, pour les dix intervalles définis par déciles, et la part de chaque intervalle dans le revenu total.

Valeurs déciles (euros)	Intervalles	Rev. Moyen	% masse totale des Rev.
$D_1 = 7304$	$< D_1$	3 845	2
$D_2 = 11091$	$[D_1, D_2[$	9 318	3
$D_3 = 14099$	$[D_2, D_3[$	12 601	5
$D_4 = 17219$	$[D_3, D_4[$	15 640	6
$D_5 = 20631$	$[D_4, D_5[$	18 863	7
$D_6 = 24653$	$[D_5, D_6[$	22 579	9
$D_7 = 29361$	$[D_6, D_7[$	26 904	11
$D_8 = 35757$	$[D_7, D_8[$	32 324	13
$D_9 = 46642$	$[D_8, D_9[$	40 548	16
	$\geqslant D_9$	69 930	28

1. Calculer le revenu annuel moyen des ménages.
2. Proposez deux indicateurs de tendance centrale, un indicateur de dispersion et un indicateur de dispersion relative. Donnez les valeurs de ces indicateurs.
3. Cette distribution de revenus est-elle symétrique ? (justifiez votre réponse).
4. Quelle est la part de l'ensemble des revenus perçus par les 4 dixièmes des ménages aux revenus les plus faibles ?
5. Soient $F_1 = 10\%$, $F_2 = 20\%$,...,$F_{10} = 100\%$ et R_i la part de l'ensemble des revenus perçus par l'ensemble des F_i ménages aux revenus les plus faibles :
   (a) Donnez les valeurs R_i correspondant à chaque fréquence cumulée F_i.
   (b) Tracez la courbe joignant, dans l'ordre, les points (F_i, R_i). Comment s'appelle cette courbe ?

(c) A quoi est égal l'indice de concentration de Gini ? (on ne demande pas sa valeur).

(d) Quelles sont les valeurs minimum et maximum de cet indice ?

(e) A quelles situations correspondent-elles ?

Exercice 3

Quel est le taux annuel moyen de croissance d'une grandeur qui double :

1. En 2 ans ?
2. En 5 ans ?
3. En 10 ans ?

Exercice 4

Soient A et B deux évènements tels que :
$P(A) = 1/4$, $P(A \cup B) = 1/3$ et $P(B) = p$.

1. Déterminer la valeur de p dans chacun des cas suivants :
   (a) Les évènements A et B sont incompatibles ;
   (b) Les évènements A et B sont indépendants ;
   (c) L'évènement A implique B ($A \subset B$).
2. Peut-on avoir B implique A ($B \subset A$) ?

Exercice 5

Le bureau des étudiants d'une école de formation professionnelle s'interroge sur l'opportunité d'un achat groupé de micro-ordinateurs en début d'année académique. Le nombre d'étudiants qui se sont déclarés intéressés est égal à 200. Mais, suite à une expérience passée, le bureau évalue la probabilité qu'un étudiant qui s'est intéressé, achète un ordinateur est égale à 0,8. D'autre part, on suppose que le comportement de chaque étudiant est indépendant de celui des autres.

Le bureau des étudiants cherche à déterminer le nombre d'ordinateurs à commander pour limiter le risque d'avoir des ordinateurs invendus. Soit X la variable aléatoire égale au nombre de micro-ordinateurs achetés par les 200 élèves intéressés.

1. Quelle est la loi exacte de X ? Par quelle loi peut-on l'approcher ? (justifiez vos réponses).
2. Calculez la probabilité qu'au plus 145 étudiants achètent un micro-ordinateur.
3. Le bureau des étudiants décide de commander 150 unités et prend en charge la garantie du matériel pendant l'année académique sous la forme suivante : chaque étudiant acheteur

verse une somme S (en euro) en plus du prix d'achat au bureau des étudiants qui assume le cout des éventuelles réparations.

On suppose que :

— les 150 ordinateurs sont vendus ;
— la probabilité qu'un ordinateur tombe en panne au cours de l'année est égale à 0,04 ;
— un ordinateur ne tombe pas plus d'une fois en panne au cours d'une année ;
— les pannes des 150 ordinateurs sont mutuellement indépendantes ;
— le cout de chaque réparation est fixé forfaitairement à 50 euros.

