Statistiques

28 janvier 2015

Introduction

Probabilités et statistiques sont les deux composantes de ce qu'on appelle les sciences de l'aléatoire. Ces deux disciplines reposent sur les mêmes objets mathématiques : mesure de probabilité, variables aléatoires, etc. Mais leurs objectifs sont parfaitement contraires.

En probabilité, on étudie les propriétés des variables aléatoires. L'objectif principal est de pouvoir ensuite prédire (en termes de probabilités) les résultats obtenus pour des tirages aléatoires de ces variables aléatoires. Par exemple, l'étude du jeu de pile ou face permet de déterminer la probabilités d'obtenir 3 piles de suite si on lance une pièce 10 fois.

En statistique, l'objet d'étude est une suite de valeurs : un échantillon. L'étude statistique consiste à considérer que ces valeurs sont les tirages aléatoires d'une certaine variable aléatoire. On cherche alors à déterminer des propriétés de cette variable à partir du seul échantillon initial.

On peut par exemple essayer de déterminer la densité de la variable. Concrètement, on considère en général que la variable associée au problème étudié suit un certain type de loi (loi exponentielle, loi normale) et on cherche alors, à partir de l'échantillon à déterminer les paramètres de la loi.

L'outil statistique permet également de tester des hypothèses : une certaine grandeur étudiée est censée avoir une certaine propriété. On dispose d'un certain nombre de valeurs expérimentales. L'étude statistique de cet échantillon peut permettre de confirmer ou d'infirmer la propriété étudiée.

Chapitre 1

Rappels de probabilités

1.1 Variables aléatoires

Définition 1.1.1. *Une* **variable aléatoire** *est une application de la forme* $X : \Omega \to \mathbf{R}$ *où* Ω *est un ensemble muni d'une mesure de probabilité.*

Concrètement, cela signifie que pour toute partie (raisonnable) A *de* $\mathbf{R}$*, on peut considérer la probabilité* $\mathbf{P}(X \in A)$*. Par définition, cette mesure de probabilité satisfait les conditions suivantes :*

- *si* A *et* B *sont des parties disjointes de* $\mathbf{R}$*,* $\mathbf{P}(X \in A \cup B) = \mathbf{P}(X \in A) + \mathbf{P}(X \in B)$ *;*
- $\mathbf{P}(X \in \mathbf{R}) = 1$.

Distinguons deux grandes familles de variables aléatoires : les variables discrètes et les variables continues. Les variables discrètes sont à valeurs dans une partie finie ou dénombrable de $\mathbf{R}$: $\exists (x_n)_{n\in\mathbf{N}},\ \sum_n \mathbf{P}(X = x_n) = 1$. Exemples : le jeu de pile ou face, le résultat d'un dé, la loi de Poisson.
Les variables continues sont à valeurs dans un intervalle de $\mathbf{R}$. Elles sont définies par une **fonction de densité** f_X ainsi :

$$\forall A \subset \mathbf{R},\ \ \mathbf{P}(X \in A) = \int_A f_X(x)dx.$$

Lorsqu'une variable est continue, la probabilité associée à une valeur précise est nulle : $\mathbf{P}(X = x) = 0$. Exemples : la loi uniforme sur un intervalle $[a, b]$, la loi exponentielle, la loi normale.

Définition 1.1.2. *Soient* X *et* Y *des variables aléatoires. On dit que* X *et* Y *sont* **indépendantes** *si pour toutes parties* A *et* B *de* $\mathbf{R}$*,*

$$\mathbf{P}(X \in A \text{ et } Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B).$$

1.1.1 Espérance et variance

Définition 1.1.3. *Soit* X *une variable aléatoire discrète à valeurs dans* $\{x_i;\ i \in I\}$*. On appelle* **espérance** *de* X *le nombre réel défini par*

$$\mathbf{E}(X) = \sum_{i\in I} x_i \mathbf{P}(x_i)$$

Si X est continue, son espérance est définie par

$$\mathbf{E}(X) = \int_{\mathbf{R}} xf(x)dx,$$

où f est la densité de X.

L'espérance représente la valeur moyenne de la variable X. C'est la moyenne des valeurs prises par X pondérées par leurs probabilités.

