NOTIONS DE STATISTIQUE

Édition 2012–2013

par
Anthony Phan
(révision : 2 octobre 2012)

Département de Mathématiques
Boulevard Marie et Pierre Curie,
Téléport 2,
BP 30179, F-86962 Chasseneuil–Futuroscope cedex.

Le propos de cette note de cours n'est pas de faire une présentation de notions statistiques telle qu'elle pourrait être faite dans une formation tournée vers « les sciences expérimentales » ni même dans une formation de mathématiques « abstraites ». Il n'est question ici que d'introduire quelques thèmes statistiques à un public familiarisé au Calcul Mathématique des Probabilités et de suivre le programme du Capes et de l'Agrégation.

Dans ces notes, nous privilégions l'expression « loi de probabilité » à celle de « mesure de probabilité ». La raison en est que le substantif « mesure » est ici employé en un sens similaire à celui de « mesure physique », c'est-à-dire l'évaluation numérique d'une certaine grandeur concrète.

12 avril 2010. Une annexe sur l'estimation d'une proportion a été ajoutée pour couvrir toutes les situations concrètes et ne pas se limiter à l'application usuelle du théorème central limite (limitation étroite voulue par les programmes forcément limités des concours).

Chapitre premier

GÉNÉRALITÉS

1. **Cadre**

On considère un espace probabilisé $(E, \mathcal{E}, \mu)$ où la loi de probabilité μ n'est pas clairement identifiée mais on sait qu'elle appartient à un certain ensemble de lois de probabilités $\{\mu_\theta : \theta \in \Theta\}$. L'ensemble d'indices Θ — qui est un sous-ensemble de $\mathbb{R}^n$ en général — est appelé *ensemble des états de la nature* et $(E, \mathcal{E}, (\mu_\theta)_{\theta\in\Theta})$ est appelé *modèle statistique*. Le problème est d'évaluer pour quels $\theta \in \Theta$ il est légitime ou acceptable de penser que $\mu = \mu_\theta$.

En général, μ est la loi d'une variable aléatoire X à valeurs dans $(E, \mathcal{E})$, qu'on nomme, dans ce contexte, *variable statistique* ou *caractère*. L'espace probabilisé sur lequel elle est définie est généralement appelé *population d'étude* mais ne jouera qu'un rôle assez anecdotique dans cette note puisque nous ne nous intéresserons essentiellement qu'à la loi μ — qui, toujours dans ce contexte, peut être qualifiée de *loi statistique*.

2. **Définitions**

Définition 1. — Soit X une variable statistique à valeurs dans $(E, \mathcal{E})$. La variable statistique X est dite :

(i) *quantitative discrète* si elle est à valeurs numériques et si l'ensemble E — qui est alors un sous-ensemble de $\mathbb{R}$ — est fini ou dénombrable ;

(ii) *quantitative continue* si elle est à valeurs numériques et si l'ensemble E — qui est alors un sous-ensemble de $\mathbb{R}$ — n'est ni fini ni dénombrable ;

(iii) *qualitative ordinale* si elle n'est pas à valeurs numériques et si l'ensemble E est muni d'un ordre ;

(iv) *qualitative nominale* si elle n'est pas à valeurs numériques et si l'ensemble E n'est pas muni d'un ordre.

En général, une variable qualitative ne peut prendre qu'un nombre fini, voire dénombrable, de valeurs. Aussi emploierons-nous l'expression « variables discrètes » pour recouvrir les cas des variables qualitatives, ordinales et nominales, et des variables quantitatives discrètes. Les valeurs possibles d'une variable discrète sont souvent appelées *modalités*.

Remarques. — *a*) On utilise aussi les expressions *variable quantitative discrète*, *variable quantitative continue*, *variable qualitative ordinale*, *variable qualitative nominale*, *variable de type quantitatif discret*, *variable de type quantitatif discret*, *variable de type qualitatif ordinal*, *variable de type qualitatif nominal*, en omettant l'adjectif « statistique ».

b) Il arrive qu'une variable qualitative prenne des valeurs codées numériquement et qu'une certaine ambigüité soit possible, c'est-à-dire qu'on puisse croire qu'il s'agit d'une variable quantitative. Dans ce cas, il faut se poser la question de savoir si l'addition ou la multiplication par un scalaire de telles valeurs a un sens. Si la réponse est positive, il s'agit certainement d'une variable quantitative, sinon, d'une variable qualitative.

c) Une variable statistique ne peut être que d'au plus un seul type. Cependant, il se peut que le type d'une variable statistique ne soit pas défini par ce qui précède — notamment dans le cas d'une variable vectorielle ou à valeurs dans un espace abstrait. Néanmoins, cette classification recouvre quasiment tous les cas pratiques (*voir* exemple ci-dessous).

Exemple. — Considérons P une certaine population d'êtres humains — en nombre fini —, muni de la tribu discrète $\mathcal{P} = \mathcal{P}(P)$ et de Pr la loi de probabilité uniforme sur $(P, \mathcal{P})$. On peut vouloir étudier les variables suivantes :

(i) le pouls (nombre de pulsations cardiaques par minutes), le nombre de dents, de cheveux, d'un individu qui sont des variables quantitatives discrètes ;

(ii) la taille, la masse et l'âge (temps écoulé depuis la naissance) d'un individu qui sont des variables quantitatives continues (même si leurs mesures sont souvent arrondies à des nombres entiers ou décimaux avec précision limitée) ;

(iii) le niveau d'études, le rang d'arrivée à une course cycliste, etc., qui sont des variables qualitatives ordinales ;

(iv) le nom de famille, la nationalité, la couleur des yeux, la catégorie socio-professionnelle, l'intention de vote, etc., qui sont des variables qualitatives nominales.

Ce qui peut nous intéresser est d'évaluer la loi, ou certaines de ses caractéristiques, d'une de ces variables sans pour autant avoir à faire cette évaluation sur la population toute entière mais seulement en en prélevant un échantillon. Nous allons donner deux présentations de la notion d'échantillon. La première se rapproche de la pratique concrète sans couvrir tous les raffinements distingués par la théorie des sondages. La seconde convient à une présentation de la Statistique mathématique dans un cadre élémentaire. C'est cette dernière qui correspondra au contexte de ce cours, la première pouvant servir d'angle d'approche critique sur des cas concrets évoqués, par exemple, en exercice.

Définition 2. — Soit P une population (finie).

(i) Un *échantillon* de taille $n \in \mathbb{N}^*$ de la population est une liste éventuellement ordonnée de n individus dans la population.

(ii) L'*échantillonnage* est la méthode utilisée pour former l'échantillon.

(iii) Un échantillonnage est *non exhaustif* lorsque les choix successifs d'individus sont indépendants (tirage avec remise), dans ce cas la liste d'individus est nécessairement ordonnée.

(iv) Un échantillonnage est *exhaustif* lorsque les choix successifs d'individus interdisent de sélectionner un individu préalablement choisi, dans ce cas la liste d'individus peut être ordonnée, mais aussi non ordonnée si on considère l'échantillon simplement comme une partie de la population (tirage sans remise).

(v) Un échantillonnage est *biaisé* lorsqu'il tend à sur-représenter une partie de la population, sinon il est *représentatif*.

(vi) Un échantillonnage est *aléatoire* s'il est opéré au hasard et si tous les individus ont même chance d'être choisis (il est donc représentatif). Il est *aléatoire simple* s'il est de plus non exhaustif (tirage avec remise classique en Calcul Élémentaire des Probabilités).

(vii) Si X est une variable statistique définie sur la population, la liste des mesures de X sur un échantillon est appelée *série d'observations*, voire série statistique.

Les définitions liées au formalisme mathématique élémentaire de la Statistique mettent à l'écart la liste d'individus, ou l'échantillonnage, pour se concentrer sur la variable statistique.

Définition 3. — Soit μ une loi de probabilité sur un espace mesurable $(E, \mathcal{E})$.

(i) Un *échantillon* de taille $n \in \mathbb{N}^*$ de la loi μ, ou d'une variable X de loi μ, est la donnée d'un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et de n variables aléatoires $X_1, \dots, X_n : (\Omega, \mathcal{A}, \mathbb{P}) \to (E, \mathcal{E})$ telle que chacune des variables $(X_i)_{i=1}^n$ est de loi μ et telle que ces variables soient indépendantes.

(ii) Un *échantillon observé*, ou une *observation d'un échantillon*, ou encore une *série d'observations*, est un n-uplet $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ où $\omega \in \Omega$ est une épreuve ; c'est donc une réalisation de l'échantillon $(X_1, \dots, X_n)$.

(iii) La *loi empirique*, ou *distribution empirique*, d'un échantillon $(X_1, \dots, X_n)$ est la loi de probabilité aléatoire définie sur $(E, \mathcal{E})$ par

$$P_{X,n}(\omega) : B \in \mathcal{E} \longmapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_B\big(X_i(\omega)\big) \qquad \text{pour tout } \omega \in \Omega.$$

(iv) La *loi empirique observée*, ou *distribution empirique observée*, d'un échantillon observé $(x_1, \dots, x_n)$ est la loi de probabilité définie sur $(E, \mathcal{E})$ par

$$P_{x,n} : B \in \mathcal{E} \longmapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_B(x_i)\,;$$

c'est donc une réalisation de la loi empirique $P_{X,n}$.

(v) Si $(X_1, \dots, X_n)$ est un échantillon d'une variable statistique à valeurs dans $\mathbb{R}$ ou l'un de ses sous-ensembles, on définit sa *fonction de répartition empirique* par

$$F_{X,n}(t, \omega) = \frac{1}{n} \operatorname{Card}\{X_i(\omega) : X_i(\omega) \leqslant t\}, \qquad t \in \mathbb{R},\ \omega \in \Omega.$$

(vi) Si $(x_1, \dots, x_n)$ est un échantillon observé d'une variable statistique à valeurs dans $\mathbb{R}$ ou l'un de ses sous-ensembles, on définit sa *fonction de répartition empirique observée* par

$$F_{x,n}(t) = \frac{1}{n} \operatorname{Card}\{x_i : x_i \leqslant t\}, \qquad t \in \mathbb{R}.$$

Remarques. — *a*) Dans la pratique, la constitution d'un échantillon d'une variable X (ou de sa loi μ) à partir d'une population $(P, \mathcal{P}, \Pr)$ s'effectue par un tirage avec remise de n individus dans P sur chacun desquels on mesure la valeur prise par la variable X. Ceci revient à considérer

$$(\Omega, \mathcal{A}, \mathbb{P}) = (P, \mathcal{P}, \Pr)^{\otimes n}, \qquad \text{et, si } \omega = (\omega_1, \dots, \omega_n) \in \Omega, \qquad X_i(\omega) = X(\omega_i).$$

Un n-uplet $\omega = (\omega_1, \dots, \omega_n)$ est alors appelé *échantillon de taille n de la population.*

b) Lorsque la population P est finie et la mesure de probabilité $\Pr$ est la mesure de probabilité uniforme sur P, cette notion d'échantillon correspond dans la pratique à l'échantillonnage aléatoire simple.

c) Si $(X_1, \dots, X_n)$ est un échantillon d'une variable statistique à valeurs dans $\mathbb{R}$ ou l'un de ses sous-ensembles, les données de sa loi empirique et de sa fonction de répartition empirique sont équivalentes ; il en est bien sûr de même pour les quantités observées correspondantes.

d) On a coutume d'utiliser des majuscules pour désigner des variables aléatoires ou statistiques et des minuscules pour leurs réalisations ou observations.

Définition 4. — Soit μ une loi de probabilité sur $(E, \mathcal{E})$. On appelle *paramètre* de la loi μ toute quantité, réelle en général, qui est fonction de la loi μ (par exemple, son espérance, sa variance — lorsqu'elles existent —, etc.).

Définition 5. — On appelle *statistique* d'un échantillon $(X_1, \dots, X_n)$, toute fonction (mesurable) de l'échantillon, c'est-à-dire toute variable Λ_n de la forme

$$\Lambda_n = f_n(X_1, \dots, X_n),$$

où f_n est une application (mesurable) de E^n dans, en général, $\mathbb{R}$.

Définition 6. — (i) On appelle *estimateur* d'un paramètre λ de la loi μ toute suite de statistiques $(\Lambda_n)_{n\geqslant 1}$ (sans condition particulière liant $(\Lambda_n)_{n\geqslant 1}$ et λ !).

(ii) Si $(\Lambda_n)_{n\geqslant 1}$ est un estimateur, pour tout $n \geqslant 1$, une réalisation

$$\ell_n = f_n(x_1, \dots, x_n) = f_n\big(X_1(\omega), \dots, X_n(\omega)\big) = \Lambda_n(\omega),$$

pour un certain $\omega \in \Omega$, d'une statistique Λ_n est appelée une *estimation*.

(iii) Un estimateur $(\Lambda_n)_{n\geqslant 1}$ est un *estimateur consistant* d'un paramètre λ de la loi π si et seulement si la suite $(\Lambda_n)_{n\geqslant 1}$ converge en probabilité vers λ lorsque n tend vers l'infini.

(iv) Un estimateur $(\Lambda_n)_{n\geqslant 1}$ est un *estimateur fortement consistant* d'un paramètre λ de la loi μ si et seulement si la suite $(\Lambda_n)_{n\geqslant 1}$ converge presque sûrement vers λ lorsque n tend vers l'infini.

(v) Un estimateur $(\Lambda_n)_{n\geqslant 1}$ est un *estimateur sans biais* d'un paramètre λ de la loi μ si et seulement si pour tout n, Λ_n est intégrable et $\mathbb{E}[\Lambda_n] = \lambda$.

Lorsqu'on désire évaluer un paramètre, on le fait avec un estimateur consistant et sans biais. Lorsqu'un estimateur n'est pas sans biais, on dit qu'il est avec biais ou *biaisé* ; lorsqu'il n'est pas consistant, on pourra le qualifier de « non consistant » plutôt que d'« inconsistant ».

Les exemples usuels de paramètres, de statistiques et d'estimateurs seront donnés des sections ultérieures.

3. **Regroupements**

Lorsqu'un échantillon observé est assez grand, on a coutume de regrouper les valeurs observées afin de présenter des données sous une forme plus synthétique.

Cas d'une variable discrète : regroupement par modalités. — Soient X une variable discrète (quantitative discrète, qualitative ordinale, ou encore qualitative nominale), et un échantillon observé $(x_1, \dots, x_i, \dots, x_n)$ de cette variable. On note $(e_1, \dots, e_j, \dots, e_k)$ les différentes valeurs observées, qu'on appelle *modalités observées*. Pour chaque valeur observée x_j, on note n_j le nombre de fois qu'elle a été observée dans l'échantillon. Ainsi, il y a k différentes valeurs observées, $n_j \geqslant 1$ pour tout $j = 1, \dots, k$, et $n_1 + \cdots + n_j + \cdots + n_k = n$ le nombre total d'observations.

Cas d'une variable quantitative continue : regroupement en classes. — Lorsqu'on étudie une variable statistique continue, un échantillon (observé) de taille n donne toujours lieu à une loi empirique discrète (portée par au plus n points).

Si on sait que la loi de probabilité μ est absolument continue par rapport à la mesure de Lebesgue, on remplace souvent la loi empirique par une loi déduite d'une *répartition en classes* : soient $a_0 < a_1 < \cdots < a_k$; on répartit les observations $(x_1, \dots, x_n)$ dans les intervalles

$$I_1 = [a_0, a_1[, \ \dots, \ I_j = [a_{j-1}, a_j[, \ \dots, \ I_k = [a_{k-1}, a_k],$$

qu'on appelle alors *classes*, pour obtenir la répartition $(I_j, n_j)_{j=1}^k$ où n_j est le nombre d'observations dans à la classe I_j et est alors appelé *effectif de la classe* ; le nombre $f_j = n_j/n$ est

alors appelé la *fréquence de la classe*. On note alors $c_j = (a_{j-1} + a_j)/2$ les *centres de classes*, et $f_j = n_j/n$ les fréquences observées correspondantes. On a alors la fonction de densité et la fonction de répartition associées

$$p_{x,n} : x \in \mathbb{R} \longmapsto \sum_{j=1}^{k} \frac{f_j}{a_j - a_{j-1}} \mathbb{1}_{[a_{j-1},a_j[}(x), \qquad F_{x,n}(x) = \int_{-\infty}^{x} p_{x,n}(y)\,\mathrm{d}y\,;$$

cette fonction de répartition est continue, croissante, affine par morceaux, et vérifie

$$F_{x,n}(a_j) - F_{x,n}(a_{j-1}) = \int_{a_{j-1}}^{a_j} p_{x,n}(x)\,\mathrm{d}x = \frac{f_j}{a_j - a_{j-1}} \times (a_j - a_{j-1}) = f_j.$$

Il est à noter qu'une fois réparties des observations en classes, on s'interdit de revenir aux valeurs originales. Le calcul de quantités telles que celles présentées à la section suivante ne doit alors se baser que sur $(I_j, n_j)_{j=1}^k$ (des remarques seront faites pour préciser ce cas) de sorte que cela revient à considérer des observations discrètes regroupées en modalités $(c_j, n_j)_{j=1}^k$. Le choix des extrémités des intervalles $(a_j)_{j=0}^k$ est assez arbitraire. Elles doivent évidemment recouvrir l'ensemble des observations et seront choisies de sorte à rendre compte de la loi de la variable X ou de ce qu'on en sait (en particulier de la forme de sa densité [si elle en a une]).

Cas d'une variable quantitative discrète : regroupement en classes. — Lorsqu'une variable quantitative discrète a un très grand nombre de modalités, il est fréquent qu'on répartisse de grands échantillons en classes de manière similaire à ce qui précède. On essaiera dans ce cas de choisir les extrémités de classes de sorte à refléter la ditribution observée, c'est-à-dire que les classes contiennent des effectifs comparables s'ils sont non nuls.

4. **Exercices**

Exercice 1. — On considère la population des malades atteints d'un cancer des poumons et la variable X qui indique si un individu est fumeur ou non : elle vaut 1 si l'individu considéré est fumeur, 0 sinon. On a observé l'échantillon :

$$(1,1,1,0,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,1,1,0,1,1)$$

Donner le type de la variable X, la répartition par modalités de l'échantillon observé, et la loi empirique observée.

Exercice 2. — Afin d'étudier une variété d'araignées d'origine mexicaine (*latrodectus hesperus* ou *veuve noire*), des œufs de cette espèce ont été prélevés sur la toile où ils étaient déposés sous la forme de sacs. À leur éclosion, les jeunes femelles et les jeunes mâles ont été placés dans des vivariums séparé afin d'empêcher leur éventuelle reproduction. Au bout de quelques mois, 20 individus de chaque sexe ont été extraits de leurs vivariums respectifs, leur taille a été mesurée, et les femelles, nettement plus grandes que les mâles, ont été délicatement

marquées.

Taille	1,49	1,17	1,29	0,37	1,36	0,40	0,44	1,05	0,58	1,11
Œufs	688	388	544	m	645	m	m	264	m	322
Taille	1,26	0,42	0,37	0,40	1,59	1,29	1,17	0,38	1,18	0,88
Œufs	503	m	m	m	847	540	382	m	404	203
Taille	1,20	0,29	0,47	1,10	0,51	0,43	1,44	0,50	1,18	1,30
Œufs	422	m	m	307	m	m	784	m	402	556
Taille	0,37	0,39	0,51	0,62	1,20	1,39	1,30	0,48	0,50	0,28
Œufs	m	m	m	m	419	690	557	m	m	m

Ces quarante araignées ont été placées dans un troisième vivarium, les femelles n'ont pas tardé à pondre et on a alors compté le nombre d'œufs correspondant.

Les données présentées indiquent la taille en centimètres de chacun des individus (tête et abdomen). Lorsqu'il s'agissait d'une femelle, on a reporté en dessous le nombre d'œufs du nid correspondant, alors que pour un mâle, on a simplement inscrit la lettre m.

La population sous-jacente est l'ensemble théorique de tous les individus de cette espèce. Préciser quelles sont les trois variables statistiques envisagées ainsi que leurs types respectifs. L'une de ces trois variables n'est pas définie sur l'ensemble de la population, ce qui oblige pour l'étudier à considérer une *sous-population*. Par quelle variable est définie cette sous-population ?

Exercice 3 (accouchements à Baltimore). — Afin de déterminer la proportion de césariennes parmi les accouchements de la ville de Baltimore, un échantillon a été constitué à partir des dossiers des services d'obstétrique de deux hôpitaux universitaires. On a calculé un taux d'accouchements par césarienne de 20 %. Par la suite, un recueil d'informations un peu plus complet a révélé que, en général, les hôpitaux de la ville présentaient des taux d'accouchements par césarienne variant entre 10 % et 12 %.

(i) Quelle est la population de cette étude ?

(ii) Pourquoi peut-on considérer qu'il y a biais d'échantillonnage ?