Soit Y, la variable aléatoire égale au nombre d'ordinateurs à réparer durant l'année universitaire.

(a) Quelle est la loi exacte de Y ? Par quelle loi peut-on l'approcher ? (justifiez vos réponses).

(b) Déterminez le plus petit nombre entier n tel que : $P(Y \leqslant n) \geqslant 0,99$.

(c) Soit Z, la variable aléatoire égale au montant (en euros) de la ligne budgétaire "Réparation des micro-ordinateurs" à la fin de l'année académique.
  — Exprimer Z en fonction de S et Y.
  — Calculer la valeur minimale de S pour que la probabilité que cette ligne budgétaire soit déficitaire, soit inférieure à 0,01.

Exercice 6

Dans cet exercice, les durées de trajets sont supposées gaussiennes.

Un directeur de société a son domicile dans la localité A. Il quitte son domicile à 8h45 et se rend en voiture à son bureau qui ouvre à 9h. La durée du trajets est, en moyenne, de 13min, avec un écart-type de 3 min.

1. Quelle est la probabilité pour le Directeur d'arriver en retard ?

2. La secrétaire du directeur a son domicile en A, mais elle se rend au bureau en empruntant en A le train de 8h32. Elle descend à la station B. Elle se rend de B à son bureau par l'autobus qui part de B à 8h50 (sans attendre le train), et qui s'arrête devant le bureau. La durée du trajet en train a pour moyenne 16 min, et pour écart-type 2 min, et la durée du trajet en autobus a pour moyenne 9 min et pour écart-type 1 min. Les durées des deux trajets sont supposées indépendantes.

   (a) Quelle est la probabilité pour la secrétaire d'arriver à l'heure ?

   (b) Quelle est la probabilité pour que le directeur ou la secrétaire arrivent à l'heure, les durées des trajets du directeur et de la secrétaire étant supposées indépendantes ?

Statistique & Probabilités - IFORD Voie A session 2015

Exercice 1

Exercice 1

Après avoir fait remplir un long questionnaire portant sur l'audience de la presse magazine à 200 individus, un institut de sondage a établi la distribution suivante pour la durée d'interview (en minutes) concernant ces 200 individus :

Durée (min)	<25	[25, 30[	[30, 35[	[35, 40[	[40, 45[	[45, 50[	$\geqslant 50$
Effectif	18	32	36	40	30	24	20

1. Définir la population, l'unité statistique, le caractère étudié et sa nature.
2. Calculez les trois quartiles de la distribution de la durée d'interview.
3. On considère que les durées minimum et maximum sont respectivement égales à 15 min et 60 min.
   (a) Tracez la courbe des fréquences cumulées et l'histogramme.
   (b) Calculez la moyenne et l'écart-type de la durée d'interview.

Exercice 2

Le tableau suivant donne les notes de quatre matières d'un étudiant de Master niveau 1 en Démographie :

Informatique	Mortalité	Sociologie	Fécondité
16	11	14	18
8		9	
11		7	
17			

La deuxième note de fécondité est illisible, mais cet étudiant se souvient que la moyenne arithmétique de l'ensemble de ses notes est égale à la moyenne arithmétique (non pondérée) des moyennes des notes par matière.

Quelle est cette note de fécondité ?

Exercice 3

Sur une période de 5 ans, une population a eu un taux de croissance annuel moyen de 3%. On a égaré la valeur du taux de croissance de la $1^{\text{ère}}$ année, mais on sait que les taux de croissance des $2^{\text{ème}}$ et $3^{\text{ème}}$ année étaient de 2%, que celui de la $5^{\text{ème}}$ année était de 1%, et que celui de la $4^{\text{ème}}$ année était de 3%.

Quel était le taux de croissance de la $1^{\text{ère}}$ année ?

Exercice 4

Une enquête exhaustive dans le pays Ifordia montre que sur les 32 564 habitants de ce pays, 23 522 lisent la revue "*Notre Ifordia*", 18 859 lisent la revue "*La vie de la population*" et 11 422 lisent les deux revues.