Proposition 1.1.1. *Soient X et Y des variables aléatoires d'espérances finies et λ un réel. Alors*

$$\mathbf{E}(X+Y) = \mathbf{E}(X) + \mathbf{E}(Y), \qquad \mathbf{E}(\lambda X) = \lambda \mathbf{E}(X).$$

Définition 1.1.4. *Soit X une variable discrète d'espérance finie.*
On appelle **variance** *de X le nombre positif défini par*

$$\mathbf{V}(X) = \sigma^2(X) = \sum_{i \in I} (x_i - \mathbf{E}(X))^2 \mathbf{P}(x_i).$$

Si X est continue de densité f, sa variance est donnée par

$$\mathbf{V}(X) = \sigma^2(X) = \int_{-\infty}^{+\infty} (x - \mathbf{E}(X))^2 f(x)dx.$$

Autrement dit, $\mathbf{V}(X) = \mathbf{E}((X - \mathbf{E}(X))^2)$.

On appelle **écart-type** *de X la racine carrée de sa variance : $\sigma(X) = \sqrt{\mathbf{V}(X)}$.*

L'écart-type mesure l'écart moyen quadratique entre les valeurs prises par X et sa moyenne.

Proposition 1.1.2. *Soit X une variable admettant une variance et a et b des réels. Alors*

$$\mathbf{V}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2, \qquad \mathbf{V}(aX+b) = a^2 \mathbf{V}(X).$$

Si X et Y sont indépendantes,

$$\mathbf{V}(X+Y) = \mathbf{V}(X) + \mathbf{V}(Y)$$

(et aussi $\mathbf{V}(X-Y) = \mathbf{V}(X) + \mathbf{V}(Y)$).

1.2 Théorèmes limites

On effectue un grand nombre de tirages indépendants d'une même variable aléatoire. On souhaite décrire les comportements limites des tirages ainsi obtenus. Par exemple, lorsqu'on lance un grand nombre de fois une pièce, on s'attend à obtenir à peu près autant de "pile" que de "face". Nous allons justifier cela et mesurer de manière probabiliste ce "à peu près autant".

Théorème 1.2.1. Loi forte des grands nombres

Soit $(X_k)_{k\in\mathbf{N}^*}$ une suite de variables aléatoires indépendantes et de même loi. Soit m leur espérance. Alors

$$\lim_{n\to+\infty} \frac{X_1+X_2+\cdots+X_n}{n} = m \quad \textit{presque sûrement.}$$

Précisemment, cela signifie

$$\mathbf{P}\left(\lim_{n\to+\infty} \frac{X_1+X_2+\cdots+X_n}{n} = m\right) = 1.$$

Théorème 1.2.2. Théorème central limite

Soit $(X_k)_{k\in\mathbf{N}^*}$ une suite de variables aléatoires i.i.d. d'espérance m et de variance σ^2 finies. Notons $S_n = X_1+X_2+\cdots+X_n$.

Quand n tend vers $+\infty$, la suite $\dfrac{S_n-nm}{\sigma\sqrt{n}}$ converge en loi vers la loi normale $\mathcal{N}(0,1)$.

Reformulations : le théorème signifie qu'à la limite, les probabilités sont données par la densité $f_{\mathcal{N}}$ de la loi normale :

$$\lim_{n\to+\infty} \mathbf{P}\left(a \le \frac{S_n-nm}{\sigma\sqrt{n}} \le b\right) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx.$$

Ou encore :

$$\lim_{n\to+\infty} \mathbf{P}\left(\frac{S_n-nm}{\sigma\sqrt{n}} \le \alpha\right) = F_{\mathcal{N}}(\alpha),$$

où $F_{\mathcal{N}}$ désigne la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. La table de la loi normale donne les valeurs de $F_{\mathcal{N}}$.

Utilisation du théorème : en pratique, quand n est assez grand ($n \ge 30$), le théorème permet d'approcher la loi de S_n par une loi normale :

$$S_n \sim \mathcal{N}(nm, n\sigma^2) \quad \text{et} \quad \frac{S_n-nm}{\sigma\sqrt{n}} = \frac{\frac{S_n}{n}-m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