(iii) Définir la variable qui permet de calculer la proportion de césariennes parmi les accouchements de la ville de Baltimore.

Exercice 4 (élections américaines de 1936). — Les premiers sondages d'opinion ont été faits aux États-Unis à l'occasion des couvertures de presse des élections présidentielles américaines. Dès 1824, le *Raleigh Star* fit des enquêtes pré-électorales par consultation d'électeurs choisis parmi ses lecteurs.

En novembre 1936, les deux candidats en présence étaient Alfred M. Landon et Franklin D. Roosevelt. Le magazine *Literary Digest* interrogea approximativement 2,4 millions électeurs.

(i) Quelle est la population d'étude ?

(ii) Définir la variable qui permet de calculer la proportion de votes en faveur d'un candidat.

(iii) Pourquoi peut-on penser qu'il y aura un biais d'échantillonnage si la majorité des électeurs interrogés sont des lecteurs du journal ?

Exercice 5 (consommation d'oxygène). — Une étude est réalisée pour déterminer la vitesse à laquelle le corps d'un homme d'âge compris entre 40 et 60 ans consomme et absorbe de l'oxygène. Pour cela, on a constitué un échantillon de 31 individus répartis en 3 groupes expérimentaux.

(i) Quelle est la population d'étude ?

(ii) L'échantillon est-il représentatif ?

(iii) Les variables de l'étude sont les suivantes : l'*âge*, le *groupe expérimental*, la *consommation d'oxygène*, le *rythme cardiaque au repos*, le *rythme cardiaque pendant la course*, la *durée*, exprimée en minutes, de la course, et, la *masse*.

Indiquer le type de chaque variable (qualitative nominale, qualitative ordinale, quantitative discrète, quantitative continue).

Exercice 6 (tirages aléatoires simples). — On lance un dé à 6 faces équilibré. Le résultat est aléatoire et est une variable notée X.

(i) Quelles peuvent être les populations sous-jacentes ?

(ii) Calculer la moyenne m_X (ou espérance) de X ainsi que sa variance σ_X^2.

(iii) Le dé est lancé n fois, les lancers successifs étant supposés ne pas dépendre des lancers antérieurs. On note $X_1, \ldots, X_n$ les résultats successifs, et

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

la variable moyenne des résultats obtenus. Quelle est la moyenne $m_{\bar{X}}$ de la variable $\bar{X}$, et quelle est sa variance $\sigma_{\bar{X}}^2$?

(iv) Prenons $n = 2$. Faire la liste des résultats possibles, y associer les valeurs de $\bar{X}$ correspondantes. Représenter par un tableau, puis graphiquement la distribution de $\bar{X}$. Calculer sa moyenne ainsi que sa variance à partir de la distribution de $\bar{X}$. Vérifier qu'on retrouve dans ce cas particulier ce qui avait été affirmé en répondant à la question précédente.

Exercice 7 (tous les échantillons possibles d'une population de 5 individus). — Une population P se compose de 5 individus i numérotés de 1 à 5. On observe sur P la variable X dont les valeurs $X(i)$ sont respectivement 8, 3, 11, 4 et 7.

(i) Déterminer tous les échantillons sans remise (on ne tient pas compte de l'ordre) de taille $n = 2$ et compléter le tableau suivant :

Numéro de l'échantillon	Couple (i, j) des unités sélectionnées	Couple des valeurs de X	Valeur de $\bar{X}(k)$
k	(i, j)	$\big(X(i), X(j)\big)$	$\dfrac{X(i) + X(j)}{2}$

(ii) Quelques propriétés que peut ou non vérifier une méthode d'échantillonnage sont :

a) tous les individus ont la même chance d'être choisis ;

b) les différents choix sont indépendants.

Montrer qu'une de ces propriétés n'est pas vérifiée par l'échantillonnage de la question (i). Calculer m_X (moyenne de la variable X sur l'ensemble de la population sur laquelle cette variable est définie), σ_X^2 (sa variance), $m_{\bar{X}}$ (moyenne de la variable $\bar{X}$ sur l'ensemble de la population sur laquelle cette variable est définie), $\sigma_{\bar{X}}^2$ (sa variance). Quelles sont les valeurs de $m_{\bar{X}}$, $\sigma_{\bar{X}}^2$ dans le cas d'un échantillonnage aléatoire simple ? Quel est l'échantillonnage le plus « précis » ?

(iii) Déterminer les quantiles $q_{10\%}$ et $q_{90\%}$ de la variable $Y = 6{,}6 - \bar{X}$ (*voir* chapitres II et III ; on utilisera la convention consistant à prendre le milieu de l'intervalle des quantiles envisageables en cas d'indétermination). Calculer

$$\mathbb{P}\{Y \in [q_{10\%}, q_{90\%}]\} \qquad \text{et} \qquad \mathbb{P}\big\{m_X \in \big[\bar{X} + q_{10\%}, \bar{X} + q_{90\%}\big]\big\}.$$

Déterminer tous les intervalles $[\bar{X}+q_{10\%}, \bar{X}+q_{90\%}]$ et retrouver le résultat précédent. (Ceci est lié à l'estimation par intervalles qui sera développée dans un chapitre ultérieur. Ici ce que l'on calcule est l'intervalle de confiance de la moyenne m_X correspondant à chaque échantillon.)

Chapitre II

STATISTIQUES ET PARAMÈTRES USUELS

Soit μ une loi de probabilité sur $(E, \mathcal{E})$ et $(X_1, \ldots, X_n)$ un échantillon de la loi μ. Un paramètre de la loi μ est, nous le rappelons, toute quantité numérique calculable à partir cette loi.

En particulier, lorsque E est fini ou dénombrable — c'est-à-dire qu'on considère une variable discrète —, $E = (e_1, \ldots, e_j, \ldots)$, les nombres $\mu_j = \mu\{e_j\}$ sont des paramètres de la loi μ souvent appelés *probabilités théoriques*, voire *fréquences théoriques*. Ses pendants empiriques sont

$$f_{X,n,j} = \frac{1}{n} \operatorname{Card}\{X_i : X_i = e_j\} \qquad \text{et} \qquad f_{x,n,j} = \frac{1}{n} \operatorname{Card}\{x_i : x_i = e_j\},$$

pour les *fréquences empiriques* et les *fréquences empiriques observées* respectivement.

Les deux sous-sections qui suivent portent presque exclusivement sur des variables quantitatives (discrètes ou continues).

1. **Statistiques et paramètres de position usuels**

Soit μ une loi de probabilité portée par $\mathbb{R}$ ou un de ses sous-ensembles. Un paramètre $\lambda = \phi(\mu)$ est un paramètre de position si, et seulement si, il s'exprime dans les mêmes unités que la variable sous-jacente et si, pour tout $c \in \mathbb{R}$, $\phi(\tau_c \mu) = \phi(\mu) + c$, où $\tau_c \mu$ est la loi de probabilité sur $\mathbb{R}$ définie par $\tau_c \mu(B) = \mu(B - c)$ et $B - c = \{x \in \mathbb{R} : x + c \in B\}$.

Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi de probabilité μ portée par $\mathbb{R}$ ou un de ses sous-ensembles. Une statistique $\Lambda_n = f_n(X_1, \ldots, X_n)$ est une statistique de position si, et seulement si, elle s'exprime dans les mêmes unités que la variable sous-jacente et si, pour tout $c \in \mathbb{R}$, $f_n(X_1 + c, \ldots, X_n + c) = f_n(X_1, \ldots, X_n) + c$.

Concrètement, *un paramètre ou une statistique est de position si et seulement si lorsque l'origine des mesures est translatée d'un nombre c, le paramètre ou la statistique l'est d'autant, et s'il s'exprime dans les mêmes unités que la variable sous-jacente.*

Ces définitions peuvent sembler assez abstraites. On vérifiera à titre d'exercice qu'elles conviennent aux exemples présentés ci-dessous.

Espérances et moyennes. — L'*espérance*, ou *moyenne*, de la loi μ est

$$\int_{\mathbb{R}} x \, \mu(\mathrm{d}x).$$

La *moyenne empirique* d'un échantillon $(X_1, \ldots, X_n)$ est

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{(prononcer « grand } X\text{-barre »).}$$

La *moyenne empirique observée* d'un échantillon observé $(x_1, \ldots, x_n)$ est

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{(prononcer « petit } x\text{-barre »).}$$

Valeurs extrêmes. — Les *valeurs extrêmes* de la loi μ sont

$$\inf \operatorname{supp} \mu \qquad \text{et} \qquad \sup \operatorname{supp} \mu,$$

où $\operatorname{supp} \mu$ est le support de μ, c'est-à-dire le plus petit sous-ensemble fermé F de $\mathbb{R}$ tel que $\mu(F) = 1$. Les *valeurs extrêmes* d'un échantillon $(X_1, \ldots, X_n)$ sont

$$\min(X_1, \ldots, X_n) \qquad \text{et} \qquad \max(X_1, \ldots, X_n).$$

Les *valeurs extrêmes observées* d'un échantillon observé $(x_1, \ldots, x_n)$ sont

$$\min(x_1, \ldots, x_n) \qquad \text{et} \qquad \max(x_1, \ldots, x_n).$$

La notion de valeurs extrêmes s'applique aussi à des lois, des échantillons, des observations, de variables qualitatives ordinales sans que ce soit alors un paramètre de position.

Modes. — Supposons la variable, ou la loi, discrète. Si la loi possède une valeur possible de plus grande probabilité, cette valeur possible est alors appelée *mode* de la loi (de même pour les lois empirique et empiriques observées).

Supposons la loi μ à densité. Si la densité de la loi μ admet un maximum local qui est aussi un maximum global, le point où est atteint ce maximum est alors appelée *mode* de la loi.

Supposons donné un échantillon réparti en classes $(I_j, n_j)_{j=1}^k$ d'une variable quantitative continue ou discrète. Si à l'une de ces classes correspond un effectif supérieurs aux autres, cette classe est alors appelée *mode* ou *classe modale* de cette répartition.

La notion de mode s'applique aussi à des lois, des échantillons, des observations, de variables qualitatives ordinales et nominales sans que ce soit alors un paramètre de position. Notons que le cas d'une loi à densité correspondrait à une répartition en classes de même longueur égale à $\mathrm{d}x$.

Les définitions qui précèdent sont un peu trop restrictives. Quitte à être imprécis, un mode est une bosse de la distribution. Si une distribution ne possède qu'une bosse marquée, elle est dite unimodale, si elle possède plusieurs bosses marquées, elle est dite plurimodale.

Quantiles. — Soient F la fonction de répartition d'une loi μ — qui peut être la loi empirique observée d'un échantillon observé $(x_1, \ldots, x_n)$, etc. —, et $\alpha \in \left]0, 1\right[$. Le *quantile* d'ordre α de la loi μ est un réel q_α défini par

$$q_\alpha \in \big[\sup\{x \in \mathbb{R} : F(x) < \alpha\},\, \inf\{x \in \mathbb{R} : F(x) > \alpha\}\big].$$

Ainsi, si $F^{-1}(\alpha)$ est réduit à un point, celui-ci est q_α ; si $F^{-1}(\alpha)$ est un intervalle, q_α est un des points de celui-ci ; et, finalement, si $F^{-1}(\alpha)$ est vide, alors α est franchi au cours d'une discontinuité de F et q_α est le point où a lieu cette discontinuité.

De manière équivalente, le quantile d'ordre α est un réel q_α tel que

$$F(q_\alpha) \geqslant \alpha \qquad \text{et} \qquad F(q_\alpha -) \leqslant \alpha.$$

Les quantiles $q_{1/4}$, $\mathrm{me} = q_{1/2}$ et $q_{3/4}$ sont appelés respectivement *premier quartile*, *médiane* et *troisième quartile* et sont les plus fréquemment calculés. On s'intéresse parfois aux *déciles* (de la forme $q_{k/10}$) et aux *centiles* (de la forme $q_{k/100}$).

Précisons qu'il n'y a indétermination du quantile d'ordre $\alpha \in \left]0, 1\right[$ que lorsque α est en face d'un palier de la fonction de répartition F. L'intervalle $F^{-1}\{\alpha\}$ est en général de la forme $[a, b[$, et de la forme $[a, b]$ si F est continue en b, avec a et b fini. Les deux principales conventions destinées à lever cette indétermination consistent à choisir pour quantile, ou bien $q_\alpha = a$ l'extrémité gauche de l'intervalle, ou bien $q_\alpha = (a + b)/2$ — qui est une convention utilisée par certains logiciels.

Quantiles, cas des échantillons. — Soit $(x_1, \dots, x_n)$ un échantillon observé qu'une variable quantitative X. Ordonnons l'échantillon pour obtenir $(x_{(1)}, \dots, x_{(n)})$ avec $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$. Notons que la fonction de répartition F ne dépend pas de l'ordonnancement de l'échantillon, et qu'elle a pour principale particularité d'être une fonction en escalier à valeurs dans $\{0, 1/n, 2/n, \dots, 1)$.

Plusieurs définitions des quantiles existent dans ce cadre. Elles vérifient toutes la définition à l'aide de la fonction de répartition et ne font qu'adopter une certaine convention dans les cas où il y a indétermination, *i.e.*, lorsque l'ordre du quantile est égal au niveau d'un palier de la fonction de répartition. La première est la suivante :

Convention 1. — Si $\alpha \in]0, 1[$, le quantile d'ordre α de l'échantillon est égal à la valeur du terme dans l'échantillon ordonné dont l'indice est le plus petit entier supérieur ou égal à $n\alpha$.

Examinons cette définition. Si α est en regard de l'intérieur d'un saut de la fonction de répartition, comme cela est toujours le cas si $n\alpha$ n'est pas entier, le quantile correspondant est alors égal au point où a lieu le saut. Ce point est une valeur observée, c'est donc un certain $x_{(j)}$. Notons $j_{\min}$ et $j_{\max}$ les plus petit et plus grand indices de la sorte. Le niveau du précédent palier est alors $(j_{\min} - 1)/n$ et celui du suivant $j_{\max}/n$ et de ce fait on a $j_{\min} - 1 < n\alpha < j_{\max}$. La définition qui précède est donc cohérente dans ce cas avec celle qui définit les quantiles par la fonction de répartition. Supposons maintenant que α soit en regard d'un palier, auquel cas $n\alpha$ est nécessairement entier. On constate que ce palier s'étend de $x_{(n\alpha)}$ à $x_{(n\alpha+1)}$. Le choix proposé est de prendre pour quantile $x_{(n\alpha)}$, c'est-à-dire de choisir l'extrémité gauche du palier.

Convention 2. — Soit $\alpha \in]0, 1[$. Le quantile d'ordre α est :

(i) si $n\alpha$ n'est pas entier, $q_\alpha = x_{(\lfloor n\alpha \rfloor + 1)}$, où $\lfloor n\alpha \rfloor$ est la partie entière de $n\alpha$;

(ii) si $n\alpha$ est entier, $q_\alpha = (x_{(n\alpha)} + x_{(n\alpha+1)})/2$.

On constate immédiatement que cette définition coïncide avec la précédente lorsque $n\alpha$ n'est pas entier. Lorsque $n\alpha$ est entier et α en regard de l'intérieur d'un saut de la fonction de répartition, alors on a notamment $x_{(n\alpha)} = x_{(n\alpha+1)}$ et, là encore, cette définition coïncide avec les précédentes. Finalement, si $n\alpha$ est entier et α en regard d'un palier de la fonction de répartition, $x_{(n\alpha)}$ et $x_{(n\alpha+1)}$ sont respectivement les extrémités gauche et droite du palier. Le quantile déterminé est alors la demi-somme de ces deux extrémités.

C'est cette dernière convention que nous avons utilisée pour réaliser les graphiques de ces notes. Il faut néanmoins insister qu'aucune convention ne semble meilleure qu'une autre.

Remarques. — *a*) Lorsqu'on a un échantillon regroupé par modalités ou en classes, la moyenne est calculée, suivant les cas, selon

$$\bar{x}_n = \frac{1}{n} \sum_{j=1}^{k} n_j \, x_j \qquad \text{ou} \qquad \bar{x}_n = \frac{1}{n} \sum_{j=1}^{k} n_j \, c_j .$$

b) Lorsqu'on a un échantillon regroupé par modalités, on effectue la procédure ci-avant en tenant compte des multiplicités et en prenant pour différentes valeurs les modalités. Pour un échantillon réparti en classes, la détermination des quantiles est faite à partir de la fonction de répartition qui est continue et affine par morceaux.

2. **Statistiques et paramètres de dispersion usuels**

Soit μ une loi de probabilité portée par $\mathbb{R}$ ou un de ses sous-ensembles. Un paramètre $\lambda = \phi(\mu)$ est un paramètre de dispersion si, et seulement si, il s'exprime dans les mêmes unités que la variable sous-jacente et si, pour tout $c \in \mathbb{R}$, $\phi(\tau_c\mu) = \phi(\mu)$, où $\tau_c\mu$ est la loi de probabilité sur $\mathbb{R}$ définie par $\tau_c\mu(B) = \mu(B - c)$ et $B - c = \{x \in \mathbb{R} : x + c \in B\}$.

Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi de probabilité μ portée par $\mathbb{R}$ ou un de ses sous-ensembles. Une statistique $\Lambda_n = f_n(X_1, \ldots, X_n)$ est une statistique de dispersion si, et seulement si, elle s'exprime dans les mêmes unités que la variable sous-jacente et, si pour tout $c \in \mathbb{R}$, $f_n(X_1 + c, \ldots, X_n + c) = f_n(X_1, \ldots, X_n)$.

Concrètement, *un paramètre ou une statistique est de dispersion si et seulement si il demeure inchangé par toute translation de l'origine des mesures, et s'il s'exprime dans les mêmes unités que la variable sous-jacente.*

Ces définitions peuvent sembler assez abstraites. On vérifiera à titre d'exercice qu'elles conviennent aux exemples présentés ci-dessous.

Écarts-type. — La *variance* de la loi μ est

$$\int_{\mathbb{R}} x^2 \, \mu(\mathrm{d}x) - \left(\int_{\mathbb{R}} x^2 \, \mu(\mathrm{d}x) \right).$$

La *variance empirique* d'un échantillon $(X_1, \ldots, X_n)$ est

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \big(\bar{X}_n\big)^2 \qquad \text{(prononcer « grand } S\text{-deux »).}$$

La *variance empirique observée* d'un échantillon observé $(x_1, \ldots, x_n)$ est

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x}_n)^2 \qquad \text{(prononcer « petit } s\text{-deux »).}$$

L'*écart-type* de la loi μ, l'*écart-type* S d'un échantillon $(X_1, \ldots, X_n)$, l'*écart-type observé* s d'un échantillon observé $(x_1, \ldots, x_n)$, sont respectivement les racines-carré des variances correspondantes.

Écarts-type corrigés. — Les variances corrigées sont respectivement les statistiques

$$\hat{S}_n^2 = \frac{n}{n-1} \, S_n^2 \qquad \text{et} \qquad \hat{s}_n^2 = \frac{n}{n-1} \, s_n^2.$$

Les écarts-type corrigés sont les racines-carrés des variances corrigées correspondantes.

Étendues. — L'*étendue* d'une loi, d'une distribution empirique, etc., est la différence de ses valeurs extrêmes (max − min).

Étendues interquartiles. — L'*étendue interquartile* d'une loi, d'une distribution empirique, etc., est la différence $q_{3/4} - q_{1/4}$.

Remarques. — *a*) Lorsqu'on a un échantillon regroupé par modalités ou en classes, la variance est calculée selon

$$s_n^2 = \frac{1}{n} \sum_{j=1}^{k} n_j \, x_j^2 - (\bar{x}_n)^2 \qquad \text{ou} \qquad s_n^2 = \frac{1}{n} \sum_{j=1}^{k} n_j \, c_j^2 - (\bar{x}_n)^2,$$

les autres statistiques s'en déduisent.

b) Les variances ne sont pas des paramètres de dispersion. En effet, elles s'expriment selon le carré des unités de la variable considérée.

c) Dans la cas d'un échantillon réparti en classes, la variance n'est pas égale à l'expression correspondante calculée avec la fonction de densité.

3. **Exercices**

Exercice 1. — Vérifier que dans le cas d'un échantillon réparti en classes, on a

$$\bar{x}_n = \int_{-\infty}^{+\infty} x\, p_{x,n}(x)\, \mathrm{d}x.$$

Exercice 2. — Vérifier qu'avec la seconde convention sur le calcul des quantiles, la médiane d'un échantillon de taille $n = 2k+1$ est le terme de rang $k+1$ dans l'échantillon ordonné, et que, si l'échantillon est de taille $n = 2k$, c'est la demi-somme des valeurs des termes de rang k et $k+1$ de l'échantillon ordonné.