1. On interroge au hasard un habitant du pays Ifordia. Calculez la probabilité :
   — que cet habitant ne lise ni "*Notre Ifordia*", ni "*La vie de la population*" ;
   — que cet habitant lise "*Notre Ifordia*", mais ne lise pas "*La vie de la population*".
2. On interroge au hasard deux habitants du pays Ifordia, et on admet que leurs réponses sont indépendantes. Calculez la probabilité :
   — que les deux habitants ne lisent aucune des deux revues ;
   — qu'un habitant lise les deux revues et que l'autre n'en lise aucune.

Exercice 5

Soit X, une variable aléatoire telle que $E(X) = 50$ et $\sigma_X = 12$ et Y, une variable aléatoire telle que : $E(Y) = 30$ et $\sigma_X = 5$. On suppose que $Cov(X, Y) = 12,5$.

1. Calculer les espérances et les écarts-type des variables : $Z = X - Y$ et $T = 3X + 2Y$.
2. Sans effectuer de calcul numérique, précisez, en justifiant votre réponse, laquelle des deux variables aléatoires $(X + Y)$ et $(X - Y)$ à l'écart-type le plus petit.

Exercice 6

Un ascenseur porte la mention "charge maximale : 350 kg", cinq personnes de sexe masculin montent ensemble dans cet ascenseur. On admettra que le poids d'un adulte de sexe masculin suit une loi normale de moyenne 60 kg et d'écart-type 10 kg.

1. Soit Y, la variable aléatoire égale au poids total des cinq personnes. Donnez, en justifiant, la loi de Y.
2. Quelle est la probabilité que l'ascenseur refuse de démarrer ?

Statistique & Probabilités - IFORD Voie B session 2014

Exercice 1

Un examen est ouvert à des étudiants de formations initiales différentes : économie, statistique, mathématiques. Le tableau ci-après présente les résultats obtenus par 286 candidats à cet examen.

Résultat de l'examen	**Formations initiales**			
	Économie	Statistique	Mathématiques	Total
Réussite	14	59	54	154
Échec	21	36	75	132
Total	62	95	129	286

1. Parmi les 154 candidats qui réussi à l'examen
   (a) Quel est le pourcentage des économistes ?
   (b) Quel est le pourcentage des statisticiens ?
   (c) Quel est le pourcentage des mathématiciens ?
2. Parmi les 286 candidats qui se sont présentés à l'examen
   (a) Quel est le pourcentage des économistes qui ont réussi àl'examen ?
   (b) Quel est le pourcentage des statisticiens qui ont réussi à l'examen ?
   (c) Quel est le pourcentage des mathématiciens qui ont réussi à l'examen ?
3. Si on désigne par taux de réussite, le pourcentage des étudiants qui ont réussi à l'examen. Déterminer le taux de réussite :
   (a) Quelle que soit la formation initiale du candidat.
   (b) Chez les économistes.
   (c) Chez les statisticiens.
   (d) Chez les mathématiciens.
   (e) Représenter graphiquement les taux de réussite selon la formation initiale.
4. Le responsable de l'examen désire savoir si la formation initiale d'un étudiant influe sur sa réussite. Quelle est sa conclusion au seuil de 5%.

Exercice 2

On dispose, pour un secteur industriel donné et sur une période de 11 années, de la série du nombre de salariés Y et du chiffre d'affaires X du secteur.

Années	Nombre de salariés (en milliers) Y	Chiffre d'affaires (en millions) X
1957	294	624
1958	271	661
1959	314	728
1960	356	782
1961	383	819
1962	369	819
1963	402	938
1964	422	1023
1965	475	1136
1966	475	1227
1967	486	1327

On veut tester la réalité d'une relation linéaire entre X et Y.

Soit $y_n = a + bx_n + \varepsilon_n$ ($n = 1$ à 11). Les hypothèse classiques du modèle linéaire simple sont supposées réalisées, c'est-à-dire que ε_n est une variable aléatoire suivant une loi normale centrée et d'écart-type σ : $\mathcal{N}(0, \sigma)$.