Par exemple, la loi binomiale étant une somme de Bernouilli indépendantes, on peut approcher pour n grand, cette loi par une loi normale :

$$\mathcal{B}(n,p) \sim \mathcal{N}(np, np(1-p)).$$

Ainsi, si $S \sim \mathcal{B}(100, \frac{1}{2})$, on peut approcher la loi de S par $S \sim \mathcal{N}(50,25)$ et on obtient alors $\frac{S-50}{5} \sim \mathcal{N}(0,1)$. Alors

$$\begin{aligned}\mathbf{P}(40 \le S \le 60) &= \mathbf{P}\left(\frac{40-50}{5} \le \frac{S-50}{5} \le \frac{60-50}{5}\right)\\ &= F_{\mathcal{N}}(2) - F_{\mathcal{N}}(-2) = 2F_{\mathcal{N}}(2) - 1 = 0,9545.\end{aligned}$$

La valeur exacte est $\sum_{k=40}^{60} \frac{1}{2^{100}}\binom{100}{k}$; elle est plus difficile à calculer.

Le théorème central limite est le résultat théorique qui permet en statistiques de définir la notion d'intervalle de confiance. L'application la plus usuelle est sans doute le sondage.

Chapitre 2

Estimation paramétrique

Notations

Dans tout ce chapitre, nous noterons $n \in \mathbf{N}$ la taille de l'échantillon étudié. Nous noterons $x_1, x_2, \ldots, x_n$ l'échantillon. C'est une suite de n valeurs réelles.

L'étude statistique de cet échantillon consiste à considérer que ces n valeurs sont les résultats de n tirages indépendants d'une même variable aléatoire X. Nous noterons m et σ^2 l'espérance et la variance de X.

2.1 Moyenne

Définition 2.1.1. Estimateur de la moyenne

Soient $X_1, \ldots X_n$ des variables aléatoires indépendante de même loi que X. On définit la variable aléatoire

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Elle est appelée estimateur de l'espérance de X.

Cela signifie concrètement que pour un échantillon $x_1, \ldots x_n$ d'une variable X, le nombre $\frac{1}{n} \sum_{i=1}^{n} x_i$ fournit une estimation de la moyenne m de X. Précisons cela :

Définition 2.1.2. *L'estimateur de la moyenne est un estimateur sans biais : il est* en moyenne *égal à ce qu'il est censé estimer.*

$$\mathbf{E}[\overline{X}] = \mathbf{E}[X] = m.$$

Définition 2.1.3. *L'estimateur de la moyenne est un estimateur convergent : si la taille de l'échantillon tend vers $+\infty$, l'estimateur converge vers ce qu'il est censé estimer. Presque sûrement*

$$\lim_{n \to +\infty} \overline{X} = m.$$

Ce résultat est simplement la loi forte des grands nombres. C'est un point essentiel de l'estimation statistique : plus l'échantillon est grand, plus l'estimation

est bonne a priori.

Il est ensuite possible de quantifier la qualité de l'approximation grâce au théorème central limite. À partir de l'estimation $\overline{X}$ effectuée, on détermine l'ensemble des valeurs exactes de m possibles pour lesquelles l'estimation obtenue est dans un intervalle de probabilité importante. Autrement dit, on cherche l'ensemble des valeurs exactes de la moyenne pour lesquelles l'échantillon étudié est vraisemblable.

Définition 2.1.4. *Soit α une probabilité (plutôt petite). Soit $\overline{x}$ l'estimation de la moyenne obtenue à partir d'un échantillon de taille n avec $n > 30$. L'***intervalle de confiance au risque** α *(ou au niveau de confiance $1-\alpha$) associé à l'estimation est l'intervalle*

$$\left[\overline{x} - \tau_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + \tau_\alpha \frac{\sigma}{\sqrt{n}}\right],$$

où τ_α est le nombre tel que $2F_{\mathcal{N}}(\tau_\alpha) - 1 = 1 - \alpha$ et σ est l'écart-type de X.

Pour $\alpha = 0,05$ on a $\tau_\alpha \approx 1,96$ et pour $\alpha = 0,01$, on a $\tau_\alpha \approx 2,58$.

Interprétation : on considère, avec le TCL, que $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit la loi $\mathcal{N}(0,1)$. Pour cette loi, l'ensemble $[-\tau_\alpha, \tau_\alpha]$ a une mesure de probabilité $1-\alpha$. On en déduit $\mathbf{P}(m - \tau_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq m + \tau_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$.

L'intervalle de confiance est l'ensemble des valeurs de m pour lesquelles l'estimation $\overline{x}$ est bien dans cet intervalle de mesure $1-\alpha$.