Exercice 3. — Les données étant celles de l'exercice portant sur des malades atteints d'un cancer des poumons, calculer la moyenne et la variance observée à partir de l'échantillon ainsi que de l'échantillon réparti en modalités, et donner le mode de la distribution observée et ses trois quartiles. En se rappelant le type de la variable étudiée, donner une légitimité à ces calculs.

Exercice 4. — Vérifier que si X est de carré intégrable et $(X_1, \ldots, X_n)$ un échantillon de la variable X, alors

$$\mathbb{E}\big[\bar{X}_n\big] = \mathbb{E}[X] \qquad \text{et} \qquad \mathrm{Var}\big(\bar{X}_n\big) = \frac{1}{n}\,\mathrm{Var}(X),$$

et aussi que

$$\mathbb{E}\big[S_n^2\big] = \frac{n-1}{n}\,\mathrm{Var}(X) \qquad \text{et qu'ainsi} \qquad \mathbb{E}\big[\widehat{S}_n^2\big] = \mathrm{Var}(X).$$

Exercice 5. — Les données étant celles portant sur la variété d'araignées mexicaines, calculer les moyennes et les variances observées de l'ensemble de l'échantillon de la taille, du sous-échantillon de la taille des femelles, et de celui de la taille des mâles.

Chapitre III

REPRÉSENTATIONS GRAPHIQUES

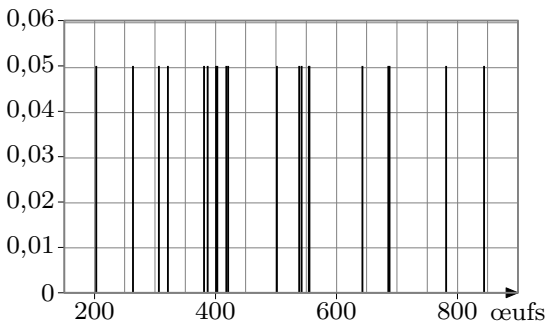
1. **Variables discrètes**

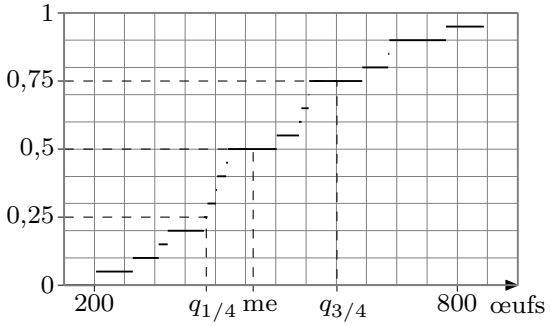
Lorsqu'on veut représenter graphiquement la distribution d'un échantillon d'une variable statistique discrète (qualitative nominale, qualitative ordinale ou quantitative discrète), on procède (si ce n'est déjà fait) à un regroupement en modalités $(x_j, n_j)_{j=1}^k$ et on représente la distribution obtenue à l'aide d'un *diagramme en bâtons*, ou *histogramme en bâtons*.

On représente sur l'axe des abscisses les différentes modalités en respectant leur ordre (si elles en ont un) et une échelle (s'il y a lieu). Au dessus de chacune de ces modalités, on trace un trait (bâton) de hauteur proportionnelle à l'effectif ou la fréquence correspondante.

Dans le cas de variables qualitatives ordinales ou quantitatives discrètes, la fonction de répartition est définie. C'est une fonction croissante en escalier dont les sauts ont lieu aux modalités x_j et y valent $f_j = n_j/n$.

Remarque. — Lorsque le nombre de modalités d'un échantillon d'une variable quantitative discrète est grand, représenter quelques observations éparses n'est guère suggestif. Ainsi pour l'échantillon du nombre d'œufs pondus par les 20 araignées femelles, on a :





À gauche on trouve le diagramme en bâtons des fréquences. Dans cet échantillon toutes les observations sont distinctes ainsi chaque modalité observée ne l'est qu'une seule fois et a une fréquence égale à $1/20 = 0{,}05$. Il serait dans ce cas préférable de regrouper les observations par classe d'amplitude 50 ou 100 et tracer un histogramme (*voir* plus loin). En revanche, la courbe représentative de la fonction de répartition empirique observée (à droite) est parfaitement évocatrice. En particulier, elle permet de retrouver les quartiles (ainsi qu'indiqué sur la figure). Dans cet exemple, les quartiles sont tous indéterminés ; ceux qui ont été choisis sont les quartiles de l'échantillon déterminés selon la seconde convention.

2. **Variables quantitatives**

Lorsqu'on veut représenter graphiquement la distribution d'un échantillon d'une variable statistique quantitative continue, on procède (si ce n'est déjà fait) à un regroupement en classes $(I_j, n_j)_{j=1}^k$ et on représente la distribution obtenue à l'aide d'un *histogramme*, ou *histogramme en bandes*. C'est aussi ce que l'on fera dans le cas d'un échantillon d'une variable statistique quantitative discrète possédant un grand nombre de modalités.

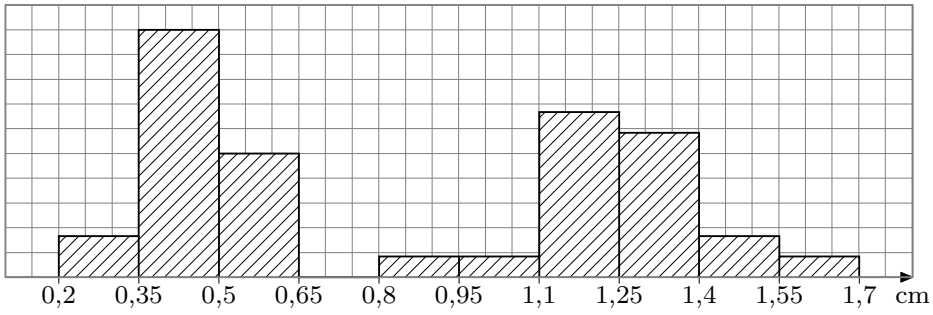
Dessiner un histogramme consiste à élever au dessus de chaque classe une bande dont la surface est proportionnelle à la fréquence l'effectif de la classe. Cela consiste donc à représenter la fonction de densité de la répartition en classes.

Représenter fonction de répartition de la distribution répartie en classes consiste à tracer la courbe représentative de sa fonction de répartition qui est la primitive de la fonction de densité nulle à gauche de la réunion des classes, et égale à 1 au-delà.

Exemple. — Nous poursuivons l'étude la taille des araignées. Constatant que la plus petite taille observée est 0,28 cm et la plus grande 1,59 cm, nous choisissons de répartir l'échantillon complet en 10 classes de même amplitude 0,15 cm :

*Classes (*cm*)*	[0,20 ; 0,35[	[0,35 ; 0,50[	[0,50 ; 0,65[	[0,65 ; 0,80[	[0,80 ; 0,95[
Effectifs	2	12	6	0	1
Eff. cumulés	2	14	20	20	21
Fréquences	0,05	0,3	0,15	0	0,025
Fré. cumulées	0,05	0,35	0,5	0,5	0,525
*Classes (*cm*)*	[0,95 ; 1,10[	[1,10 ; 1,25[	[1,25 ; 1,40[	[1,40 ; 1,55[	[1,55 ; 1,7]
Effectifs	1	8	7	2	1
Eff. cumulés	22	30	37	39	40
Fréquences	0,025	0,2	0,175	0,05	0,025
Fré. cumulées	0,55	0,75	0,925	0,975	1

L'histogramme de cette distribution est

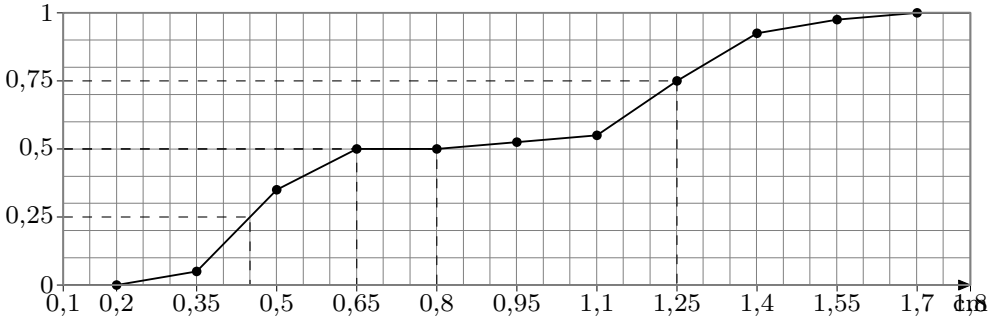


On s'aperçoit que le choix précédent n'est pas le plus adapté à la distribution observée. Il serait notamment nécessaire de raffiner l'amplitude de classes entre 0,2 cm et 0,65 cm, et donc que le choix de classes de même amplitude est à revoir.

On constate sur l'histogramme la présence de deux classes modales. La distribution est donc bimodale. La raison de cette bimodalité est évidente : l'échantillon est composé de la taille de 20 mâles et de 20 femelles ; il apparaît que les mâles sont généralement nettement plus petits que les femelles de cette espèce. Dans ce genre de circonstances, on doit étudier les raisons d'une plurimodalité, en déduire éventuellement des sous-échantillons (ici correspondant aux mâles et aux femelles) et effectuer une étude séparée de ces derniers (si on dispose de renseignements permettant d'effectuer cette séparation).

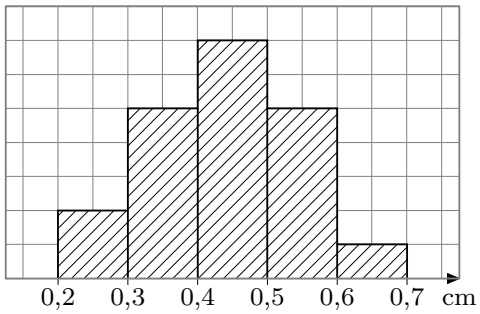
La représentation de la fonction de répartition est appelée *polygone des fréquences cu-*

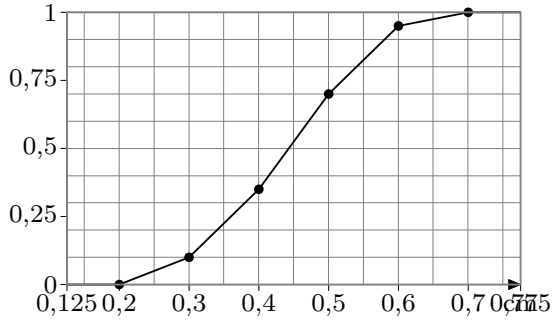
mulées et est ici



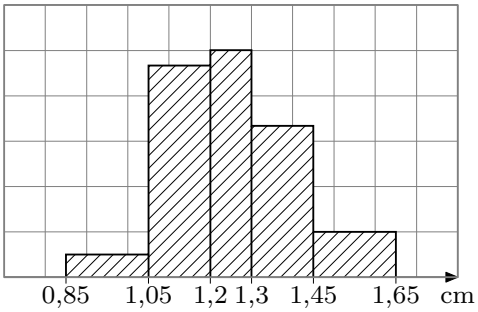
Nous avons indiqué sur le graphique la détermination des quartiles qui sont alors définis à l'aide de la fonction de répartition obtenue après regroupement en classes. Ceux-ci diffèrent généralement des quartiles calculés sur l'échantillon — quartiles qui seront calculés à la sous-section suivante. Notons que la médiane est indéterminée et que toutes les valeurs de l'intervalle [0,65 ; 0,8] conviennent (on prend alors souvent le centre de l'intervalle pour médiane).

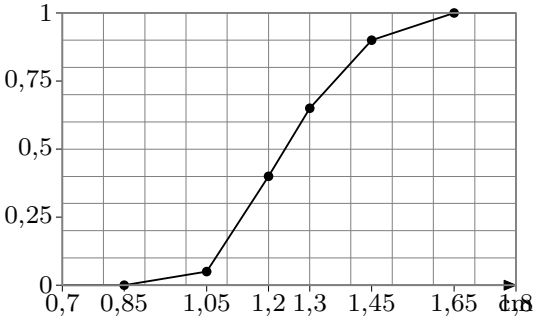
Ci-dessous, nous avons représenté la distribution de la taille des mâles puis de celle des femelles avec de nouvelles répartitions en classes.





Sur les figures ci-dessous, on remarquera que les classes n'ont pas la même amplitude et que la hauteur des bandes est proportionnelle à la fréquence divisée par l'amplitude (comme il se doit).





3. **Diagrammes en boîtes**

La représentation de distributions d'échantillons ou de lois de probabilité de variables quantitatives (discrètes ou continues) par des *diagrammes en boîtes*, ou *boîtes à moustaches*, est devenue un classique incontournable. Notre expérience personnelle nous les a fait découvrir dans le milieu des années 90 lorsque quelques rares enseignants-chercheurs en faisait une promotion *a posteriori* timide à un public d'étudiants biologistes. Quelques années plus tard, des logiciels dits professionnels proposaient ce type de représentations graphiques pour vendre

leur produit auprès de laboratoires centrés autour des Sciences du Vivant. Au début du XXI$^{\text{ème}}$ siècle, ces graphiques ont été inscrits au programme des lycées. Émettons une mise en garde : ce qui marche bien dans un contexte finit par devenir une mode et être utilisé dans des situations inadaptées et finalement être décrié pour cela. Un diagramme simple a l'ambition de sa simplicité.

Un diagramme en boîte est une représentation extrêmement synthétique de distributions *unimodales*. Elle est basée sur un profil de référence, celui des lois normales ; de sorte que toute déviation par rapport à ce profil de référence oblige à suspecter que la distribution considérée ne correspond pas à une loi normale — il faut prendre garde à ne pas faire de réciproque hâtive. Ce type de représentations graphiques est de plus particulièrement adaptée à la comparaison de plusieurs échantillons ou distributions. Il suffit de feuilleter une revue médicale ou pharmaceutique pour en avoir l'illustration.

Un diagramme en boîtes c'est une boîte avec des pattes ou moustaches : la boîte est composée de trois traits placés aux niveaux des quartiles $q_{1/4}$, me et $q_{3/4}$, et est refermée par deux autres traits pour former la boîte. Quant aux pattes ou moustaches, plusieurs conventions sont possibles : les deux moustaches partent de $q_{1/4}$ et $q_{3/4}$ pour atteindre respectivement :

(i) $\max(\min(x_i), q_{1/4} - 1{,}5(q_{3/4} - q_{1/4}))$ et $\min(\max(x_i), q_{3/4} + 1{,}5(q_{3/4} - q_{1/4}))$;

(ii) les quantiles $q_{0{,}01}$ et $q_{0{,}99}$;

(iii) le minimum et le maximum de la distribution.

On ajoute souvent au diagramme obtenu la moyenne de l'échantillon, et, lorsque une extrémité des moustaches ne coïncide pas avec l'extremum correspondant dans l'échantillon, les valeurs observées supérieures à cette extrémité.

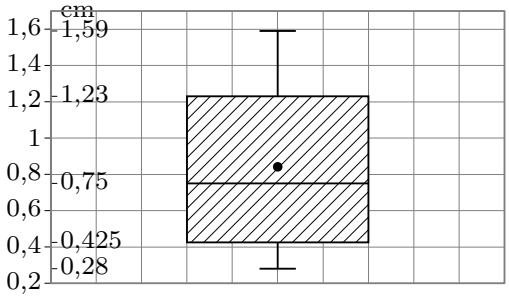
Les diagrammes satisfaisant la première convention sont parfois appelés *diagrammes de Tuckey* et sont très couramment employés. L'amplitude des moustaches y est telle que pour une loi normale leur extrémités sont les quantiles $q_{0{,}01}$ et $q_{0{,}99}$.

Il est à noter qu'alors les moustaches ne s'étendent pas toujours jusqu'aux mininimum et maximum de l'échantillon car leur amplitude est au plus $1{,}5(q_{3/4} - q_{1/4})$. La raison de cette limitation est d'éviter de représenter des valeurs observées qui pourraient être exceptionnelles, voire aberrantes.

Nous privilégierons la première convention, la dernière étant, selon nous, à rejeter.

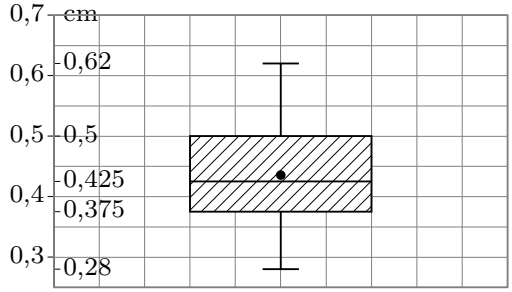
Martelons que de part la forme même de ce type de diagramme, il n'est adapté qu'à des distributions unimodales assez régulières ayant plus ou moins une forme de cloche.

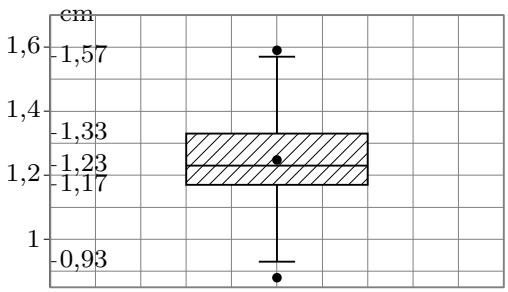
Exemple. — Nous reprenons les données portant sur les veuves noires. L'échantillon étudié dans un premier temps est celui de la taille de l'ensemble des 40 sujets étudiés. La représentation en boîte à moustaches est alors la suivante :



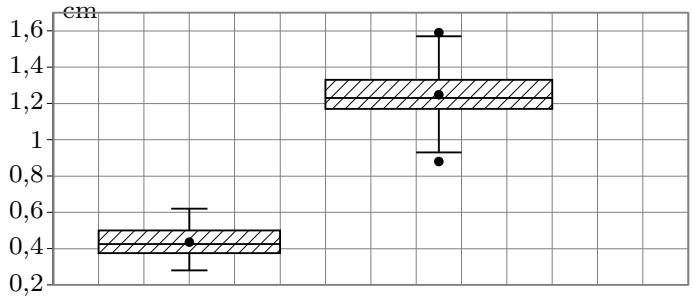
Compte-tenu de la différence de taille entre mâles et femelles, le diagramme précédent n'a guère de sens. On doit étudier séparément les distributions des tailles des mâles et des femelles. Les boîtes à moustaches qui suivent distinguent ces deux sous-échantillons. Celle de gauche

porte sur la taille des mâles, celle de droite, sur celle des femelles.





Nous constatons une grande symétrie de ces distributions par rapport à leurs médianes, médianes qui sont proches des moyennes — bien que la distribution observée de la taille des mâles présente une légère dissymétrie positive. Cependant, les deux représentations précédentes sont faites dans des échelles différentes. Pour comparer les différences entre mâles et femelles, nous sommes amenés à représenter les deux boîtes à moustaches dans une échelle commune :



La boîte à moustaches de la taille des mâles est représentée à gauche, et celle des femelles, à droite. La différence de distribution de taille étant si criante entre les deux sexes que nous nous abstiendrons de tout commentaire.

Exercices

Exercice 1. — Les données étant celles de l'exercice portant sur des malades atteints d'un cancer des poumons, représenter la distribution des observations.

Exercice 2. — Représenter la distribution du nombre d'œufs pondus après une répartition en classes d'extrémités $200, 300, \ldots, 900$; tracer sa fonction de répartition et déterminer les quartiles (selon la première ou la seconde convention), la moyenne et l'écart-type. Comparer ces résultats avec ceux de l'échantillon non réparti en classes.

Exercice 3. — Pour les données portant sur les araignées, calculer les quartiles (selon la première ou la seconde convention), les moyennes, les extrémités de moustaches et retrouver les boîtes à moustaches précédentes. Réaliser la boîte à moustaches du nombre d'œufs pondus par les femelles.

Chapitre IV

ANALYSE STATISTIQUE MULTIVARIÉE, UN EXEMPLE : LES RÉGRESSIONS LINÉAIRES

1. **Introduction**

Jusqu'à présent, nous nous sommes plus particulièrement focalisé sur ce qui pouvait être dit ou étudié sur une variable statistique définie sur un population. L'analyse statistique repose sur les informations que l'on peut obtenir d'un échantillon, c'est-à-dire d'une liste d'individus dans la population. La formation d'un échantillon est l'étape la plus difficile et coûteuse d'une telle étude. Il est souvent profitable de mesurer sur chaque individu plus d'une variable.

Exemple. — Lors des enquêtes téléphoniques portant sur la couverture maladie des individus, il n'est pas simplement demandé le nom de la mutuelle de la personne interrogée, mais aussi un certain nombre d'autres variables qui peuvent avoir l'un des quatre type présentés au premier chapitre (sexe, âge, habitation, catégorie socio-professionnelle [pour cadrer le revenu], etc).