1. Donner les estimations $\hat{a}$ et $\hat{b}$ des paramètres a et b obtenus par la méthode des moindres carrés ordinaires.
2. Calculer :
   (a) L'estimateur de la variance σ^2.
   (b) L'estimation de la variance de l'estimateur $\hat{b}$.
   (c) Le coefficient de détermination R^2 de la régression.
3. On se pose le problème de test :
$$\begin{cases} \mathrm{H}_0 \ : b = 0 \\ \mathrm{H}_1 \ : b \neq 0 \end{cases}$$
   (a) A quelle question ce test répond-t-il ?
   (b) Peut-on dire que b est significativement différent de zéro au risque $\alpha = 0,05$?
4. On suppose que, connu de façon exogène, le chiffre d'affaires en 1970 est de 1772 millions.

(a) Sur la base de la relation estimée jusque là à partir des données fournies, à combien de salariés peut-on s'attendre en 1970 ?

(b) Sur la seule base de la relation linéaire estimée jusque là à partir des données fournies, à combien de salariés peut-on s'attendre en 1970 ?

(c) Trouver un intervalle de confiance de niveau 0,95 pour l'effectif salarié du secteur.

(d) A posteriori, on démontre en 1970, 514 000 salariés. Que peut-on déduire de cette observation supplémentaire ?

Exercice 3

Une compagnie d'assurance A assure 1000 clients pour un sinistre de type S. Chaque client indépendamment des autres a, au moins au cours d'une année, une probabilité de 0,001 d'être sinistré. Le cout unitaire du remboursement du sinistre s'élève à 10 millions de francs. La compagnie d'assurance règle le 31 décembre, sur ses réserves, le montant du remboursement des sinistres de l'année et souhaite honorer ses engagements avec une probabilité de 0,999.

1. (a) Déterminer la loi exacte de nombre X de sinistres constaté en une année.
   (b) Donner l'espérance et la variance de X.
2. Par quelle loi, cette loi exacte peut-elle être approximée ?
3. Après avoir déterminé, à partir de cette loi, la valeur de x_0 telle que : $P(X \geqslant x_0) > 0,001$ et $P(X \geqslant x_0 + 1) < 0,001$, donner le montant des réserves que doit posséder la compagnie d'assurance pour honorer ses engagements avec une probabilité de 0,999.
4. Une deuxième compagnie, la compagnie d'assurance B, ayant les mêmes caractéristiques propose à la compagnie A de fusionner avec elle.
   (a) Déterminer la loi exacte du nombre Z des sinistres constatés en une année parmi l'ensemble des 2000 clients, ainsi que la loi approchée de Z.
   (b) — Déterminer le montant des réserves que doivent posséder ensemble, les deux compagnies si elles fusionnent et continuent d'honorer leurs engagements avec une probabilité de 0,999.
   — Comparer ce montant à celui des réserves dont devraient disposer les deux compagnies non fusionnées.

Exercice 4

Une urne contient n jetons numérotés de 1 à n. On effectue p tirages avec remise. Soit X la variable aléatoire réelle égale au plus grand numéro tiré. Déterminer la loi de X.

Exercice 5

Soit (X_n), n étant un entier naturel non nul, une suite de variables aléatoires indépendantes telles que $\mathrm{E}(X_n) = m$ et $\mathrm{V}(X_n) = \sigma^2$. Montrer que si m est inconnu alors :

$$T_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum X_j\right)^2 \text{ est un estimateur sans biais de } \sigma^2.$$

Exercice 6

Les assistants sociaux travaillant pour une clinique psychiatrique sont si occupés qu'en moyenne seuls 60% des patients prospectifs téléphonant pour la première fois obtiendront une communication avec l'un des assistants. On demande aux autres de laisser leur numéro de téléphone.

Trois fois quatre, un assistant trouve le temps de rappeler encore le jour même, autrement le rappel a lieu le lendemain.

L'Expérience a montré que dans cette clinique, la probabilité que le patient prospectif demande une consultation est 0,8 s'il a pu parler immédiatement à m assistants, tandis qu'elle tombe à 0,6 et 0,4 respectivement s'il y a eu rappel du patient le jour même ou le lendemain.

1. Quel est le pourcentage parmi les gens qui appellent demanderont-ils une consultation ?
2. Quel pourcentage des gens en consultation n'ont pas eu à attendre qu'on les rappelle ?