Remarque : en général, lorsqu'on cherche à estimer une moyenne, on ne connaît pas non plus l'écart-type. On utilise alors dans la formule ci-dessus l'estimation $\overline{\sigma}$ de σ que l'on va présenter dans la suite.

Lorsque la variable X suit une loi normale, on peut préciser les intervalles de confiance, même avec des échantillons de petite taille :

Proposition 2.1.1. *Supposons que X suive la loi normale $\mathcal{N}(m, \sigma^2)$. Alors, si σ est connu,*

$$\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

Si σ n'est pas connu et est estimée par $\bar{\sigma}$,

$$\frac{\overline{X} - m}{\frac{\bar{\sigma}}{\sqrt{n}}} \sim T_{n-1},$$

où T_{n-1} désigne la loi de Student à $n-1$ degrés de liberté. Lorsque n est grand, la loi de Student est très proche de la loi normale centrée réduite.

Ces lois permettent de déterminer des intervalles de confiance pour m.

2.2 Variance

Définition 2.2.1. Estimateur de la variance
Soient $X_1, \ldots, X_n$ des variables aléatoires indépendante de même loi que X.
– Si l'espérance m de X est connue, on définit la variable aléatoire

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

– Si l'espérance de X est inconnue, on définit la variable aléatoire

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Ces variables sont appelées estimateurs de la variance de X. L'estimateur de l'écart-type est la racine carrée de ces estimateurs.

Hormis le facteur $\frac{1}{n-1}$, on reconnaît la définition discrète de la variance. Ces estimations sont simplement les variances empiriques.

Proposition 2.2.1. *Dans les deux cas, l'estimateur est sans biais et convergent :*

$$\mathbf{E}[\bar{\sigma}^2] = \sigma^2(X), \qquad et \qquad \lim_{n\to\infty} \bar{\sigma}^2 = \sigma^2.$$

C'est grâce au facteur $\frac{1}{n-1}$ que le second estimateur est bien sans biais.

Comme pour tout estimateur, on souhaite donner un intervalle de confiance pour la variance estimée. On peut le faire si la variable X suit une loi normale.

Proposition 2.2.2. *Supposons que X est une variable de loi normale $\mathcal{N}(m, \sigma^2)$ et considérons un échantillon de taille n.*
Si m est connu, notons $\bar{\sigma}_1^2$ l'estimation de σ^2. Si m est estimée par $\overline{X}$, notons $\bar{\sigma}_2^2$ l'estimation.
Alors les variables $\frac{n\bar{\sigma}_1^2}{\sigma^2}$ et $\frac{(n-1)\bar{\sigma}_2^2}{\sigma^2}$ suivent la loi du χ^2 à respectivement n et $n-1$ degrés de liberté.

Il est ainsi possible de déterminer un intervalle de confiance au risque α pour la variance de X : on trouve des bornes a et b telles que pour la loi du χ^2, $\mathbf{P}([a,b]) = 1-\alpha$. L'intervalle de confiance pour σ^2 est alors

$$\left[\frac{n\bar{\sigma}_1^2}{b}, \frac{n\bar{\sigma}_1^2}{a}\right] \qquad ou \qquad \left[\frac{(n-1)\bar{\sigma}_2^2}{b}, \frac{(n-1)\bar{\sigma}_2^2}{a}\right].$$

2.3 Maximum de vraisemblance

On dispose d'un échantillon $x_1, \ldots, x_n$ et on suppose qu'il est la réalisation de n tirages indépendants d'une même loi de densité f_θ dépendant d'un paramètre θ. On souhaite estimer θ à partir de l'échantillon.

La méthode du maximum de vraisemblance consiste à estimer θ par la valeur pour laquelle la probabilité d'observer le tirage $(x_1, \ldots, x_n)$ est maximale.

Supposons dans un premier temps la loi discrète. la probabilité d'observer l'échantillon est, d'après l'hypothèse d'indépendance :

$$\mathbf{P}_\theta(x_1, \ldots, x_n) = \mathbf{P}_\theta(x_1) \times \mathbf{P}_\theta(x_2) \times \cdots \mathbf{P}_\theta(x_n),$$

les probabilités dépendant du paramètre θ. On cherche alors la valeur de θ pour laquelle cette probabilité est maximale.