L'étude de chaque variable est qualifiée d'étude statistique univariée, l'étude des variables dans leur ensemble, d'étude statistique multivariée. Ceci est à rapprocher de ce qu'il se passe lorsqu'en Probabilité on introduit la notion de vecteur aléatoire. Un vecteur aléatoire est juste la donnée d'une liste de variables aléatoires définies sur un même espace probabilisé et chacune prenant ses valeurs dans certains espaces mesurables. Le vecteur aléatoire est une variable aléatoire à part entière à valeurs dans l'espace produit correspondant. Parler de vecteur aléatoire annonce qu'on va s'intéresser aux lois respectives de chaques composantes (lois marginales) et essayer de comprendre comment la loi du vecteur (loi conjointe) est reliée à ces lois marginales.

Ne serait-ce qu'avec un exemple aussi banal que celui qui précède, il est évident que la vie réelle présente situations d'analyses statistiques multivariées mélangeant les quatre types de variables statistiques de notre nomenclature. Ces types sont hétérogènes. Analyser de telle données nécessite des outils plus ou moins spécifiques, plus ou moins adaptés. De telles études sont parfois critiquées. Il n'est que des situations très simples où l'accord général peut être obtenu.

Ici, nous n'allons considérer que deux variables, il s'agira donc d'une analyse bivariée. Elles seront quantitatives, continues, notées X et Y. La question qui se posera est : en supposant qu'il existe une relation fonctionnelle entre les deux variables, comment estimer la fonction qui les relie. Nous allons même simplifier notre propos en oubliant toutes sortes de variations aléatoires et simplement faire de l'ajustement de données à un modèle prescrit. Quelques remarques essaieront de compléter cette présentation.

2. **Notations**

On appelle nuage de points la donnée d'une suite $(x_i, y_i)_{i=1}^n$ dans $\mathbb{R}^2$. Celle-ci pourrait être pondérée par des effectifs $(n_i)_{i=1}^n$ (cas de regroupements), des fréquences $(f_i)_{i=1}^n$, ou de probabilités $(p_i)_{i=1}^n$. Nous laissons le cas d'une pondération pour une seconde lecture.

La barre signifiant qu'on considère la moyenne arithmétique, nous notons

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

et aussi

$$s_x^2 = \overline{x^2} - (\bar{x})^2, \qquad s_y^2 = \overline{y^2} - (\bar{y})^2, \qquad \operatorname{cov}(x,y) = \overline{xy} - \bar{x} \times \bar{y}.$$

Essayer de mettre en évidence aussi tôt que possible ces quantités dans les calculs qui suivront nous paraît simplificateur.

3. **Régression linéaire de y en x**

Il est des situations où l'abscisse X est parfaitement contrôlée par l'expérimentateur (on préfère alors utiliser la minuscule x), mais la mesure Y correspondante est perturbée par un bruit aléatoire. La variable x est alors une variable de contrôle ou variable explicative, la variable y étant alors une variable réponse ou expliquée. Pour une valeur x_i donnée par l'expérimentateur, on mesure $Y_i = \phi(x_i) + \varepsilon_i$ où ϕ est une fonction reliant x et y et ε_i est le bruit qui apparaît lors de cette mesure. Le type de relation le plus simple qu'on puisse envisager est celui où ϕ est une fonction affine, $\phi(x) = \alpha + \beta x$. Nous avons alors

$$Y_i = \alpha + \beta x + \varepsilon_i.$$

Si le nuage de points $(x_i, y_i)_{i=1}^n$ est issu de mesures de ce type, il doit plus ou moins se répartir près d'une droite dans le plan. Le problème est d'estimer les coefficients de cette droite. Considérons la fonction

$$\begin{aligned} F : \mathbb{R}^2 &\longrightarrow \mathbb{R}_+ \\ (a,b) &\longmapsto \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - y_i)^2 \end{aligned}$$

Si nous connaissions les coefficients α et β, nous pourrions évaluer

$$F(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

où e_i est le bruit observé pour la mesure de y_i. En supposant le bruit de moyenne nulle et de variance faible, cette quantité devrait être petite et même plus petite que toutes les autres valeurs $F(a,b)$. Ainsi, le nuage de points étant donné nous cherchons à minimiser F. Puisque cette fonction est une somme de carrés, cette minimisation est dite « au sens des moindres carrés ».

Théorème. — *Soit $(x_i, y_i)_{i=1}^n$ un nuage de points. Si $s_x^2 > 0$, alors les coefficients $(\hat{a}, \hat{b})$ minimisant le problème de moindres carrés précédent sont donnés par*

$$\begin{cases} \hat{b} = \operatorname{cov}(x,y)/s_x^2 \\ \hat{a} = \bar{y} - \hat{b} \times \bar{x} \end{cases}$$

Démonstration. — Nous avons donc à trouver $\hat{a}$ et $\hat{b}$ minimisant la fonction

$$F(a,b) = \frac{1}{n}\sum_{i=1}^{n}(a + b \times x_i - y_i)^2, \qquad a,\, b \in \mathbb{R},$$

qui correspond à une forme quadratique définie positive (son graphe est un paraboloïde) et admet donc un unique minimum qui est atteint là où les dérivées partielles de F en a et b s'annulent simultanément :

$$\begin{cases} \dfrac{\partial F}{\partial a}(\hat{a},\hat{b}) = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n}(\hat{a}+\hat{b}\times x_i - y_i) = 0 \\ \dfrac{\partial F}{\partial b}(\hat{a},\hat{b}) = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n} x_i(\hat{a}+\hat{b}\times x_i - y_i) = 0 \end{cases} \quad \text{soit} \quad \begin{cases} 0 = \hat{a} + \dfrac{\hat{b}}{n}\displaystyle\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \\ 0 = \dfrac{\hat{a}}{n}\displaystyle\sum_{i=1}^{n} x_i + \dfrac{\hat{b}}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{cases}$$

et donc avec les notations convenues

$$\begin{cases} 0 = \hat{a} + \hat{b}\times\bar{x} - \bar{y} \\ 0 = \hat{a}\times\bar{x} + \hat{b}\times\overline{x^2} - \overline{xy} \end{cases}$$

puis en substituant $\hat{a}$ dans la seconde équation par son expression donnée par la première

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\times\bar{x} \\ 0 = (\bar{y} - \hat{b}\times\bar{x})\times\bar{x} + \hat{b}\times\overline{x^2} - \overline{xy} \end{cases} \qquad \begin{cases} \hat{a} = \bar{y} - \hat{b}\times\bar{x} \\ 0 = \hat{b}\times\big(\overline{x^2} - (\bar{x})^2\big) - \big(\overline{xy} - \bar{x}\times\bar{y}\big) \end{cases}$$

ce qui donne finalement

$$\begin{cases} \hat{b} = \operatorname{cov}(x,y)/s_x^2 \\ \hat{a} = \bar{y} - \hat{b}\times\bar{x} \end{cases}$$

□

Remarques. — *a*) Avoir $s_x^2 = 0$ signifie que tous les x_i sont égaux. Le nuage de points est alors aligné sur une droite verticale.

b) D'après le théorème de Cauchy–Schwarz, on a

$$r = \frac{\operatorname{cov}(x,y)}{s_x s_y} \in [-1,1]$$

et $|r| = 1$ si et seulement si le nuage de points est exactement aligné sur une droire. Ce coefficient r est appelé coefficient de corrélation linéaire. Suivant les contextes, la proximité de son carré r^2 (le coefficient de détermination) avec 1 peut être quantifié à l'aide d'outils probabilistes pour des modèles convenables. Nous ne le ferons pas.

c) Une fois que $\hat{a}$ et $\hat{b}$ ont été calculés, les différences $(y_i - \hat{a} - \hat{b}x_i)_{i=1}^n$ sont les « résidus de la régression ». Le problème de minimisation considéré consiste donc en la minimisation de la somme des carrés des résidus.

d) Si on modifie les échelles de mesures en x ou en y, les coefficients $\hat{a}$ et $\hat{b}$ en sont affectés d'autant. La régression linéaire de y en x s'avère donc être stable par changement d'unités de mesure. Nous pourrons voir que ce n'est pas le cas de toutes les méthodes de régression.

e) Si on veut faire de la régression linéaire de x en y, il suffit d'adapter les formules. Cependant, la dissymétrie (variables explicative/expliquée) entre les deux types de coordonnées fait que si on a déjà choisi un type de régression, l'autre sera forcément inadapté.

4. **Généralisations de la régression linéaire**

Soient $\phi_0, \dots, \phi_d$ une famille (libre) de fonctions, $(x_1, y_1), \dots, (x_n, y_n)$ un nuage de points (la première coordonnée x pouvant être multi-dimensionnelle). On cherche $\hat{a}_0, \dots, \hat{a}_d \in \mathbb{R}$ tels que $y(x) = \hat{a}_0\phi_0(x) + \cdots + \hat{a}_d\phi_d(x)$ minimise

$$F(a_0, \dots, a_d) = \sum_{j=1}^{n} \big(y_j - y(x_j)\big)^2.$$

Une condition nécessaire d'extremum est ici l'annulation de la différentielle de F : pour $i = 0, \dots, d$,

$$\sum_{j=1}^{n} \phi_i(x_j)\big(y_j - y(x_j)\big) = 0 \qquad \text{soit} \qquad \sum_{j=1}^{n} \phi_i(x_j) \times y(x_j) = \sum_{j=1}^{n} \phi_i(x_j) \times y_j$$

Notons Φ la matrice dont les $d+1$ colonnes correspondent aux fonctions $(\phi_i)_{i=0}^{d}$:

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \dots & \phi_i(x_1) & \dots & \phi_d(x_1) \\ \vdots & & \vdots & & \vdots \\ \phi_0(x_j) & \dots & \phi_i(x_j) & \dots & \phi_d(x_j) \\ \vdots & & \vdots & & \vdots \\ \phi_0(x_n) & \dots & \phi_i(x_n) & \dots & \phi_d(x_n) \end{pmatrix}$$

qui est une matrice de dimensions $(n, d+1)$. Le système précédent s'écrit alors

$$\Phi' \times \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_j) \\ \vdots \\ y(x_n) \end{pmatrix} = \Phi' \times \begin{pmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{pmatrix} = \Phi' \times Y.$$

avec $\Phi' = {}^t\Phi$ la matrice transposée de Φ. Comme on a

$$\Phi \times a = \Phi \times \begin{pmatrix} a_0 \\ \vdots \\ a_i \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} a_0\phi_0(x_1) + \cdots + a_i\phi_i(x_1) + \cdots + a_d\phi_d(x_1) \\ \vdots \\ a_0\phi_0(x_j) + \cdots + a_i\phi_i(x_j) + \cdots + a_d\phi_d(x_j) \\ \vdots \\ a_0\phi_0(x_n) + \cdots + a_i\phi_i(x_n) + \cdots + a_d\phi_d(x_n) \end{pmatrix} = \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_j) \\ \vdots \\ y(x_n) \end{pmatrix}$$

le système s'écrit finalement

$$\Phi' \times \Phi \times a = \Phi' \times \Phi \times \begin{pmatrix} a_0 \\ \vdots \\ a_i \\ \vdots \\ a_d \end{pmatrix} = \Phi' \times \begin{pmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{pmatrix} = \Phi' \times Y.$$

Soit

$$\Phi' \times \Phi \times a = \Phi' \times Y.$$

Ce système peut paraître simple, il l'est, mais son inversion (calcul de $(\Phi' \times \Phi)^{-1}$) pose d'assez gros problèmes numériques lorsque de grands écarts se creusent dans la matrice Φ et donc aussi dans la matrice $\Phi' \times \Phi$ (on parle de problème de conditionnement en analyse numérique).

Remarques. — *a*) Lorsque la régression est polynomiale, soit à ajuster

$$y = a_0 + a_1 x + \cdots + a_i x^i + \cdots a_d x^d,$$

on a $\phi_i(x) = x^i$, $0 \leqslant i \leqslant d$, les matrices Φ' et Φ sont des matrices rectangulaires de Vandermonde (on peut les voir comme sous-matrices de matrices carré de Vandermonde dont les propriétés de rang ou d'inversibilité sont bien connues)

$$\Phi = \begin{pmatrix} 1 & \dots & x_1^i & \dots & x_1^d \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & x_j^i & \dots & x_j^d \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & x_n^i & \dots & x_n^d \end{pmatrix}, \qquad \Phi' = \begin{pmatrix} 1 & \dots & 1 & \dots & 1 \\ \vdots & & \vdots & & \vdots \\ x_1^i & \dots & x_j^i & \dots & x_n^i \\ \vdots & & \vdots & & \vdots \\ x_1^d & \dots & x_j^d & \dots & x_n^d \end{pmatrix}.$$

Le rang de ces matrices est au plus $\min(d+1, n)$. Si tous les x_j sont distincts, et $d+1 \leqslant n$, alors ce rang est $d+1$. La matrice produit $\Phi' \times \Phi$, qui est de dimensions $(d+1, d+1)$, est alors de rang $d+1$. On peut inverser le système. La difficulté est que dès que d est grand (supérieur à 4 par exemple), la matrice Φ est mal conditionnée, le produit $\Phi' \times \Phi$ l'est aussi et les résultats numériques sont souvent assez surprenants.

b) Lorsque $x \in \mathbb{R}^d$, $\phi_0(x) = 1$, $\phi_i(x) = x^{(i)}$ la i-ème coordonnée de x pour $1 \leqslant i \leqslant d$, le problème précédent est nommée régression linéaire multiple. Il consiste simplement à ajuster en fonction des données une relation de la forme

$$y = a_0 + a_1 x^{(1)} + \cdots + a_i x^{(i)} + \cdots + a_d x^{(d)}.$$

5. **Fausses généralisations de la régression linéaire**

Lorsqu'on souhaite ajuster $y = \phi(x)$ avec ϕ non linéaire mais d'un type simple (exponentiel, logarithmique, puissance), il est fréquent de se ramener à une régression linéaire. La non linéarité de la transformation qui permet ce passage fait que le problème de minimisation ne sera pas du type « moindres carrés », et sa validité sera très discutable.

On le fait faute de mieux. L'idéal serait de recourrir à des méthodes numériques (méthode de Newton–Raphson) pour ajuster les coefficients dans ces situations plutôt que de faire n'importe quoi (exemple : voir le manuel de sa calculatrice).

6. **Régression orthogonale**

Étant donné le nuage de points $(x_i, y_i)_{i=1}^n$, ce qu'on cherche à minimiser est

$$F(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} d\big(D(a,b), (x_i, y_i)\big)^2 = \frac{1}{n} \sum_{i=1}^{n} \Big(x_i^2 + (y_i - a)^2 - \frac{1}{1+b^2} (x_i + b y_i - ab)^2 \Big),$$

où $D(a,b)$ est la droite d'équation $y = a + bx$ et d la distance euclidienne dans le plan. Nous cherchons donc la somme des carrés des distances des points $(x_i, y_i)_{i=1}^n$ à la droite d'équation $y = a + bx$.

Pour comprendre comment on obtient l'expression de la fonction F, considérons le cas où $a = 0$. Nous avons à faire à la distance de la droite $D(0,b)$ d'équation $y = bx$ qui passe par l'origine avec un point (x_i, y_i). Le vecteur unitaire directeur de la droite est $u = (1,b)/\|(1,b)\| = (1,b)/\sqrt{1+b^2}$. Le projeté orthogonal de (x_i, y_i) sur $D(0,b)$ est alors $\langle u, (x_i, y_i)\rangle u = (x_i + by_i)/\sqrt{1+b^2} \times u$ dont le carré de la norme euclidienne vaut $(x_i + by_i)^2/(1+b^2)$. D'après le théorème de Pythagore, le carré de la distance $D(0,b)$ avec le point (x_i, y_i) est

$$d\big(D(0,b),(x_i,y_i)\big)^2 = \|(x_i,y_i)\|^2 - (x_i+by_i)^2/(1+b^2) = x_i^2 + y_i^2 - \frac{1}{1+b^2}(x_i+by_i)^2.$$

Pour $a \in \mathbb{R}$, on a alors

$$d\big(D(a,b),(x_i,y_i)\big)^2 = d\big(D(0,b),(x_i,y_i-a)\big)^2 = x_i^2 + (y_i-a)^2 - \frac{1}{1+b^2}\big(x_i+b(y_i-a)\big)^2,$$

ce qui explique l'expression de F.

Théorème. — *Soit $(x_i, y_i)_{i=1}^n$ un nuage de points. Si $\operatorname{cov}(x,y) \neq 0$, alors les coefficients (a_1, b_1) minimisant le problème de moindres carrés précédent sont donnés par*

$$\begin{cases} c = (s_y^2 - s_x^2)/2\operatorname{cov}(x,y) \\ b_1 = c + \sqrt{c^2+1} \\ a_1 = \bar{y} - b_1 \times \bar{x} \end{cases}$$

La droite obtenue est appelée axe principal du nuage de points. La droite de coefficients

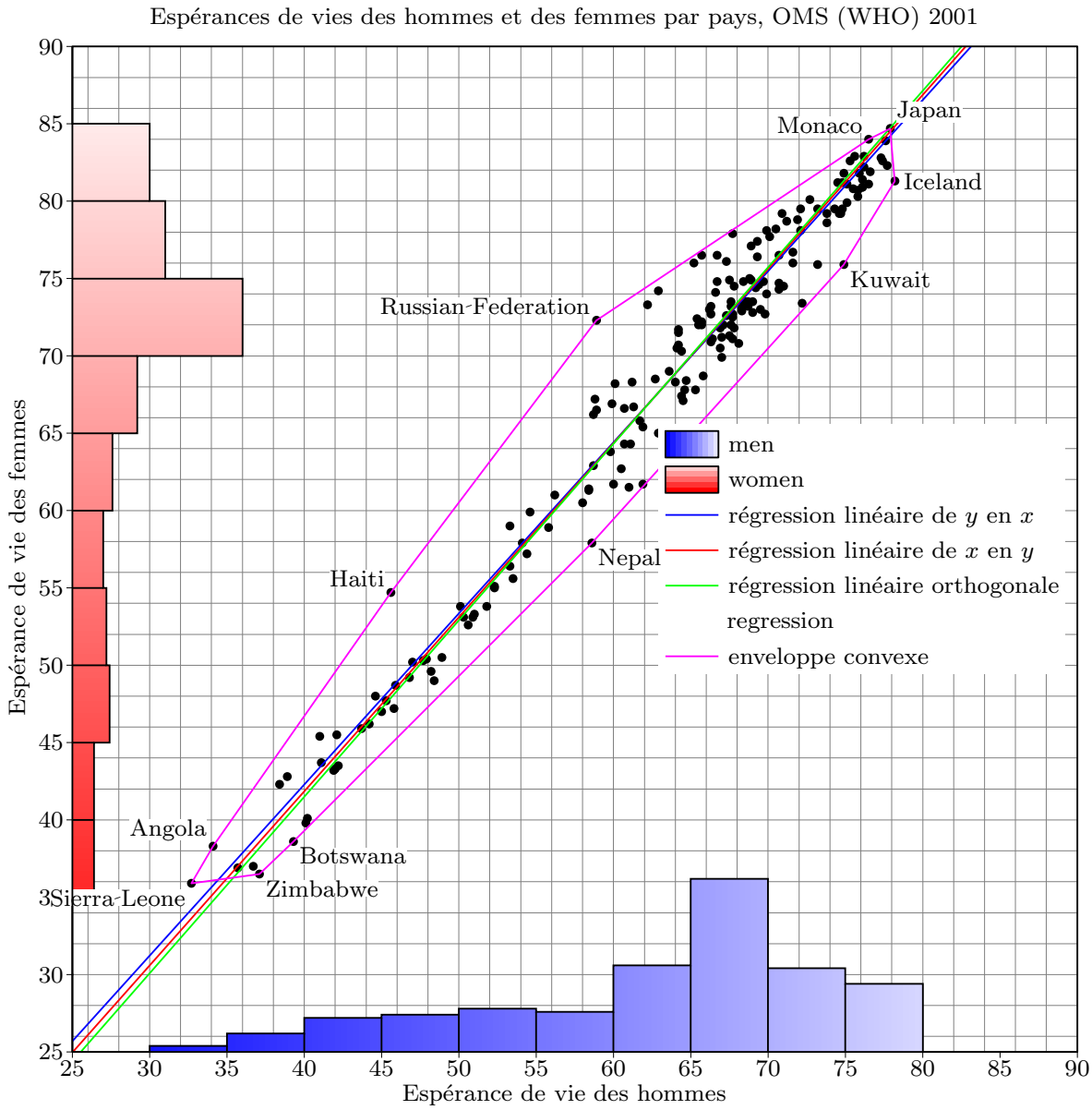
$$\begin{cases} b_2 = c - \sqrt{c^2+1} \\ a_2 = \bar{y} - b_2 \times \bar{x} \end{cases}$$

est orthogonale à l'axe principal et est nommée axe secondaire du nuage de points.

Démonstration. — À rédiger.