Statistique & Probabilités - IFORD Voie A session 2014

Exercice 1

Une entreprise industrielle vend des machines-outils. On s'intéresse au nombre de machines vendues en une journée. Pour cela, on définit la variable statistique X associée au caractère "nombre de machines vendues dans la journée". On observe les ventes pour 60 jours et on dresse le tableau des ventes ci-après :

Nombre de machines vendues dans la journée	Nombre de jours de ventes
0	98
1	232
2	119
3	85
4	50
5	16

1. Calculer les fréquences des ventes et les fréquences cumulées croissantes.
2. Donner la définition et calculer les caractéristiques de tendance centrale suivantes : le mode, la médiane, la moyenne.
3. Calculer les caractéristiques de dispersion suivantes : la variance, l'écart-type, le coefficient de variation.

Exercice 2

Dans une maternité, on a observé un échantillon de 20 naissances selon le poids en kilogramme et la parité :

N^o	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Poids	2,9	2,8	3,5	3,9	4,0	4,3	2,2	2,4	3,6	3,0	2,9	2,8	3,5	3,9	4,0	4,3	2,2	2,4
N^o	M	P	M	M	M	M	P	M	M	P	M	P	M	M	M	M	P	M

Notes : P = primipare ; M=multipare.

1. Un chercheur qui s'intéresse au poids seulement classe les naissances en deux catégories : Poids inférieur à 3 kg et poids supérieur ou égal à 3 kg.
   (a) Présenter le tableau d'analyse.
   (b) Déterminer le pourcentage des enfants de chaque catégorie de poids.
2. Si un chercheur s'intéresse à la parité seulement (primipare ou multipare),
   (a) Présenter le tableau d'analyse.
   (b) Déterminer le pourcentage des enfants de chaque parité.
3. Si un chercheur s'intéresse au lien entre le poids (classement en deux catégorie de la question 1.) et la parité,
   (a) Présenter le tableau d'analyse.
   (b) Déterminer le pourcentage :
      — des enfants multipares qui ont un poids inférieur à 3 kg.
      — des enfants primipares qui ont un poids inférieur à 3 kg.
      — des enfants multipares qui ont un poids supérieur ou égal à 3 kg.
      — des enfants primipares qui ont un poids supérieur ou égal à 3 kg.
4. Au seuil de 5%, peut-on dire que le poids dépend de la parité ? Commenter votre réponse.

Exercice 3

Votre jeune frère de 4 ans, qui ne sait pas lire, veut vous aider à aménager votre chambre d'étudiant. Il se propose pour placer vos livres dans la bibliothèque. Vous en avez 10 différents dont 3 de mathématiques, 4 de statistique et 3 de démographie.

1. Quelle est la probabilité que les 4 livres de statistique soient au début de la tablette ?
2. Quelle est la probabilité que les livres soient tous regroupés par matière ?

   Après avoir placé les 10 livres sur la tablette, votre jeune frère s'ennuie et décide de choisir 3 livres au hasard pour faire semblant d'étudier.

3. Quelle est la probabilité qu'il ait choisit des livres tous de matières différentes ?
4. Quelle est la probabilité qu'il ait choisit au moins 2 livres de la même matière ?

Exercice 4

Un jeu télévisé consiste en 10 questions successives auxquelles il peut être répondu par "oui" ou par "non" (on répond obligatoirement). Une réponse est donc faite d'une succession de "oui" et de "non".

1. De combien de façons différentes la suite de 10 réponses peut-elle se présenter ?
2. De combien de façons différentes la réponse "oui" peut-elle figurer exactement 7 fois dans la suite des dix réponses ?
3. De combien de façons différentes la réponse "non" peut-elle figurer exactement 3 fois dans la suite de dix réponses ?

Exercice 5

Soit Z une variable statistique définie à partir d'une variable statistique X par : $z_i = \dfrac{x_i - \bar{X}}{\sigma_x}$.

1. Calculer la moyenne de Z.
2. Calculer la variance de Z.

Exercice 6

X suit une loi normale de paramètres m et $\sigma = 2$. On considère un échantillon de taille $n = 30$. On obtient $\sum_{i=1}^{30} x_i = 360$. Donner un intervalle de confiance pour un niveau de confiance 0,95.