Dans le cas continu, la probabilité infinitésimale d'obtenir l'échantillon est

$$f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) dx_1 dx_2 \cdots dx_n.$$

On oublie les dx_i qui ne dépendent pas de θ et on cherche à maximiser le reste. Notons $L(\theta) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$; elle est appelée fonction de vraisemblance. On peut commencer par chercher θ tel que

$$\frac{\partial L(\theta)}{\partial \theta} = 0.$$

Il est souvent plus simple de passer par le *log*. En effet, maximiser L_θ est équivalent à maximiser $\ln(L_\theta)$. Ainsi, rechercher le maximum de vraisemblance revient à chercher θ tel que

$$\frac{\partial \ln(L(\theta))}{\partial \theta} = \sum_{i=1}^{n} \frac{\frac{\partial f_\theta(x_i)}{\partial \theta}}{f_\theta(x_i)} = 0.$$

Le plus souvent, cette équation n'a qu'une seule solution que l'on note $\hat{\theta}$. Ce $\hat{\theta}$ est alors l'estimateur du maximum de vraisemblance de θ.

Prenons un exemple. On cherche à estimer le paramètre λ d'une loi exponentielle $\mathcal{E}(\lambda)$ dont la densité est $f_\lambda(x) = \lambda e^{-\lambda x}$ pour $x > 0$. Pour un échantillon $x_1, \ldots, x_n$ donné, la fonction de vraisemblance L_λ est alors

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

On cherche λ maximisant $L(\lambda)$: $\ln(L(\lambda)) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$ et sa dérivée en λ est

$$\frac{\partial \ln(L(\theta))}{\partial \theta} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Cette dérivée est nulle pour

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

On obtient ainsi l'estimation de λ associée à l'échantillon $(x_1, \ldots, x_n)$.
Remarquons que l'estimateur obtenu ici est raisonnable. L'espérance de la loi $\mathcal{E}(\lambda)$ est $\frac{1}{\lambda}$ et notre estimateur est justement $\hat{\lambda} = \frac{1}{\overline{X}}$.

Chapitre 3

Tests d'hypothèse

3.1 Généralités

On considère une variable aléatoire dont on cherche à décrire un des paramètres. On formule une hypothèse quant à la valeur de ce paramètre.

Le but d'un test statistique est de confronter cette hypothèse aux résultats obtenus avec un échantillon afin d'accepter ou de rejeter l'hypothèse.

Nous noterons H_0 l'hypothèse formulée ; elle est appelée **hypothèse nulle**. Toute autre hypothèse à laquelle on peut la confronter sera notée H_1 et sera appelée **hypothèse alternative**.

Un **test d'hypothèse** est une procédure qui à partir de la donnée d'un échantillon permet de choisir entre les hypothèses H_0 et H_1.

Les deux hypothèses n'ont pas le même statut. Dans le cadre d'un test statistique, on considère toujours que H_0 est a priori vraie. C'est cette hypothèse qui est véritablement soumise au test. Si le test aboutit au choix de H_1, cela signifie qu'on rejette H_0. S'il aboutit au choix de H_0, cela signifie qu'on ne rejette pas H_0.

Comme la décision repose sur un échantillon considéré aléatoire, le test peut aboutir à une erreur. On peut estimer la probabilités d'erreur.

Il y a deux erreurs possibles. On appelle **erreur de première espèce** le fait de rejeter H_0 alors que H_0 est vraie. On appelle **risque de première espèce** la probabilité de cette erreur :

$$\alpha = \mathbf{P}(\text{rejeter } H_0 | H_0 \text{ est vraie}).$$

On appelle **erreur de seconde espèce** le fait d'accepter H_0 alors que H_0 est fausse. On appelle **risque de seconde espèce** la probabilité de cette erreur :

$$\beta = \mathbf{P}(\text{accepter } H_0 | H_0 \text{ est fausse}).$$

Quand on élabore un test, on tient à contrôler la valeur du risque α. En général on impose $\alpha = 0,05$ ou $\alpha = 0,01$. Mais on ne peut dans ce cas peu ou mal contrôler le risque β. Plus α est choisi petit, plus β sera grand. Le nombre $1 - \beta$ est appelé **puissance du test**.

3.2 Test de conformité d'une moyenne

On considère une variable aléatoire d'espérance m inconnue et de variance σ^2. On pense que la moyenne m est égale à m_0. On cherche à construire un test permettant, à partir d'un échantillon, d'accepter cette hypothèse ou de la rejeter.