Exemple. — En 2002, l'Organisation Mondiale de la Santé (OMS), qui plus tard a été rebaptisée Wolrd Health Organisation (WHO), a publié les espérances de vies des hommes et des femmes de chaques pays membres de l'ONU. Chercher si des deux variables l'une est explicative et l'autre expliquée n'est peut-être pas très judicieux. Les deux variables étant

apparemment homogènes, une régression orthogonale est plus adaptée.



Diagonalisation de la matrice de covariance. — à rédiger.

Dimensions supérieures, introduction à l'analyse en composantes principales. — à rédiger.

Exercices

Exercice 1 (pour l'auteur). — Poursuivre la rédaction de ce chapitre, corriger les fautes de frappe, simplifier les calculs, illustrer d'exemples.

Exercice 2. — Calculer le carré de la distance euclidienne d'un point (x, y) à une droite d'équation $y = a + bx$. On pourra commencer par le cas où $a = 0$.

Exercice 3. — Les données suivantes décrivent l'évolution de la population des États-Unis

au cours de la seconde moitié du XIX^{e} siècle.

Année	1860	1870	1880	1890	1900
Population (en millions d'habitants)	31,4	39,8	50,2	63,0	76,0

On désignera par P l'effectif de la population (en millions d'habitants) et par X le temps écoulé (en années) depuis 1850.

(i) Calculer le coefficient de corrélation linéaire entre X et P et donner l'équation de la droite de régression de P en X. Représenter le nuage de points et la courbe de régression sur un même graphique.

(ii) On se propose d'ajuster le nuage de points par un modèle de type exponentiel

$$P = \exp(a'X + b').$$

Pour ce faire, on se ramène à un modèle linéaire au moyen du changement de variable $Y = \ln P$. Ceci vous paraît-il mieux adapté aux données ? Justifier la réponse (on pourra par exemple comparer des quantités telles que les coefficients de corrélation linéaire et les sommes des carrés des résidus obtenus par les deux modèles).

(iii) Estimer avec ces deux méthodes la population des États-Unis en 1930 et calculer les résidus correspondants (la population réelle en 1930 était de 112,8 millions d'habitants).

Exercice 4. — On considère la série suivante :

t_i	1	2	3	4	5	6	7	8	9	10
y_i	58	40	31	15	18	15	9	9	10	8

(i) Représenter graphiquement cette série.

(ii) On se propose d'ajuster une tendance f de la forme $f(t) = \dfrac{1}{a + b \times t}$. Justifier ce choix.

(iii) Déterminer les coefficients a et b en utilisant un changement de variables approprié :

a) par la méthode des deux points (en les choisissant judicieusement) ;

b) par régression linéaire.

(iv) Représenter les deux tendances ainsi obtenues sur le graphique précédent et comparer les résultats. Est-ce que les résidus ont une allure irrégulière ?

Chapitre V

ESTIMATION ET INTERVALLES DE CONFIANCE

1. **Estimation ponctuelle**

Le théorème suivant dresse la liste des estimateurs consistants, et, à une exception près, sans biais qu'il est impératif de connaître. Celui-ci repose essentiellement sur l'application de la loi faible des grands nombres (LGN) (convergence en probabilité des moyennes de Césaro vers l'espérance), bien que la loi forte des grands nombres (LFGN) (convergence presque sûre des moyennes de Césaro vers l'espérance) s'applique aussi compte-tenu des hypothèses envisagées — ces estimateurs sont donc aussi fortement consistants. Cependant, la convergence en probabilité est la seule convergence réellement pertinente dans le cadre de l'estimation statistique.

Théorème. — *Soit $(X_1, \dots, X_n)$ un échantillon d'une variable X, ou d'une loi μ, quantitative (réelle) de moyenne m et de variance σ^2.*

(i) *Pour tout $x \in \mathbb{R}$,*

$$\mathbb{E}[F_{X,n}(x)] = \mu(]-\infty, x]) = \mathbb{P}\{X \leqslant x\} = F_X(x),$$

et, lorsque $n \to \infty$, les variables aléatoires $F_{X,n}(x)$ tendent en probabilité vers le réel $F_X(x)$.

(ii) *On a*

$$\mathbb{E}\big[\bar{X}_n\big] = \mathbb{E}[X] = m,$$

et, lorsque $n \to \infty$, les variables aléatoires $\bar{X}_n$ tendent en probabilité vers le réel m.

(iii) *On a*

$$\mathbb{E}\big[S_n^2\big] = \frac{n-1}{n}\operatorname{Var}(X) = \frac{n-1}{n}\,\sigma^2,$$

et, lorsque $n \to \infty$, les variables aléatoires S_n^2 tendent en probabilité vers le réel σ^2.

Démonstration. — Pour le point (i), on a

$$\begin{aligned}\mathbb{E}\big[F_{X,n}(x)\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{1}_{]-\infty,x]}(X_i)\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\{X_i \leqslant x\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\{X \leqslant x\} = \mathbb{P}\{X \leqslant x\}\,;\end{aligned}$$

les variables $Y_i = \mathbb{1}_{]-\infty,x]}(X_i)$ sont i.i.d, intégrables, comme $\bar{Y}_n = F_{X,n}(x)$, par la LGN, on en déduit que $F_{X,n}(x)$ tend en probabilité vers $\mathbb{P}\{X \leqslant x\}$ quand n tend vers l'infini. Pour le point (ii), il est clair, par linéarité, que $\mathbb{E}[\bar{X}_n] = m$ et la LGN permet de conclure. Pour le point (iii), rappelons que $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$. Puisque les X_i^2 sont des variables aléatoires i.i.d. intégrables, par la LGN, $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ tend en probabilité vers $\mathbb{E}[X^2]$. La fonction $x \mapsto x^2$ étant continue et $\bar{X}_n$ tendant en probabilité vers la constante m, alors $(\bar{X}_n)^2$ tend en probabilité vers m^2. Ainsi S_n^2 tend en probabilité vers $\mathbb{E}[X^2] - m^2 = \sigma^2$. Le calcul de $\mathbb{E}[S_n^2]$ aura été laissé en exercice.

Remarques. — *a*) La statistique S_n^2 est un estimateur consistant mais biaisé de σ^2. Ainsi, on préfère utiliser $\widehat{S}_n^2$ qui est un estimateur consistant et sans biais de σ^2.

b) Une méthode mathématique permet d'associer à un paramètre un estimateur. Il s'agit du *Principe du Maximum de Vraisemblance*. L'étude de cette méthode sort du propos de ces notes.

2. **Estimation par intervalle**

Généralement, une valeur ponctuelle ℓ_n d'une statistique Λ_n n'est pas considérée comme étant une estimation suffisante ou pertinente d'un paramètre λ car le plus souvent Λ_n a une loi proche d'être diffuse (n grand) et donc $\mathbb{P}\{\Lambda_n = \ell_n\} \approx 0$. On remplace alors cette *estimation ponctuelle* par une *estimation par intervalle* qui est la donnée d'un intervalle aléatoire, le plus souvent centré autour de l'estimateur, qui a une forte probabilité de contenir le paramètre λ, c'est-à-dire que son amplitude a été déterminée de telle sorte que cette probabilité soit grande.

Définition 7. — Soit $\alpha \in]0,1[$. On appelle *intervalle de confiance* du paramètre λ au niveau de confiance $1-\alpha$, ou au seuil α, tout intervalle aléatoire $[\Lambda_{\min}, \Lambda_{\max}]$, d'extrémités des statistiques $\Lambda_{\min}$ et $\Lambda_{\max}$ de l'échantillon, tel que

$$\mathbb{P}\big\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\big\} = \mathbb{P}\{\Lambda_{\min} \leqslant \lambda \leqslant \Lambda_{\max}\} \geqslant 1-\alpha.$$

On définit les notions voisines suivantes :

(i) l'intervalle est un *intervalle de confiance exact* si $\mathbb{P}\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\} = 1-\alpha$;

(ii) l'intervalle est un *intervalle de confiance approximatif* si $\mathbb{P}\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\} \approx 1-\alpha$;

(iii) l'intervalle est un *intervalle de confiance asymptotiquement exact*, ou plus simplement *asymptotique*, si

$$\mathbb{P}\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\} \to 1-\alpha \qquad \text{lorsque } n \text{ tend vers l'infini.}$$

Évidemment, on souhaite pouvoir prendre α petit pour avoir une forte probabilité que le paramètre soit dans l'intervalle, mais aussi que l'amplitude de l'intervalle soit petite. Le coût de cette double exigence est d'avoir à considérer de grands échantillons puisque, si l'estimateur Λ_n est consistant, la loi de Λ_n se concentre autour du paramètre λ lorsque n est grand.

On pourrait penser que l'idéal serait toujours disposer d'intervalles de confiance exacts. Même si c'était possible, il existerait certainement des situations pour lesquelles les déterminations numériques seraient trop lourdes en regard du gain de précision apporté dans la pratique. De plus, il existe des situations très communes (comme celle qui suit) où par nature même les intervalles de confiance ne peuvent être qu'approximatifs. On essaie alors de s'assurer que l'inégalité $\mathbb{P}\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\} = 1-\alpha$ est satisfaite.

Noter que qu'un intervalle de confiance asymptotiquement exact est approximatif. Cependant, il est rare que l'inégalité précédente soit garantie dans ce cadre. Il est fréquent de qualifier d'« exact » des intervalles de confiance approximatifs non asymptotiques (validité de l'approximation à n fixé). Cet usage sera précisé par l'expression « dit exact » dans la suite.

3. **Estimation d'une proportion**

Nous considérons un réel $p \in]0,1[$ qui peut apparaître comme la proportion d'individus dans une population vérifiant une certaine propriété, ou plus simplement une probabilité. La variable statistique X prend les valeurs 0 ou 1, et a pour loi $\mathcal{B}(1,p)$, la loi de Bernoulli de paramètre p, paramètre qu'on cherche à estimer. Rappelons au passage que $\mathbb{E}[X] = p$ et $\mathrm{Var}(X) = p(1-p)$.

Soit $(X_1,\ldots,X_n)$ un échantillon de la loi de Bernoulli de paramètre p (ou de X). La statistique $\bar{X}_n$ est un estimateur (consistant et sans biais) de p et une valeur observée $\bar{x}_n = k/n$ est une estimation (ponctuelle) de la proportion p. Dans ce contexte, l'expression de la statistique S_n^2 est assez particulière : puisque chaque X_i est à valeurs dans $\{0,1\}$,

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

Par conséquent, dans le cas de l'étude d'une proportion, la variance observée s_n^2 ne nécessite pas de calcul spécifique, mais n'apporte pas non plus d'information supplémentaire à la donnée de la proportion observée $\bar{x}_n$.

Nous ne proposons ici que la détermination d'un intervalle de confiance asymptotique de la proportion p. Une discussion plus détaillée est donnée en annexe.

Remarque. — Le modèle statistique $(E,\mathcal{E},(\mu_\theta)_{\theta\in\Theta})$ considéré est $E = \{0,1\}$, $\mathcal{E} = \mathcal{P}(E) = \{\varnothing,\{0\},\{1\},\{0,1\}\}$, $\Theta =]0,1[$ et pour $\theta \in \Theta$, $\mu_\theta = \mathcal{B}(1,\theta)$.

Une approche (un peu trop) rapide. — Pour $p \in]0,1[$, le théorème central limite affirme que

$$\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$$

est de loi proche de la loi normale $\mathcal{N}(0,1)$ lorsque n est grand. On a en particulier

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leqslant \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \leqslant z_{1-\alpha/2}\right\} \approx 1-\alpha \qquad (*)$$

où $z_{1-\alpha/2} = -z_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$. Puisque $\bar{X}_n$ approche p, on devrait avoir aussi

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leqslant \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n}} \leqslant z_{1-\alpha/2}\right\} \approx 1-\alpha$$

ce qui s'écrit

$$\mathbb{P}\left\{\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leqslant p \leqslant \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right\} \approx 1-\alpha$$

D'où l'intervalle de confiance asymptotique usuel de la proportion p :

$$\left[\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\,,\ \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

Ce type d'intervalle de confiance asymptotique d'une proportion était déjà connu de Laplace[1] puisqu'il en est fait mention dans son traité *Théorie analytique des probabilités* (1812), p. 283.

1. Pierre-Simon de Laplace (23 mars 1749 – 5 mars 1827), mathématicien et astronome français.

Pratique. — Si on dispose d'un échantillon observé $(x_1, \ldots, x_n)$ comportant k réponses positives tel que

$$n \geqslant 50, \qquad k \geqslant 10 \qquad \text{et} \qquad n - k \geqslant 10,$$

on détermine $\bar{x}_n = k/n$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance asymptotique de la proportion p.

Remarque. — Les conditions d'utilisation de l'intervalle de confiance asymptotique ci-dessus sont données généralement pour $\alpha = 0{,}05$. D'autres conditions peuvent être demandées, en particulier pour d'autres valeurs de α.

Une approche plus prudente. — Revenons à $(*)$ qui s'écrit aussi

$$\mathbb{P}\left\{ \frac{\big(\bar{X}_n - p\big)^2}{p(1-p)/n} \leqslant z_{1-\alpha/2}^2 \right\} \approx 1 - \alpha$$

où encore

$$\mathbb{P}\Big\{ \big(1 + z_{1-\alpha/2}^2/n\big)p^2 - \big(2\bar{X}_n + z_{1-\alpha/2}^2/n\big)p + \bar{X}_n^2 \leqslant 0 \Big\} \approx 1 - \alpha.$$

On voit apparaître un polynôme du second degré en la variable p qui est de coefficient dominant strictement positif. Ainsi, il est négatif lorsque p est entre ses racines :

$$P_{\min/\max} = \frac{\bar{X}_n + z_{1-\alpha/2}^2/2n \mp \sqrt{z_{1-\alpha/2}^2/n \times \big(z_{1-\alpha/2}^2/4n + \bar{X}_n(1-\bar{X}_n)\big)}}{1 + z_{1-\alpha/2}^2/n}$$

qui définissent donc l'intervalle de confiance asymptotique cherché. Celui-ci est meilleur que le précédent si on se fonde uniquement sur l'approximation du théorème central limite. Il présente de plus le grand avantage de fournir des bornes qui sont toujours dans $[0, 1]$ contrairement à ce qu'il se passe pour l'approche rapide. Cette méthode semble due à Wilson[2] (1927).

Pratique. — Identique à celle qui précède avec des calculs en plus.

Remarque. — Il est à remarquer que si on néglige les termes d'ordre strictement supérieur à $1/\sqrt{n}$, on a

$$[P_{\min}, P_{\max}] \approx \left[\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n\big(1-\bar{X}_n\big)}{n}}\,,\ \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n\big(1-\bar{X}_n\big)}{n}} \right],$$

c'est-à-dire qu'on retrouve l'intervalle de confiance obtenu précédemment de manière un peu rapide. On trouvera dans la section portant sur les tests paramétriques un raisonnement semblable. Dans la sous-section suivante, nous retiendrons aussi l'approche rapide, et ce, par seul souci de simplicité.

2. Edwin Bidwell Wilson (25 avril 1979 – 28 décembre 1964), mathématicien et érudit américain.

4. **Estimation de la loi d'une variable discrète**

Soient $(X_1, \ldots, X_n)$ un échantillon d'une loi μ discrète supportée à un nombre fini ou dénombrable de points $e_1, \ldots, e_j, \ldots$, et $\mu_j = \mu(e_j)$. Fixons j et considérons, pour $i = 1, \ldots, n$, $Y_i = \mathbb{1}_{\{X_i = e_j\}}$. Ce sont des variables aléatoires indépendantes et identiquement distribuées de loi de Bernoulli de paramètre μ_j et on a $\bar{Y}_n = f_{X,n,j}$ la fréquence empirique de la valeur e_j qui est un estimateur consistant et sans biais de la probabilité μ_j. D'après la sous-section précédente, si $0 < \mu_j < 1$, et *si n est grand*, on a pour intervalle de confiance approximatif de la probabilité μ_j au seuil α

$$\left[f_{X,n,j} - z_{1-\alpha/2}\sqrt{\frac{f_{X,n,j}(1-f_{X,n,j})}{n}}\,,\; f_{X,n,j} + z_{1-\alpha/2}\sqrt{\frac{f_{X,n,j}(1-f_{X,n,j})}{n}} \right],$$

où $z_{1-\alpha/2}$ est, comme précédemment, le quantile d'ordre $1-\alpha/2$ de la loi Normale.

Pratique. — Si on dispose d'un échantillon observé $(x_1, \ldots, x_n)$, on détermine $f_{x,n,j}$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance pour chaque j, ou pour un j particulier suivant son intérêt.

Remarques. — *a*) Si pour un certain j, $\mu_j = 0$, alors e_j n'appartient pas au support de la loi discrète μ et n'est donc pas à considérer. Si pour un certain j, $\mu_j = 1$ alors le support de la loi μ est réduit à e_j et alors (presque sûrement) toutes les observations sont égales à e_j. Dès lors, toute estimation conduira (presque sûrement) au seul résultat possible.

b) Le modèle statistique $(E, \mathcal{E}, (\mu_\theta)_{\theta\in\Theta})$ considéré est $E = \{e_1, \ldots, e_j, \ldots\}$, $\mathcal{E} = \mathcal{P}(E)$, Θ l'ensemble des suites de nombres positifs indexées par E dont la somme vaut 1, et pour $\theta \in \Theta$, μ_θ est la mesure de probabilité associée à la suite de nombres θ.

5. **Estimation de la moyenne d'une loi normale**

Soient $(X_1, \ldots, X_n)$ un échantillon d'une loi normale μ de moyenne m et de variance σ^2. Rappelons que $\bar{X}_n$ est un estimateur consistant et sans biais de la moyenne m. On sait que

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \qquad \text{et} \qquad \frac{\bar{X}_n - m}{S_n/\sqrt{n-1}} = \frac{\bar{X}_n - m}{\widehat{S}_n/\sqrt{n}}$$

suivent respectivement la loi normale $\mathcal{N}(0,1)$ et la loi de Student à $n-1$ degrés de liberté. Ces statistiques permettent de déterminer des intervalles de confiance suivant qu'on ait à sa disposition la valeur de σ ou non.

Si σ est connu. — Nous pouvons considérer la statistique $(\bar{X}_n - m)/(\sigma/\sqrt{n})$ dont la loi est la loi normale $\mathcal{N}(0,1)$. Comme auparavant, soit $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi Normale. Ainsi, on a exactement

$$\mathbb{P}\left\{ \left| \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \right| \leqslant z_{1-\alpha/2} \right\} = 1 - \alpha,$$

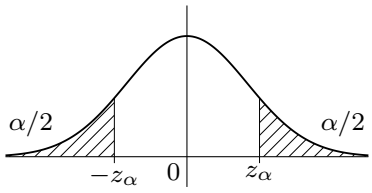
d'où on déduit l'intervalle de confiance exact au seuil α de la moyenne m :

$$\left[\bar{X}_n - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\,,\; \bar{X}_n + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \right].$$

Pratique. — Si on dispose d'un échantillon observé $(x_1, \ldots, x_n)$, on détermine $\bar{x}_n$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — Le modèle statistique $(E, \mathcal{E}, (\mu_\theta)_{\theta\in\Theta})$ considéré est $E = \mathbb{R}$, $\mathcal{E} = \mathcal{B}(\mathbb{R})$, $\Theta = \mathbb{R}$ et pour $\theta \in \Theta$, $\mu_\theta = \mathcal{N}(\theta, \sigma^2)$.

Si σ est inconnu. — Nous pouvons considérer la statistique $(\bar{X}_n - m)/(S_n/\sqrt{n-1})$ dont la loi est la loi de Student à $n-1$ degrés de liberté. De même qu'auparavant, soit $t_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté.



Ainsi, on a exactement

$$\mathbb{P}\left\{\left|\frac{\bar{X}_n - m}{S_n/\sqrt{n-1}}\right| \leqslant t_{1-\alpha/2}\right\} = 1-\alpha,$$

d'où on déduit l'intervalle de confiance exact au seuil α de la moyenne m :

$$\left[\bar{X}_n - t_{1-\alpha/2}\,\frac{S_n}{\sqrt{n-1}},\ \bar{X}_n + t_{1-\alpha/2}\,\frac{S_n}{\sqrt{n-1}}\right].$$

Pratique. — Si on dispose d'un échantillon observé $(x_1, \ldots, x_n)$, on détermine $\bar{x}_n$, s_n et $t_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — Le modèle statistique $(E, \mathcal{E}, (\mu_\theta)_{\theta\in\Theta})$ considéré est $E = \mathbb{R}$, $\mathcal{E} = \mathcal{B}(\mathbb{R})$, $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ et pour $\theta = (\theta_1, \theta_2) \in \Theta$, $\mu_\theta = \mathcal{N}(\theta_1, \theta_2^2)$.