On pose :

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m \neq m_0.$$

On considère un échantillon $x_1, \ldots, x_n$ de taille n de la variable X et on note $\overline{x}$ l'estimation de m associée.

On sait que sous l'hypothèse H_0 et si $n \geq 30$, $\frac{\overline{X}-m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Pour un risque α donné, on a alors

$$\mathbf{P}(-t_\alpha \leq \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \leq t_\alpha) = 1 - \alpha,$$

où t_α est obtenu avec les tables de la loi normale. Ainsi, si H_0 est vraie, on devrait obtenir avec une grande probabilité

$$m_0 - t_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{x} \leq m_0 + t_\alpha \frac{\sigma}{\sqrt{n}}.$$

On peut ainsi proposer le test suivant : au risque α

$$\begin{cases} \text{si } \overline{x} \in \left[m_0 - t_\alpha \frac{\sigma}{\sqrt{n}}, m_0 + t_\alpha \frac{\sigma}{\sqrt{n}}\right] & \text{on accepte } H_0 \\ \text{si } \overline{x} \notin \left[m_0 - t_\alpha \frac{\sigma}{\sqrt{n}}, m_0 + t_\alpha \frac{\sigma}{\sqrt{n}}\right] & \text{on rejette } H_0 \end{cases}$$

L'intervalle définissant le test ressemble beaucoup à un intervalle de confiance. Il est cependant centré en m_0 et non en $\overline{x}$.

Le principe de ce test est simple : on regarde l'écart entre la moyenne théorique m_0 et la moyenne empirique $\overline{x}$. L'analyse statistique permet de dire si cet écart est significativement important. S'il l'est, on peut se permettre de remettre en cause la valeur de m_0.

Si la variance σ^2 n'est pas connue mais est estimée par $\overline{\sigma}^2$ à partir de l'échantillon et si n est grand, on peut établir le même test en effectuant simplement le remplacement.

Si n est petit et si on suppose que X suit une loi normale, on peut également définir un test à partir de la loi de Student.

3.3 Test de conformité d'une variance

On suppose que X suit une loi normale (m, σ^2). On cherche à tester la valeur de σ et on pense qu'elle est égale à σ_0 :

$$H_0 : \sigma = \sigma_0 \quad \text{et} \quad H_1 : \sigma \neq \sigma_0.$$

On sait que sous l'hypothèse H_0, $\frac{(n-1)\overline{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-1}$, où $\overline{\sigma}^2$ est l'estimation (sans biais) de la variance à partir de l'échantillon.

On détermine avec les tables un intervalle $[a, b]$ de mesure $1-\alpha$ pour la loi χ^2_{n-1} et on établit le test suivant au risque α :

$$\begin{cases} \text{si } \overline{\sigma}^2 \in \left[\dfrac{a\sigma_0^2}{n-1}, \dfrac{b\sigma_0^2}{n-1}\right] & \text{on accepte } H_0 \\ \text{si } \overline{\sigma}^2 \notin \left[\dfrac{a\sigma_0^2}{n-1}, \dfrac{b\sigma_0^2}{n-1}\right] & \text{on rejette } H_0 \end{cases}$$

3.4 Test de comparaison de deux moyennes

On dispose de deux échantillons de tailles n_1 et n_2. On considèrent qu'ils proviennent de deux variables aléatoires de lois normales $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$. On cherche à tester le fait que leurs espérances sont égales :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 \neq m_2.$$

Si σ_1 et σ_2 sont connues et si hypothèse H_0 est vraie, alors

$$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

On en tire le test suivant au risque α :

$$\begin{cases} \text{si } \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in [-t_\alpha, t_\alpha] & \text{on accepte } H_0 \\ \text{si } \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \notin [-t_\alpha, t_\alpha] & \text{on rejette } H_0 \end{cases}$$

où t_α est tel que $\mathbf{P}([-t_\alpha, t_\alpha]) = 1-\alpha$ pour la loi $\mathcal{N}(0,1)$.

Si les variances ne sont pas connues et si $n \geq 30$, le test est le même en remplaçant simplement σ_1 et σ_2 par leurs estimations.