Exemple. — Entre 1851 et 1860, L.-Ad. Bertillon releva la taille de 1 101 178 conscrits. Il put observer que la distribution des tailles était indubitablement normale et observa une moyenne de 163,814 cm et un écart-type corrigé de 6,158 cm. On relève dans un journal que les jeunes gens français durant cette période avait une taille moyenne de 1,65 m. Peut-on être plus précis ?

Nous devons estimer la moyenne d'une loi normale d'écart-type inconnu. Comme $n = 1\,101\,178$ est très grand, la loi de Student à $n-1$ degrés de liberté peut être remplacée par la loi normale $\mathcal{N}(0,1)$. On obtient ainsi pour intervalle de confiance de niveau de confiance $\alpha = 95\,\%$

$$\left[\bar{x} - z_{0,975} \times \hat{s}_n/\sqrt{n}\,,\ \bar{x} + z_{0,975} \times \hat{s}_n/\sqrt{n}\right] \approx [163{,}802\,;\ 163{,}826].$$

Le niveau de confiance étant assez large, il semble bien qu'on puisse préciser que la moyenne devait être de 1,638 m. Néanmoins, ça ne fait guère de différence.

6. **Estimation de la moyenne d'une loi de probabilité**

Soient $(X_1, \ldots, X_n)$ un échantillon d'une loi de probabilité μ admettant une moyenne m et une variance σ^2. Rappelons que $\bar{X}_n$ est un estimateur consistant et sans biais de la moyenne m. Lorsque n est grand, les statistiques

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \qquad \text{et} \qquad \frac{\bar{X}_n - m}{S_n^2/\sqrt{n}} \approx \frac{\bar{X}_n - m}{S_n^2/\sqrt{n-1}}$$

sont proches et suivent approximativement la loi normale $\mathcal{N}(0,1)$. Ces statistiques permettent de déterminer des intervalles de confiance suivant qu'on ait à sa disposition la valeur de σ ou non.

Si σ est connu. — Nous pouvons considérer la statistique $(\bar{X}_n - m)/(\sigma/\sqrt{n})$ dont la loi est proche de la loi normale $\mathcal{N}(0,1)$. Comme auparavant, soit $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi Normale. Ainsi, on a approximativement

$$\mathbb{P}\left\{\left|\frac{\bar{X}_n - m}{\sigma/\sqrt{n}}\right| \leqslant z_{1-\alpha/2}\right\} \approx 1-\alpha,$$

d'où on déduit l'intervalle de confiance approximatif au seuil α de la moyenne m :

$$\left[\bar{X}_n - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\,,\ \bar{X}_n + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right].$$

Pratique. — Si on dispose d'un échantillon observé $(x_1,\dots,x_n)$, on détermine $\bar{x}_n$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — Le modèle statistique $(E,\mathcal{E},(\mu_\theta)_{\theta\in\Theta})$ considéré est $E=\mathbb{R}$, $\mathcal{E}=\mathcal{B}(\mathbb{R})$, Θ l'ensemble des lois de probabilité sur $(E,\mathcal{E})$ de variances finies égales à σ^2.

Si σ est inconnu. — Nous pouvons considérer la statistique $(\bar{X}_n - m)/(S_n/\sqrt{n})$ dont la loi est proche de la loi normale $\mathcal{N}(0,1)$. De même qu'auparavant, soit $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi Normale. Ainsi, on a approximativement

$$\mathbb{P}\left\{\left|\frac{\bar{X}_n - m}{S_n/\sqrt{n}}\right| \leqslant t_{1-\alpha/2}\right\} \approx 1-\alpha,$$

d'où on déduit l'intervalle de confiance approximatif au seuil α de la moyenne m :

$$\left[\bar{X}_n - t_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}},\ \bar{X}_n + t_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}}\right].$$

Pratique. — Si on dispose d'un échantillon observé $(x_1,\dots,x_n)$, on détermine $\bar{x}_n$, s_n et $t_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — Le modèle statistique $(E,\mathcal{E},(\mu_\theta)_{\theta\in\Theta})$ considéré est $E=\mathbb{R}$, $\mathcal{E}=\mathcal{B}(\mathbb{R})$, Θ l'ensemble des lois de probabilité sur $(E,\mathcal{E})$ de variances finies.

Exercice 1. — Les données étant celles de l'exercice portant sur des malades atteints d'un cancer des poumons, donner une estimation de la proportion de fumeurs dans la population. Donner un intervalle de confiance au niveau de confiance 0,9 de cette proportion.

Exercice 2. — Avant une élection, 1 200 personnes ont été interrogées sur leur intention de vote. Parmi elles, 219 ont exprimé leur intention de s'abstenir ou de voter blanc. Donner, en pourcentages, un intervalle de confiance du nombre d'abstentionnistes au seuil 5 % ($\alpha = 0{,}05$).

Exercice 3. — La taille d'une femelle veuve noire est distribuée suivant une loi normale dont l'écart-type est connu et vaut 0,16 cm. Déterminer un intervalle de confiance pour la moyenne de cette loi au niveau de confiance 0,95.

La taille d'un mâle veuve noire est distribuée suivant une loi normale dont l'écart-type est n'est pas précisément connu. Déterminer un intervalle de confiance pour la moyenne de cette loi au niveau de confiance 0,95.

Chapitre VI

TESTS D'HYPOTHÈSES

Nous commençons cette section par des généralités relativement abstraites qui nous semblent être cependant un juste milieu entre une présentation simpliste et une exposition excessivement théorique, toutes deux ayant leurs défauts propres. On pourra en avoir une première lecture pour se faire une vague idée de la problématique et se lancer dans la lecture des sous-sections suivantes. On y reviendra par la suite afin de comprendre, fort(e) de la connaissance d'exemples de tests particuliers, la démarche générale.

1. **Généralités**

Soient $(E, \mathcal{E}, (\mu_\theta)_{\theta\in\Theta})$ un modèle statistique et $\mathcal{H} : \Theta \to \{\text{vrai}, \text{faux}\}$. Le problème est que, pour $\theta \in \Theta$ donné, on n'accède pas directement à $\mathcal{H}(\theta)$ mais seulement à des observations d'échantillons $(X_1^\theta, \dots, X_n^\theta)$ de la loi μ_θ. Pour décider si $\mathcal{H}(\theta)$ est ou bien vraie ou bien fausse, on formule une procédure, appelée *règle de décision*, qui en fonction de $(X_1^\theta, \dots, X_n^\theta)$ conduit à l'acceptation ou le rejet de l'hypothèse selon laquelle $\mathcal{H}(\theta)$ est vraie. Il est évidemment possible que la décision obtenue soit en contradiction avec la réalité, aussi se donne-t-on certaines marges d'erreurs.

Définition 8. — Tester au seuil $\alpha \in]0, 1[$, ou au niveau de confiance $1 - \alpha$,

$$H_0 : \mathcal{H}(\theta) \text{ est vraie} \qquad \text{contre} \qquad H_1 : \mathcal{H}(\theta) \text{ est fausse}$$

— l'hypothèse nulle contre l'hypothèse alternative —, consiste en :

(i) la donnée d'une *règle de décision* au seuil α

$$f_n : E^n \to \mathbb{R}, \qquad A_n^\alpha \subset \mathbb{R} \text{ (un intervalle)} \quad \text{et} \quad R_n^\alpha = (A_n^\alpha)^c,$$

où A_n^α est appelée *région d'acceptation* et R_n^α *région de rejet* au seuil α ;

(ii) la donnée d'un échantillon $(X_1^\theta, \dots, X_n^\theta)$ de la loi μ_θ ;

(iii) en posant $\Lambda_n^\theta = f(X_1^\theta, \dots, X_n^\theta)$ la *statistique du test*, décider d'accepter H_0 si $\Lambda_n^\theta \in A_n^\alpha$, décider de rejeter H_0 sinon, c'est-à-dire si $\Lambda_n^\theta \in R_n^\alpha$.

La règle de décision $(f_n, A_n^\alpha, R_n^\alpha)$ dépend du type d'hypothèse à tester, du seuil α fixé pour opérer ce test et de la taille n de l'échantillon considéré. Définir une règle de décision pour un type d'hypothèses donné est du ressort du statisticien qui se doit de respecter les contraintes suivantes :

$$\begin{aligned} &\text{si } \mathcal{H}(\theta) = \text{vrai} \quad \text{alors} \quad \beta_\theta = \mathbb{P}\{\Lambda_n^\theta \in R_n^\alpha\} \leqslant \alpha, \\ &\text{si } \mathcal{H}(\theta) = \text{faux} \quad \text{alors} \quad 1 - \beta_\theta = \mathbb{P}\{\Lambda_n^\theta \in A_n^\alpha\} \text{ aussi petit que possible.} \end{aligned}$$

Le réel β_θ est une fonction de $\theta \in \Theta$ à valeurs dans $[0, 1]$ (en toute rigueur, il dépend aussi de α, de n et de la règle de décision mais ceux-ci sont fixés). On souhaite que sa restriction à $\Theta_0 = \{\theta \in \Theta : \mathcal{H}(\theta) = \text{vrai}\}$ soit proche de 0, c'est-à-dire que la probabilité de rejeter H_0 alors que H_0 est vraie est faible (*erreur de première espèce*), et que sa restriction à $\Theta_1 = \{\theta \in \Theta : \mathcal{H}(\theta) = \text{faux}\}$ soit proche de 1, c'est-à-dire, par passage au complémentaire,

que la probabilité d'accepter H_0 alors que H_0 est fausse est faible (*erreur de seconde espèce*). Le tableau ci-dessous illustre ces diverses éventualités.

Décision \ θ	Θ_0	Θ_1
Acceptation de H_0	$1-\beta_\theta$	$1-\beta_\theta$ Erreur de 2^{nde} espèce
Rejet de H_0	$\beta_\theta \quad (\beta_\theta \leqslant \alpha)$ Erreur de $1^{\text{ère}}$ espèce	β_θ

La restriction de $\theta \mapsto \beta_\theta$ à $\Theta_1 = \{\theta \in \Theta : \mathcal{H}(\theta) = \text{faux}\}$ est appelée *puissance du test* : plus celle-ci est proche de 1, plus le test est puissant puisque la probabilité de commettre une erreur de seconde espèce est faible.

Remarques. — *a*) Les nombres n et α sont souvent des données du problème et on se dispense, en général, de les mentionner — sauf si, bien sûr, l'étude porte sur plusieurs valeurs de α ou de n. De même, θ est « l'état de la nature » du système étudié, aussi n'en fait-on pas mention dans un test concret puisqu'il n'est, en général, pas amené à varier.

b) L'erreur de première espèce est généralement bien contrôlée par la formulation même de la règle de décision et il arrive souvent qu'elle soit égale à α — le test est alors exactement au seuil α, sinon, il n'est qu'au seuil au plus α. En revanche, ça n'est pas toujours le cas de l'erreur de seconde espèce. L'usage veut qu'on confonde les erreurs de première et de seconde espèces avec leurs valeurs maximales, mais nous ne nous étenderons pas plus sur ces questions.

c) La plupart des tests sont formulés de sorte que H_0 est supposée vraie *a priori* et que seules des observations contredisant fortement cette idée de départ autorisent à rejeter cette opinion première. Selon le principe qu'il est plus facile de conserver son opinion *a priori* que de l'abandonner, on constate qu'une procédure de test est dissymétrique parce qu'elle favorise H_0. Dans la vie courante ceci peut être fondamental : il est préférable — notamment pour l'intéressé(e) — de tester

$$H_0 : \text{l'étudiant(e) est bon(ne)} \qquad \text{contre} \qquad H_1 : \text{l'étudiant(e) est nul(le)},$$

alors qu'un laboratoire pharmaceutique testera (enfin, on l'espère)

$$H_0 : \text{cette drogue est nocive} \qquad \text{contre} \qquad H_1 : \text{cette drogue n'est pas nocive.}$$

Ainsi, ce qui est véritablement significatif est le rejet de H_0. En général, si on a été amené à accepter H_0, une étude plus fine doit suivre — étude qui n'est pas nécessairement du ressort de la Statistique et peut être plus pragmatique (contrôle des fondements scientifiques ou techniques de l'évaluation ou de l'élaboration des objets d'étude).

2. **Généralités sur les tests paramétriques**

On cherche à comparer un paramètre réel $m = m_\theta$ d'une loi $\mu = \mu_\theta$, par exemple sa moyenne, à une valeur donnée m_0.

On appelle *test bilatéral* du paramètre m relativement à la valeur m_0, un test dont les hypothèses sont

$$H_0 : m = m_0 \qquad \text{contre} \qquad H_1 : m \neq m_0.$$

On appelle *test unilatéral* du paramètre m relativement à la valeur m_0, un test dont les hypothèses sont, ou bien

$$H_0 : m \leqslant m_0 \qquad \text{contre} \qquad H_1 : m > m_0,$$

ou bien

$$H_0 : m \geqslant m_0 \qquad \text{contre} \qquad H_1 : m < m_0,$$

La forme de la statistique $\Lambda = f(X_1, \ldots, X_n)$ est, en général, déduite de celle du test bilatéral, et seules les régions d'acceptation et de rejet changent entre ces différents tests.

3. **Test du paramètre p d'une loi de Bernoulli**

Test bilatéral. — Soient $(X_1, \ldots, X_n)$ un échantillon de la loi $\mathcal{B}(1, p)$, $p \in]0, 1[$ inconnu, et $\alpha \in]0, 1[$ le seuil de test. Soit $p_0 \in]0, 1[$. On teste

$$H_0 : p = p_0 \qquad \text{contre} \qquad H_1 : p \neq p_0.$$

On introduit la statistique

$$\Lambda = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \times \frac{\sqrt{p(1-p)}}{\sqrt{p_0(1-p_0)}} + \frac{p - p_0}{\sqrt{p_0(1-p_0)/n}}$$

D'après le théorème central limite, lorsque n est grand ($n \geqslant 50$, $np \geqslant 10$, $n(1-p) \geqslant 10$), le premier facteur du premier terme est de loi proche de $\mathcal{N}(0, 1)$. Ainsi, si H_0 est fausse, Λ a tendance à prendre des valeurs éloignées de 0. Aussi est-on amené à définir la région d'acceptation de la forme

$$A = [-z_{1-\alpha/2}, z_{1-\alpha/2}],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale $\mathcal{N}(0, 1)$. La décision se fait par le calcul de la valeur observée de Λ :

$$\ell = \frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)/n}}\,.$$

Si $\ell \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$, on accepte H_0, sinon, on rejette H_0.

Remarques. — *a*) Le modèle statistique est donné par $E = \{0, 1\}$, $\Theta =]0, 1[$ et $\mu_\theta = \mathcal{B}(1, \theta)$. Ce test se fonde sur une propriété d'approximation, et, pour cette raison, c'est un test asymptotique. Le seuil n'est réellement respecté que pour $n = \infty$, c'est-à-dire que, pour $\theta = p_0$, $\mathbb{P}\{\Lambda_n^{p_0} \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\}$ converge vers $1 - \alpha$ quand n tend vers l'infini.

b) La variable aléatoire $n\bar{X}_n$ a pour loi la loi binomiale $\mathcal{B}(n, p)$. Le test suppose que les conditions d'approximation d'une telle loi par une loi normale sont satisfaites. Si ce n'est pas le cas, la situation peut devenir nettement plus complexe. On peut, en l'occurence, avoir recours à des approximations poissonniennes, voire à aucune approximation du tout, mais nous n'avons pas souhaité développer cette problématique.

c) La condition $\Lambda \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$ s'écrit aussi

$$(\bar{X}_n - p_0)^2 \leqslant z_{1-\alpha/2}^2 \, \frac{p_0(1-p_0)}{n},$$

c'est-à-dire

$$\big(1 + z_{1-\alpha/2}^2/n\big)p_0^2 - \big(2\bar{X}_n + z_{1-\alpha/2}^2/n\big)p_0 + \bar{X}_n^2 \leqslant 0.$$

À l'aide des racines du polynôme en p_0 ci-dessus :

$$P_\pm = \frac{2\bar{X}_n + z_{1-\alpha/2}^2/n \pm \sqrt{z_{1-\alpha/2}^4/n^2 + 4(z_{1-\alpha/2}^2/n)\bar{X}_n(1-\bar{X}_n)}}{2(1 + z_{1-\alpha/2}^2/n)},$$

on en déduit que cette condition équivaut à $p_0 \in [P_-, P_+]$. On trouve ainsi un intervalle de confiance approximatif $[P_-, P_+]$ pour le paramètre inconnu p. En négligeant les termes d'ordre strictement supérieur à $1/\sqrt{n}$, on a

$$[P_-, P_+] \approx \left[\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right]$$

et on retrouve l'estimation par intervalle de la section précédente. Ceci n'est pas fortuit, cette dernière correspondant dans le test de paramètre au choix de la statistique

$$\Lambda' = \frac{\bar{X}_n - p_0}{\sqrt{\bar{X}_n\big(1-\bar{X}_n\big)/n}}.$$

L'intervalle de confiance obtenu à l'aide de la statistique Λ apparaît, expérimentalement, meilleur que celui obtenu à l'aide de Λ', sa détermination nécessite seulement un peu plus de calculs et est plus difficile à mémoriser.

Tests unilatéraux. — Dans le cas du test

$$H_0 : p \geqslant p_0 \qquad \text{contre} \qquad H_1 : p < p_0,$$

on constate que si H_0 est fausse, la statistique Λ a tendance à prendre des valeurs éloignées de 0 par valeurs négatives. On est amené à définir la région d'acceptation de la forme

$$A = [-z_{1-\alpha}, +\infty[,$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale $\mathcal{N}(0,1)$. Si $\ell \in [-z_{1-\alpha}, +\infty[$, on accepte H_0, sinon, on rejette H_0.

Dans le cas du test

$$H_0 : p \leqslant p_0 \qquad \text{contre} \qquad H_1 : p > p_0,$$

on constate que si H_0 est fausse, la statistique Λ a tendance à prendre des valeurs éloignées de 0 par valeurs positives. On est amené à définir la région d'acceptation de la forme

$$A =]-\infty, z_{1-\alpha}],$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale $\mathcal{N}(0,1)$. Si $\ell \in]-\infty, z_{1-\alpha}]$, on accepte H_0, sinon, on rejette H_0.

Exemple. — Considérons la population des personnes atteintes d'un cancer des poumons. Un échantillon de taille 30 de tels personnes comprend une proportion égale à 0,8 de fumeurs. Cette proportion observée est-elle en accord avec l'affirmation selon laquelle 90 % des malades atteints d'un tel cancer sont fumeurs ?

Soit p la proportion des fumeurs dans cette population. Nous testons au seuil $\alpha = 0{,}05$

$$H_0 : p = 0{,}9 \qquad \text{contre} \qquad H_1 : p \neq 0{,}9.$$

La statistique de décision observée est

$$\ell = \frac{0{,}8 - 0{,}9}{\sqrt{0{,}9(1-0{,}9)/30}} \approx -1{,}824 \qquad (\ell' \approx -1{,}368)$$

qui est dans l'intervalle d'acceptation $[-1{,}96\,;\ 1{,}96]$ au seuil $\alpha = 0{,}05$. Avec ce seuil, nous sommes amené à conclure que les observations ne sont pas en contradiction significative avec l'hypothèse $p = 0{,}9$.

En revanche, si le seuil est $\alpha = 0{,}1$, alors l'intervalle d'acceptation devient $[-1{,}645\,;\ 1{,}645]$ et on conclut alors que les observations sont en contradiction significative avec l'hypothèse $p = 0{,}9$.

4. **Test de la moyenne d'une loi normale**

Soient $(X_1, \dots, X_n)$ un échantillon de la loi $\mathcal{N}(m, \sigma^2)$, $m \in \mathbb{R}$ inconnu, $\sigma \in \mathbb{R}_+^*$, et $\alpha \in \,]0,1[$ le seuil de test. Soit $m_0 \in \mathbb{R}$. On considère le test unilatéral

$$H_0 : m = m_0 \qquad \text{contre} \qquad H_1 : m \neq m_0.$$

Cas où σ est connu. — On introduit la statistique

$$\Lambda = \frac{\bar{X}_n - m_0}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} + \frac{m - m_0}{\sigma/\sqrt{n}}$$

Le premier terme a pour loi la loi $\mathcal{N}(0,1)$, ce qui résulte du fait que $\bar{X}_n$ a pour loi $\mathcal{N}(m, \sigma^2/n)$. Ainsi, si H_0 est fausse, Λ a tendance à prendre des valeurs éloignées de 0. Aussi définit-on la région d'acceptation par

$$A = [-z_{1-\alpha/2}, z_{1-\alpha/2}],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale $\mathcal{N}(0,1)$. La décision se fait par le calcul de la valeur observée de Λ :

$$\ell = \frac{\bar{x}_n - m_0}{\sigma/\sqrt{n}}\,.$$

Si $\ell \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$, on accepte H_0, sinon, on rejette H_0.

Remarque. — Le modèle statistique est donné par $E = \mathbb{R}$, $\Theta = \mathbb{R}$ et $\mu_\theta = \mathcal{N}(\theta, \sigma^2)$. Ce test ne fait appel à aucune approximation de lois. L'erreur de première espèce est par construction $\mathbb{P}\{\Lambda^{m_0} \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}]\} = \alpha$. Quant à l'erreur de seconde espèce, pour $\theta \neq m_0$,

$$\begin{aligned}
\beta_\theta &= \mathbb{P}\big\{\Lambda^\theta \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\big\} \\
&= \mathbb{P}\left\{\frac{\bar{X}_n^\theta - \theta}{\sigma/\sqrt{n}} \in \Big[-z_{1-\alpha/2} - \frac{\theta - m_0}{\sigma/\sqrt{n}}, z_{1-\alpha/2} - \frac{\theta - m_0}{\sigma/\sqrt{n}}\Big]\right\} \\
&= \Phi\Big(z_{1-\alpha/2} - \frac{\theta - m_0}{\sigma/\sqrt{n}}\Big) - \Phi\Big(-z_{1-\alpha/2} - \frac{\theta - m_0}{\sigma/\sqrt{n}}\Big)
\end{aligned}$$

dont le supremum vaut $1 - \alpha$. On constate qu'en choisissant α petit, c'est-à-dire en se fixant une erreur de première espèce faible, on accroît l'erreur de seconde espèce.

Exemple. — Le diamètre d'une certaine pièce manufacturée suit une loi normale dont la moyenne m n'est pas connue mais d'écart-type σ connu égal à 0,1 cm. Sur 100 pièces issues de la chaîne de fabrication, on a relevé un diamètre moyen de 10,02 cm. Le diamètre normal étant de 10 cm, peut-on penser qu'il y ait un défaut dans la chaîne de fabrication ?

Nous testons la moyenne d'une loi normale d'écart-type connu :

$$H_0 : m = 10\,\text{cm} \qquad \text{contre} \qquad H_1 : m \neq 10\,\text{cm}.$$

La statistique de décision observée est alors

$$\ell = \frac{10{,}02 - 10}{0{,}1/\sqrt{100}} = 2.$$

La région d'acceptation au seuil $\alpha = 0{,}05$ étant $[-1{,}96\,;\ 1{,}96]$ et puisque $\ell = 2$ n'y appartient pas, nous rejetons H_0 : les observations conduisent significativement à penser qu'il y a un défaut de fabrication.

Cas où σ est inconnu, test de Student. — On introduit la statistique

$$\Lambda = \frac{\bar{X}_n - m_0}{S_n/\sqrt{n-1}} = \frac{\bar{X}_n - m}{S_n/\sqrt{n-1}} + \frac{m - m_0}{S_n/\sqrt{n-1}},$$

ce qui revient à substituer dans la statistique précédente le réel inconnu σ par l'écart-type corrigé

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

qui en est un estimateur consistant et sans biais. Le premier terme a pour loi la loi de Student à $n-1$ degrés de liberté $\mathcal{T}(n-1)$. Ceci n'est pas un fait immédiat.

Ainsi, si H_0 est fausse, Λ a tendance à prendre des valeurs éloignées de 0. Aussi définit-on la région d'acceptation par

$$A = [-t_{1-\alpha/2}, t_{1-\alpha/2}],$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté. La décision se fait par le calcul de la valeur observée de Λ :

$$\ell = \frac{\bar{x}_n - m_0}{s_n/\sqrt{n-1}} = \frac{\bar{x}_n - m_0}{\hat{s}_n/\sqrt{n}}\,.$$

Si $\ell \in [-t_{1-\alpha/2}, t_{1-\alpha/2}]$, on accepte H_0, sinon, on rejette H_0.

Remarque. — Le modèle statistique est donné par $E = \mathbb{R}$, $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ et $\mu_\theta = \mathcal{N}(m_\theta, \sigma_\theta^2)$ où $\theta = (m_\theta, \sigma_\theta)$. Ce test ne fait appel à aucune approximation de lois. L'erreur de première espèce est par construction $\mathbb{P}\{\Lambda^\theta \notin [-t_{1-\alpha/2}, t_{1-\alpha/2}]\} = \alpha$ pour tout θ tel que $m_\theta = m_0$. Quant à l'erreur de seconde espèce, elle est calculée de la même manière que précédemment à l'aide de la fonction de répartition d'une loi de Student et son supremum est là encore $1-\alpha$. Pour les tests qui suivent nous laissons au lecteur (à la lectrice) le soin de préciser le modèle statistique et la valeur des erreurs correspondants.

Remarques. — *a*) Les tests unilatéraux correspondant aux deux tests précédents suivent le même principe qu'à la sous-section précédente.

b) Si on considère un test portant sur la moyenne d'une loi de probabilité quelconque, on supposera n grand et, via le théorème central limite, on mènera un test semblable aux deux tests précédents avec pour loi sous H_0 de la statistique Λ correspondante la loi $\mathcal{N}(0,1)$.

5. **Test comparatif de deux moyennes de lois normales**

Soient $(X_1, \ldots, X_n)$ un échantillon de la loi $\mathcal{N}(m_X, \sigma^2)$, $(Y_1, \ldots, Y_m)$ un échantillon de la loi $\mathcal{N}(m_Y, \sigma^2)$, m_X et $m_Y \in \mathbb{R}$ inconnus, $\sigma = \sigma_X = \sigma_Y \in \mathbb{R}_+^*$, et $\alpha \in\,]0,1[$ le seuil de test. On considère le test unilatéral

$$H_0 : m_X = m_Y \qquad \text{contre} \qquad H_1 : m_X \neq m_Y.$$

Cas où σ est connu. — On introduit la statistique

$$\Lambda = \frac{\bar{X}_n - \bar{Y}_m}{\sigma} \times \sqrt{\frac{n \times m}{n+m}}$$

Si $m_X = m_Y$, on montre facilement que la statistique a pour loi la loi $\mathcal{N}(0,1)$. Ainsi, si H_0 est fausse, Λ a tendance à prendre des valeurs éloignées de 0. Aussi est-on amené à définir la région d'acceptation de la forme

$$A = [-z_{1-\alpha/2}, z_{1-\alpha/2}],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$. La décision se fait par le calcul de la valeur observée de Λ :

$$\ell = \frac{\bar{x}_n - \bar{y}_m}{\sigma} \times \sqrt{\frac{n \times m}{n+m}}\,.$$

Si $\ell \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$, on accepte H_0, sinon, on rejette H_0.

Cas où σ est inconnu. — On introduit la statistique

$$\Lambda = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{nS_X^2 + mS_Y^2}} \times \sqrt{\frac{n \times m \times (n+m-2)}{n+m}}$$

Si $m_X = m_Y$, on montre que cette statistique a pour loi la loi de Student à $n+m-2$ degrés de liberté $\mathcal{T}(n+m-2)$. Ceci n'est pas un fait immédiat et fait l'objet d'une démonstration à part entière.

Ainsi, si H_0 est fausse, Λ a tendance à prendre des valeurs éloignées de 0. Aussi est-on amené à définir la région d'acceptation de la forme

$$A = [-t_{1-\alpha/2}, t_{1-\alpha/2}]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n+m-2$ degrés de liberté. La décision se fait par le calcul de la valeur observée de Λ :

$$\ell = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{ns_x^2 + ms_y^2}} \times \sqrt{\frac{n \times m \times (n+m-2)}{n+m}}\,.$$

Si $\ell \in [-t_{1-\alpha/2}, t_{1-\alpha/2}]$, on accepte H_0, sinon, on rejette H_0.

Remarques. — *a*) Il n'y a rien de nouveau quant aux tests unilatéraux. Il est à noter que nous avons supposé $\sigma_X = \sigma_Y$ dans les deux cas. Des tests plus généraux existent, mais sont bien plus complexes.

b) Ces tests de comparaison de moyennes de lois normales de même variance inconnue sont souvent appelés *tests de Student* et leur statistique est souvent notée T ou T_{m+n-2}.

Exercice 1. — Le rédacteur d'une page web portant sur les araignées vivant sur le continent américain affirme que la taille moyenne du corps des femelles *latrodecti hesperi* est égale à 1 pouce 1/4. En tenant compte des informations données dans les exercices et exemples précédents, tester cette affirmation aux seuils 0,1, 0,05 et 0,01. Remarquer le rôle de la valeur du seuil pour le test, et, de manière complémentaire, du niveau de confiance. Proposer des explications de son erreur manifeste.

ANNEXES

A. **Intervalles de confiance d'une proportion**

Soit $p \in [0,1]$ un réel représentant une probabilité ou une proportion d'individus dans une population. Le caractère correspondant est une variable statistique X de loi de Bernoulli de paramètre p qui a pour moyenne $\mathbb{E}[X] = p$ et variance $\mathrm{Var}(X) = p(1-p)$.

Soit $(X_1, \ldots, X_n)$ un échantillon de la loi de Bernoulli de paramètre p. La statistique $\bar{X}_n$ est un estimateur (consistant et sans biais) de p et une valeur observée $\bar{x}_n = k/n$ est une estimation (ponctuelle) de la proportion p. Rappelons que dans ce contexte on a $S_n^2 = \bar{X}_n(1-\bar{X}_n)$ et donc que la connaissance de la variance observée n'apporte rien à celle de la proportion observée.

Pour faire une estimation par intervalle de la proportion p, on se donne un échantillon observé $(x_1, \ldots, x_n)$, on note k le nombre d'apparitions du nombre 1 et $n-k$ celui de 0, l'estimation ponctuelle de p obtenue étant $\bar{x}_n = k/n$. On distingue plusieurs situations vis-à-vis du calcul numérique.

1. La taille de l'échantillon n est grande, ainsi que $k = n\bar{x}_n$ et $n-k = n(1-\bar{x}_n)$:

$$n \geqslant 50, \qquad k \geqslant 10 \qquad \text{et} \qquad n-k \geqslant 10\,;$$

on déterminera un intervalle de confiance asymptotique de p (*voir* la sous-section *deux approches asymptotiques par le théorème central limite*).

2.a. La taille de l'échantillon n est grande, mais $k = n\bar{x}_n > 0$ est petit :

$$n \geqslant 50, \qquad \text{et} \qquad 0 < k < 10\,;$$

on déterminera un intervalle de confiance asymptotique de p (*voir* la sous-section *une approche asymptotique par la loi des événements rares*).

2.b. La taille de l'échantillon n est grande, mais $n-k = n(1-\bar{x}_n) > 0$ est petit :

$$n \geqslant 50, \qquad \text{et} \qquad 0 < n-k < 10\,;$$

la situation est symétrique du 2.a et lui est donc semblable.

3. La taille de l'échantillon n est petite et la proportion non triviale :

$$n < 50 \qquad \text{et} \qquad 0 < k < n\,;$$

on déterminera un intervalle de confiance approximatif (dit exact) (*voir* la sous-section *une approche exacte*).

4. La proportion observée est triviale :

$$k = 0 \qquad \text{ou} \qquad k = n\,;$$

on déterminera un intervalle de confiance approximatif (dit exact) (*voir* la remarque de la sous-section *une approche exacte*).

Les valeurs seuil 50 et 10 sont données à titre indicatif et sont utilisées généralement avec $\alpha = 0{,}05$. D'autres valeurs seuil peuvent être proposées, en particulier pour d'autres valeurs de α.

Remarque. — Si $p \in \{0,1\}$, alors quelque soit l'échantillon obtenu, la proportion observée sera triviale. Et réciproquement, une proportion observée non triviale implique que $p \neq 0$ et $p \neq 1$.

A.1. Deux approches asymptotiques par le théorème central limite

Une approche (un peu trop) rapide. — Pour $p \in]0,1[$, le théorème central limite affirme que

$$\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$$

est de loi proche de la loi normale $\mathcal{N}(0,1)$ lorsque n est grand. On a en particulier

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leqslant \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \leqslant z_{1-\alpha/2}\right\} \approx 1-\alpha \qquad (E_\alpha)$$

où $z_{1-\alpha/2} = -z_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$. Puisque $\bar{X}_n$ approche p, on devrait avoir aussi

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leqslant \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n}} \leqslant z_{1-\alpha/2}\right\} \approx 1-\alpha$$

ce qui s'écrit

$$\mathbb{P}\left\{\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leqslant p \leqslant \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right\} \approx 1-\alpha$$

D'où l'intervalle de confiance asymptotique usuel d'une proportion :

$$\left[\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\,,\, \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]. \qquad (L_\alpha)$$

Pratique. — On détermine $\bar{x}_n$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — On doit garder à l'esprit qu'on ne peut utiliser ce type d'intervalle qu'avec des conditions sur la taille n de l'échantillon et aussi sur la valeur de la proportion observée. Pour se faire une idée, majorons $x(1-x)$ par $1/4$ pour $x \in [0,1]$. L'intervalle de confiance est d'amplitude majorée par, voire comparable à, $z_{1-\alpha/2}/\sqrt{n} \approx 1{,}96/\sqrt{n}$ pour $\alpha = 0{,}05$. Ainsi, pour n trop petit, il est fréquent que l'intervalle sorte de $[0,1]$ — ce qui n'a pas de sens puisque nous estimons une proportion —, en particulier si la proportion observée est proche de 0 ou de 1.

Une approche plus prudente. — Revenons à (E_α) qui s'écrit aussi

$$\mathbb{P}\left\{\frac{\big(\bar{X}_n - p\big)^2}{p(1-p)/n} \leqslant z_{1-\alpha/2}^2\right\} \approx 1-\alpha$$

où encore

$$\mathbb{P}\Big\{\big(1 + z_{1-\alpha/2}^2/n\big)p^2 - \big(2\bar{X}_n + z_{1-\alpha/2}^2/n\big)p + \bar{X}_n^2 \leqslant 0\Big\} \approx 1-\alpha.$$

On voit apparaître un polynôme du second degré en la variable p qui est de coefficient dominant strictement positif. Ainsi, il est négatif lorsque p est entre ses racines :

$$P_{\min/\max} = \frac{\bar{X}_n + z_{1-\alpha/2}^2/2n \mp \sqrt{z_{1-\alpha/2}^2/n \times \big(z_{1-\alpha/2}^2/4n + \bar{X}_n(1-\bar{X}_n)\big)}}{1 + z_{1-\alpha/2}^2/n} \qquad (W_\alpha)$$

qui définissent donc l'intervalle de confiance asymptotique cherché. Celui-ci est meilleur que le précédent si on se fonde uniquement sur l'approximation du théorème central limite. Il présente de plus le grand avantage de fournir des bornes qui sont toujours dans $[0,1]$ contrairement à ce qu'il se passe pour l'approche rapide.

Pratique. — On détermine $\bar{x}_n$ et $z_{1-\alpha/2}$ et en déduit l'intervalle de confiance.

Remarque. — En négligeant les termes d'ordre strictement supérieur à $1/\sqrt{n}$, on retrouve l'intervalle de confiance asymptotique usuel.

A.2. Une approche asymptotique par la loi des événements rares

Grossièrement, la loi des événements rares permet de dire que pour p petit et n grand, la loi binomiale de paramètres n et p est proche de la loi de Poisson de paramètre $\lambda = np$. Plus exactement, si $(p_n)_{n\geqslant 1}$ est une suite dans $[0,1]$ telle que np_n tend vers un réel $\lambda \geqslant 0$, alors les lois binomiales $(\mathcal{B}(n,p_n))_{n\geqslant 1}$ convergent vers la loi de Poisson de paramètre λ.

Supposons n grand et avoir observé $0 < k \ll n$. Puisque $k/n \ll 1$ estime p, on est dans le cadre de la loi des événements rares. La loi de la variable $n\bar{X}_n = X_1 + \cdots + X_n$, qui est la loi binomiale de paramètre n et p, est proche de la loi de Poisson de paramètre $\lambda = np$.

On détermine un intervalle de confiance dit exact $[\lambda_{\min}(k,\alpha), \lambda_{\max}(k,\alpha)]$ du paramètre d'une loi de Poisson pour une observation unique égale à k (*voir* la section *intervalle de confiance du paramètre d'une loi de Poisson*). On prendra alors pour intervalle de confiance asymptotique de la proportion p :

$$[p_{\min}, p_{\max}] = \Big[\frac{\lambda_{\min}(k,\alpha)}{n}\,,\,\frac{\lambda_{\max}(k,\alpha)}{n}\Big]. \qquad (ER_{\mathrm{a},\alpha})$$

Supposons n grand et avoir observé $0 < n-k \ll n$. Dans ce cas, l'application de la loi des événements rares consiste à dire que la loi de la variable $n - n\bar{X}_n$, qui est la loi binomiale de paramètres n et $1-p$, est proche de la loi de Poisson de paramètre $\lambda = n(1-p)$. On obtient pour intervalle de confiance asymptotique de la proportion $1-p$:

$$[1-p_{\max}, 1-p_{\min}] = \Big[\frac{\lambda_{\min}(n-k,\alpha)}{n}\,,\,\frac{\lambda_{\max}(n-k,\alpha)}{n}\Big].$$

Soit, pour la proportion p :

$$[p_{\min}, p_{\max}] = \Big[1 - \frac{\lambda_{\max}(n-k,\alpha)}{n}\,,\,1 - \frac{\lambda_{\min}(n-k,\alpha)}{n}\Big]. \qquad (ER_{b,\alpha})$$

Pratique. — On dispose d'un échantillon observé $(x_1,\ldots,x_n)$. Si n est grand et $k > 0$ (resp. $n-k > 0$) est petit, on détermine à l'aide d'une table les bornes de l'intervalle de confiance du paramètre d'une loi de Poisson associé à l'observation k (resp. $n-k$). En divisant les bornes par n, on obtient l'intervalle de confiance de p (resp. de $1-p$).

Remarques. — *a*) Même si on dispose d'outils numériques pour la détermination dite exacte de l'intervalle de confiance d'une proportion, l'approche par les événements rares demeure pertinente quand n est grand (et k, ou $n-k$, petit), même si elle est faite sur ordinateur.

b) Le lien numérique direct entre l'approche exacte et celle par la loi des événements rares est qu'à x et k fixés, quand $n \to \infty$,

$$P_{n,k}(x/n) \longrightarrow P_k(x) = 1 - \mathrm{e}^{-x}\sum_{\ell=0}^{k-1}\frac{x^\ell}{\ell\,!} \qquad \text{et} \qquad Q_{n,k}(x/n) \longrightarrow Q_k(x) = \mathrm{e}^{-x}\sum_{\ell=0}^{k}\frac{x^\ell}{\ell\,!},$$

polynômes intervenant dans la détermination des bornes des intervalles de confiance dits exacts d'une proportion et du paramètre d'une loi de Poisson (*voir* plus loin).

c) Supposons que la proportion observée soit triviale, $k = 0$ par exemple. L'intervalle de confiance de λ pour $k = 0$ est $[0, -\ln(\alpha/2)]$ (ou $[0, -\ln\alpha]$). On en déduit l'intervalle de confiance asymptotique $[0, -\ln(\alpha/2)/n]$ pour p. Or, l'approche exacte donne pour intervalle de confiance $[0, 1-(\alpha/2)^{1/n}]$ (*voir* plus loin). Mais, puisque n est grand,

$$1-(\alpha/2)^{1/n} = 1-\exp\left(\frac{\ln(\alpha/2)}{n}\right) = 1-\left(1+\frac{\ln(\alpha/2)}{n}+o(1/n)\right) = -\frac{\ln(\alpha/2)}{n}+o(1/n).$$

Ainsi, dans le cas d'une proportion observée triviale, l'approche par la loi des événements rares n'apporte rien de nouveau.

A.3. Une approche exacte

Nous recherchons un intervalle de confiance exact, c'est-à-dire des bornes $P_{\min/\max} = \phi_{\min/\max}(X_1, \ldots, X_n)$ telles que $\mathbb{P}\{P_{\min} \leqslant p \leqslant P_{\max}\} = 1-\alpha$, mais n'obtiendrons qu'un intervalle de confiance approximatif (cela vient de la nature même du problème). La détermination que nous présentons est celle de Clopper–Pearson qui est la plus commune.

Nous ne connaissons pas p mais observons k « réussites » sur n essais.

La probabilité qu'une variable aléatoire de loi $\mathcal{B}(n, x)$ prenne des valeurs aussi grandes que k est

$$P_{n,k}(x) = \sum_{\ell=k}^{n} C_n^\ell\, x^\ell (1-x)^{n-\ell} = 1-\sum_{\ell=0}^{k-1} C_n^\ell\, x^\ell (1-x)^{n-\ell}.$$

Lorsque $k \geqslant 1$, cette fonction polynomiale, positive sur $[0,1]$, est nulle en $x = 0$ et on a presque immédiatement

$$P'_{n,k}(x) = \sum_{\ell=k}^{n} (\ell - nx) C_n^\ell\, x^{\ell-1}(1-x)^{n-\ell-1}$$

et on constate que pour tous $x \in [0, k/n]$, $\ell \geqslant k$, on a $\ell - nx \geqslant 0$, et qu'ainsi $P'_{n,x}$ est positive sur $[0, k/n]$ et donc que $P_{n,k}$ est croissante sur $[0, k/n]$, et même strictement croissante sur $]0, k/n[$, et on a grossièrement $P_{n,k}(k/n) \approx 1/2$. Pour α assez petit, il existe alors un unique $p_{\min} = p_{\min}(k, n, \alpha) \in [0, k/n]$ tel que $P_{n,k}(p_{\min}) = \alpha/2$.

Le choix de la borne inférieure de l'intervalle de confiance se justifie ainsi : puisque nous avons observé k, la probabilité d'observer des valeurs aussi grandes que k doit être important, c'est-à-dire $P_{n,k}(p) \geqslant \alpha/2$, ou encore $p \geqslant p_{\min}$. La borne inférieure $p_{\min}$ est bien une fonction des observations $k = x_1 + \cdots + x_n$ (et aussi de la taille n de l'échantillon et du seuil α).

Pour la borne supérieure, on procède de même, ou bien en considérant les probabilités d'observer des valeurs aussi petites que k lorsque $k < n$, et la fonction polynomiale

$$Q_{n,k}(x) = \sum_{\ell=0}^{k} C_n^\ell\, x^\ell (1-x)^{n-\ell} = 1-\sum_{\ell=k+1}^{n} C_n^\ell\, x^\ell (1-x)^{n-\ell} = 1-P_{n,k+1}(x),$$

ou bien en passant formellement au complémentaire succès/échec. En résumé :

$$\begin{cases} P_{n,k}(p_{\min}) = \alpha/2 & \text{avec } p_{\min} = p_{\min}(k,n,\alpha) \in [0, k/n], \\ Q_{n,k}(p_{\max}) = \alpha/2 & \text{avec } p_{\max} = p_{\max}(k,n,\alpha) \in [k/n, 1]. \end{cases} \qquad (CP_\alpha)$$

Le problème de détermination des bornes est numérique : il faut inverser un polynôme dont les valeurs sont déjà lourdes à calculer.

Remarque (cas d'une proportion observée triviale). — Lorsque la proportion observée $\bar{x}_n$ est égale à 0 ou 1, la seule approche possible ou souhaitée est une approche (dite) exacte, qui, heureusement, est alors très simple à mettre en œuvre. Si $x_n = 0$, donc si $k = 0$, on veut obtenir un intervalle de la forme $[0, p_{\max}]$ avec $Q_{n,0}(p_{\max}) = (1 - p_{\max})^n = \alpha/2$, soit $p_{\max} = 1 - (\alpha/2)^{1/n}$. De même, si $\bar{x}_n = 1$, donc si $k = n$, on veut obtenir un intervalle de la forme $[p_{\min}, 1]$ avec $P_{n,n}(p_{\min}) = (p_{\min})^n = \alpha/2$, soit $p_{\min} = (\alpha/2)^{1/n}$. Donc

$$\begin{array}{ll} p_{\min}(0, n, \alpha) = 0, & p_{\max}(0, n, \alpha) = 1 - (\alpha/2)^{1/n} \\ p_{\min}(n, n, \alpha) = (\alpha/2)^{1/n}, & p_{\max}(n, n, \alpha) = 1. \end{array} \qquad (0{-}1_\alpha)$$

On pourra cependant préférer dans ces cas des intervalles de confiance unilatéraux en remplaçant $\alpha/2$ par α.

Nous pouvons préciser certaines propriétés qui ne sont peut-être pas immédiates.

PROPOSITION. — *Pour tout* $\alpha \in]0, 1[$ *et* $0 \leqslant k \leqslant n$ *entier,* $p_{\min}(k, n, \alpha) < p_{\max}(k, n, \alpha)$.

Démonstration. — La démonstration est identique ou presque à celle qui lui correspond à la section *intervalle de confiance du paramètre d'une loi de Poisson.* □

Soit X une variable aléatoire de loi binomiale de paramètres n et p, et notons $P_{\min} = p_{\min}(X, n, \alpha)$ et $P_{\max} = p_{\max}(X, n, \alpha)$.

THÉORÈME. — *L'intervalle* $[P_{\min}, P_{\max}]$ *est un intervalle de confiance au niveau* $1 - \alpha$ *du paramètre* p. *Plus précisément,* $\mathbb{P}\{p \in [P_{\min}, P_{\max}]\} > 1 - \alpha$.

Démonstration. — La démonstration est identique ou presque à celle qui lui correspond à la section *intervalle de confiance du paramètre d'une loi de Poisson.* □

Pratique. — On détermine $\bar{x}_n = k/n$. S'il n'est pas possible de recourir à une détermination asymptotique (en particulier si $n \leqslant 20$), on se reporte à une table spécifique, à une abaque, ou on réalise le calcul des bornes à l'aide d'un logiciel adapté.

Remarque. — À tout intervalle de confiance $[P_{\min}, P_{\max}]$ construit à partir de X de loi $\mathcal{B}(n, p)$, est associé un sous-ensemble I de $\{0, 1, \ldots, n\}$ tel qu'on ait l'égalité d'événements $\{p \in [P_{\min}, P_{\max}]\} = \{X \in I\}$ (car le premier événement est $\sigma(X)$-mesurable). Ainsi,

$$\mathbb{P}\{p \in [P_{\min}, P_{\max}]\} = \mathbb{P}\{X \in I\} = \sum_{\ell \in I} C_n^\ell \, p^\ell (1 - p)^{n-\ell}.$$

Il n'y a qu'un nombre fini de choix possibles pour I et un nombre plus restreint encore de choix tels que la probabilité précédente soit supérieure ou égale à $1 - \alpha$. On ne doit pas s'attendre à pouvoir trouver un sous-ensemble I tel que l'égalité soit réalisée, et même si on trouvait un ensemble I optimal pour approcher l'égalité, il ne lui correspondrait peut-être pas un intervalle de confiance.

B. **Intervalle de confiance du paramètre d'une loi de Poisson**

Nous considérons une variable aléatoire suivant une loi de Poisson de paramètre $\lambda \geqslant 0$. On observe une valeur entière $k \geqslant 0$ et on souhaite en déduire un intervalle de confiance de λ (sur la base d'une unique observation !). Notre approche suit l'esprit de la détermination de Clopper–Pearson pour l'intervalle de confiance d'une proportion. Elle fournira aussi un intervalle de confiance dit exact.

La probabilité qu'une variable aléatoire de loi de Poisson de paramètre $x \geqslant 0$ prenne des valeurs aussi grandes que $k > 0$ — le cas $k = 0$ sera examiné en remarque plus loin — est

$$P_k(x) = \mathrm{e}^{-x} \sum_{\ell \geqslant k} \frac{x^\ell}{\ell\,!} = 1 - \mathrm{e}^{-x} \sum_{\ell=0}^{k-1} \frac{x^\ell}{\ell\,!}\,.$$

C'est une fonction régulière de la variable x, nulle en 0 et tendant vers 1 en $+\infty$, et on a

$$P_k'(x) = -\,\mathrm{e}^{-x} \sum_{\ell \geqslant k} \frac{x^\ell}{\ell\,!} + \mathrm{e}^{-x} \sum_{\ell \geqslant k} \frac{\ell x^{\ell-1}}{\ell\,!} = \mathrm{e}^{-x} \frac{x^{k-1}}{(k-1)\,!}$$

qui strictement positive pour $x > 0$. Ainsi, il existe un unique $\lambda_{\min} = \lambda_{\min}(k, \alpha) \in \,]0, \infty[$ tel que $P_k(\lambda_{\min}) = \alpha/2$. Symétriquement, la probabilité qu'une variable aléatoire de loi de Poisson de paramètre $x \geqslant 0$ prenne des valeurs aussi petites que $k > 0$ est

$$Q_k(x) = \mathrm{e}^{-x} \sum_{\ell=0}^{k} \frac{x^\ell}{\ell\,!} = 1 - \mathrm{e}^{-x} \sum_{\ell \geqslant k+1} \frac{x^\ell}{\ell\,!} = 1 - P_{k+1}(x).$$

C'est une fonction continue et strictement décroissante qui vaut 1 en 0 et tend vers 0 en $+\infty$. Ainsi, il existe un unique $\lambda_{\max} = \lambda_{\max}(k, \alpha) \in \,]0, \infty[$ tel que $Q_k(\lambda_{\max}) = \alpha/2$. L'intervalle de confiance de λ pour une observation $k > 0$ est alors $[\lambda_{\min}, \lambda_{\max}]$ donné par

$$\lambda_{\min} = \lambda_{\min}(k, \alpha) = P_k^{-1}(\alpha/2) \qquad \text{et} \qquad \lambda_{\max} = \lambda_{\max}(k, \alpha) = Q_k^{-1}(\alpha/2). \qquad (P_\alpha)$$

Remarque. — Pour $k = 0$, l'intervalle de confiance (bilatéral) est alors de la forme $[0, \lambda_{\max}]$ avec $Q_0(\lambda_{\max}) = \mathrm{e}^{-\lambda_{\max}} = \alpha/2$, autrement dit $[0, -\ln(\alpha/2)]$ ou encore

$$\lambda_{\min}(0, \alpha) = 0 \qquad \text{et} \qquad \lambda_{\max}(0, \alpha) = -\ln(\alpha/2).$$

Notons au passage que si α est voisin de 0, on autorise presque toutes les valeurs possibles pour λ, alors que si α est proche de 1 on se restreint à λ voisin de 0. Par ailleurs, pour $k = 0$, il semble plus pertinent de considérer l'intervalle de confiance unilatéral, autrement dit de remplacer $\alpha/2$ par α ci-dessus.

Nous pouvons préciser certaines propriétés qui ne sont peut-être pas immédiates.

Proposition. — *Pour tout* $\alpha \in \,]0, 1[$ *et* $k > 0$ *entier,* $\lambda_{\min}(k, \alpha) < \lambda_{\max}(k, \alpha)$.

Démonstration. — Puisque $P_k(x) \geqslant P_{k+1}(x)$ pour tout $x \geqslant 0$, on a $\lambda_{\min}(k, \alpha) \leqslant \lambda_{\min}(k + 1, \alpha)$. Puisque Q_k est une fonction décroissante, $Q_k(\lambda_{\min}(k, \alpha)) \geqslant Q_k(\lambda_{\min}(k + 1, \alpha)) = 1 - P_{k+1}(\lambda_{\min}(k + 1, \alpha)) = 1 - \alpha/2 > \alpha/2$. Puisque $Q_k(\lambda_{\min}(k, \alpha)) > \alpha/2$, $Q_k(\lambda_{\max}(k, \alpha)) = \alpha/2$ et Q_k est décroissante, on a $\lambda_{\min}(k, \alpha) < \lambda_{\max}(k, \alpha)$. □

Soit X une variable aléatoire de loi de Poisson de paramètre $\lambda \geqslant 0$, et notons $\Lambda_{\min} = \lambda_{\min}(X, \alpha)$ et $\Lambda_{\max} = \lambda_{\max}(X, \alpha)$.

Théorème. — *L'intervalle* $[\Lambda_{\min}, \Lambda_{\max}]$ *est un intervalle de confiance au niveau* $1 - \alpha$ *du paramètre* λ. *Plus précisément,* $\mathbb{P}\big\{\lambda \in [\Lambda_{\min}, \Lambda_{\max}]\big\} > 1 - \alpha$.

Démonstration. — Nous allons regarder ce qu'il se passe pour la borne inférieure. Pour $k > 0$ donné, comme P_k est une fonction strictement croissante de la variable $x \geqslant 0$, nous avons : $\lambda_{\min}(k, \alpha) \leqslant \lambda \iff P_k(\lambda_{\min}(k, \alpha)) \leqslant P_k(\lambda) \iff \alpha/2 \leqslant P_k(\lambda)$ par définition de $\lambda_{\min}(k, \alpha)$, en notant que ceci reste vrai pour $k = 0$. Alors,

$$\mathbb{P}\{\Lambda_{\min} \leqslant \lambda\} = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k\,!} \times \mathbb{1}_{\lambda_{\min}(k,\alpha) \leqslant \lambda} = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k\,!} \times \mathbb{1}_{\alpha/2 \leqslant P_k(\lambda)}.$$

Puisque pour $\lambda \geqslant 0$ fixé, la suite $(P_k(\lambda))_{k\geqslant 0}$ de valeur initiale 1 décroît vers 0, soit $\kappa(\lambda,\alpha) = \max\{k : P_k(\lambda) \geqslant \alpha/2\}$. On a alors

$$\mathbb{P}\{\Lambda_{\min} \leqslant \lambda\} = \mathrm{e}^{-\lambda} \sum_{k=0}^{\kappa(\lambda,\alpha)} \frac{\lambda^k}{k\,!} = 1 - P_{\kappa(\lambda,\alpha)+1}(\lambda) > 1 - \alpha/2$$

puisque $P_{\kappa(\lambda)+1,\alpha}(\lambda) < \alpha/2$. De la même manière, nous avons pour la borne supérieure

$$\mathbb{P}\{\Lambda_{\max} \geqslant \lambda\} > 1 - \alpha/2.$$

Comme $\lambda_{\min}(k,\alpha) < \lambda_{\max}(k,\alpha)$ pour tout $k \geqslant 0$, on a toujours $\Lambda_{\min} < \Lambda_{\max}$. Alors

$$\begin{aligned}\mathbb{P}\big\{\Lambda_{\min} \leqslant \lambda \leqslant \Lambda_{\max}\big\} &= 1 - \mathbb{P}\big(\{\Lambda_{\min} > \lambda\} \cup \{\lambda > \Lambda_{\max}\}\big)\\ &= 1 - \mathbb{P}\{\Lambda_{\min} > \lambda\} - \mathbb{P}\{\lambda > \Lambda_{\max}\}\\ &= \mathbb{P}\{\Lambda_{\min} \leqslant \lambda\} + \mathbb{P}\{\Lambda_{\max} \geqslant \lambda\} - 1 > 1 - \alpha.\end{aligned}$$

D'où la conclusion. □

Supposons maintenant que $(X_1,\dots,X_n)$ soit un échantillon de la loi de Poisson de paramètre $\lambda \geqslant 0$ et $(k_1,\dots,k_n)$ un échantillon observé. La somme $X = X_1 + \cdots + X_n$ a pour loi la loi de Poisson de paramètre $n\lambda$ puisque les variables aléatoires X_i, $i = 1,\dots,n$, sont indépendantes et toutes de loi de Poisson de paramètre λ. La valeur observée de X est $k = k_1 + \cdots + k_n$. Avec ce qui précède, nous obtenons un intervalle de confiance $[\lambda_{\min}(k,\alpha), \lambda_{\max}(k,\alpha)]$ de $n\lambda$. L'intervalle de confiance de λ est donc

$$\Big[\frac{\lambda_{\min}(k_1+\cdots+k_n,\alpha)}{n}\,,\,\frac{\lambda_{\max}(k_1+\cdots+k_n,\alpha)}{n}\Big].$$

Ceci n'est pas totalement étonnant puisque $\bar{x}_n = k/n$ est l'estimation ponctuelle habituelle de λ et que les lois de Poisson forme un semi-groupe à 1 paramètre seulement.

Pratique. — Si $(k_1,\dots,k_n)$ est un échantillon observé d'une loi de Poisson de paramètre λ, on calcule $k = k_1 + \cdots + k_n$, détermine $\lambda_{\min}(k,\alpha)$ et $\lambda_{\max}(k,\alpha)$ à l'aide d'une table ou d'un logiciel approprié, et en déduit l'intervalle de confiance de λ.

Remarque. — Tout comme pour l'intervalle de confiance dit exact d'une proportion, avoir l'égalité $\mathbb{P}\{\Lambda_{\min} \leqslant \lambda \leqslant \Lambda_{\max}\} = 1 - \alpha$ ne semble pas réalisable.

C. **Programmation**

En Scilab, on définit une fonction `proportionCLI`, où `CLI` fait référence aux expressions *confidence limits/interval*, ainsi qu'une fonction `poissonCLI`. Pour mettre en œuvre la méthode de Clopper–Wilson, nous nous sommes servi de l'implémentation du calcul des fonctions de répartition des lois binomiales par Scilab, alors que pour le cas poissonnien, nous avons tout programmé « à la main ».

```
accuracy = 1E-12;

function [pmin, pmax] = proportionCLI(k, n, alpha);
  if k == 0 then pmin = 0; pmax = 1-(alpha/2)^(1/n);
  elseif k == n then pmin = (alpha/2)^(1/n); pmax = 1;
  elseif n < 50 then
    global accuracy;
    // Clopper-Pearson
```

```
    // local p, lb, ub;
    lb = 0; ub = k/n;
    while ub-lb > accuracy;
      p = .5*(lb+ub);
      if 1-cdfbin("PQ", k-1, n, p, 1-p) < alpha/2 then lb = p;
      else ub = p;
      end
    end
    pmin = 0.5*(lb+ub);
    lb = k/n; ub = 1;
    while ub-lb > accuracy;
      p = 0.5*(lb+ub);
      if cdfbin("PQ", k, n, p, 1-p) < alpha/2 then ub = p;
      else lb = p;
      end
    end
    pmax = 0.5*(lb+ub);
  elseif k < 10 then [pmin, pmax] = poissonCLI(k, alpha)/n;
  elseif n-k < 10 then [p, q] = poissonCLI(n-k, alpha)/n;
    pmin = 1-q; pmax = 1-p;
  else
    // Wilson
    // local p, z, t;
    p = k/n;
    z = cdfnor("X", 0, 1, 1-alpha/2, alpha/2);
    z = z*z/n; t = sqrt(z*(z/4+p*(1-p)));
    pmin = (p+z/2-t)/(1+z); pmax = (p+z/2+t)/(1+z);
  end
endfunction

function p = poissoncdf(x, lambda);
  if x < 0 then p = 0;
  elseif lambda > 0 then
    // local k;
    p = 1.0;
    for k = floor(x):-1:1; p = 1+p/k*lambda; end
    p = p*exp(-lambda);
  else p = 1;
  end
endfunction

function [lmin, lmax] = poissonCLI(k, alpha);
  if k == 0 then lmin = 0; lmax = -log(alpha/2);
  else
    global accuracy;
    // local lb, ub;
    // lowerbound
    lb = 0.0; ub = 1.0*k;
    while ub-lb > accuracy;
      if 1-poissoncdf(k-1, (lb+ub)/2) < alpha/2 then lb = (lb+ub)/2;
```

```
      else ub = (lb+ub)/2;
      end
    end
    lmin = (lb+ub)/2;
    // upperbound
    lb = 0.0; ub = 1.0*k;
    while poissoncdf(k, ub) > alpha/2; ub = 2*ub; end
    while ub-lb > accuracy
      if poissoncdf(k, (lb+ub)/2) > alpha/2 then lb = (lb+ub)/2;
      else ub = (lb+ub)/2;
      end
    end
    lmax = (lb+ub)/2;
  end
endfunction
```

TABLE DES MATIÈRES

Chapitre premier. Généralités 1
1. Cadre 1
2. Définitions 1
3. Regroupements 4
4. Exercices 5

Chapitre II. Statistiques et paramètres usuels 9
1. Statistiques et paramètres de position usuels 9
2. Statistiques et paramètres de dispersion usuels 12
3. Exercices 13

Chapitre III. Représentations graphiques 14
1. Variables discrètes 14
2. Variables quantitatives 14
3. Diagrammes en boîtes 16
Exercices 18

Chapitre IV. Analyse statistique multivariée, un exemple : les régressions linéaires 19
1. Introduction 19
2. Notations 19
3. Régression linéaire de y en x 20
4. Généralisations de la régression linéaire 22
5. Fausses généralisations de la régression linéaire 23
6. Régression orthogonale 23
Exercices 25

Chapitre V. Estimation et intervalles de confiance 27
1. Estimation ponctuelle 27
2. Estimation par intervalle 28
3. Estimation d'une proportion 29
4. Estimation de la loi d'une variable discrète 31
5. Estimation de la moyenne d'une loi normale 31
6. Estimation de la moyenne d'une loi de probabilité 32

Chapitre VI. Tests d'hypothèses 34
1. Généralités 34
2. Généralités sur les tests paramétriques 35
3. Test du paramètre p d'une loi de Bernoulli 36
4. Test de la moyenne d'une loi normale 38
5. Test comparatif de deux moyennes de lois normales 39

Annexes 41
A. Intervalles de confiance d'une proportion 41
A.1. *Deux approches asymptotiques par le théorème central limite* 42
A.2. *Une approche asymptotique par la loi des événements rares* 43
A.3. *Une approche exacte* 44
B. Intervalle de confiance du paramètre d'une loi de Poisson 45
C. Programmation 47

Table des matières 50