3.5 Test du χ^2 d'adéquation

Tous les tests précédents sont des tests paramétriques : on cherche à tester la valeur d'un paramètre. Nous allons ici tester le fait qu'une variable suive une certaine loi de probabilité. L'idée est de comparer la répartition des valeurs de l'échantillon avec la répartition théorique décrite par la densité de la loi considérée.

$$H_0 : X \text{ suit la loi de densité } f \quad \text{et} \quad H_1 : X \text{ ne suit pas cette loi.}$$

Pour élaborer le test, on commence par découper $\mathbf{R}$ en intervalles disjoints : $\mathbf{R} = I_1 \cup I_2 \cup \cdots \cup I_k$. Pour chaque intervalle I_j, on note N_j le nombre de termes de l'échantillon appartenant à l'intervalle I_j ($\sum_{j=1}^{k} N_j = n$).

Pour que le test fonctionne correctement, il faut que pour tout j, $N_j \geq 5$. Autrement dit, il faut choisir le découpage de $\mathbf{R}$ en intervalles de manière à ce que chaque morceau contiennent suffisamment de termes de l'échantillon.

Nous allons comparer ces effectifs aux effectifs théoriques : pour tout j, on note p_j la mesure de probabilité de l'intervalle I_j pour la loi testée : $p_j = \mathbf{P}(I_j) = \int_{I_j} f(x)\mathrm{d}x$. Si on ramène ces valeurs à la taille n de l'échantillon, on peut considérer que les effectifs théoriques des intervalles I_j sont les nombres $np_1, \ldots, np_k$.

On calcule alors $T = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$. Ce nombre mesure un écart entre les effectifs théoriques et empiriques. Si l'hypothèse H_0 est vraie, alors la variable T suit la loi χ^2_{k-1} (loi du chi-deux à $k-1$ degrés de liberté). On peut ainsi établir un test : pour un risque α imposé, on cherche t_α tel que $\mathbf{P}([0, t_\alpha]) = 1 - \alpha$ pour la loi χ^2_{k-1}. Alors

$$\begin{cases} \text{si } T \in [0, t_\alpha] & \text{on accepte } H_0 \\ \text{si } T \notin [0, t_\alpha] & \text{on rejette } H_0 \end{cases}$$

Si certains paramètres de la loi testée ne sont pas connus et doivent être estimés, la variable T suit alors la loi χ^2_{k-1-s} où s désigne le nombre de paramètres à estimer pour pouvoir calculer les probabilités p_j. Par exemple, si on cherche à tester le fait que X suive une loi normale (sans rien préciser d'autres), il faudra estimer les paramètres m et σ^2 de cette loi normale à partir de l'échantillon et la variable T suivra la loi χ^2_{k-1-2}.

3.6 Test du χ^2 d'indépendance

On considère maintenant des échantillons où deux grandeurs sont mesurées simultanément. On dispose donc d'un échantillon de la forme $(x_1, y_1), \ldots, (x_n, y_n)$. On suppose que les valeurs x_i sont des tirages d'une même loi X et que les y_i sont des tirages d'une même loi Y. Notre échantillon est donc une suite de tirages du couple (X, Y).
On se demande si X et Y sont des variables indépendantes.

L'hypothèse du test est donc

H_0 : X et Y sont indépendantes et H_1 : X et Y ne sont pas indépendantes.

Le test d'indépendance du χ^2 ressemble beaucoup au test précédent. Il repose sur le résultat suivant : si X et Y dont indépendantes, alors pour tout événement de la forme $A \times B$, $\mathbf{P}(X \in A \text{ et } Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B)$. Nous allons confronter l'échantillon à cette propriété.

On classe les valeurs x_k observées dans r intervalles $I_1, \ldots, I_r$ et les valeurs y_k observées dans s intervalles $J_1, \ldots, J_s$. Pour tous i et j, on note O_{ij} le nombre

de couples (x_k, y_k) appartenant à $I_i \times J_j$. Pour que le test fonctionne correctement, il faut que la plupart des effectifs O_{ij} soient supérieurs à 5.
Notons également pour tout i, $O_{i+} \sum_{j=1}^{s} O_{ij}$ le nombre de valeurs x_k appartenant à I_i et pour tout j, $O_{+j} = \sum_{i=1}^{i} O_{ij}$ le nombre de valeurs y_k appartenant à J_j. Si les variables X et Y étaient indépendantes, on devrait idéalement obtenir $\frac{O_{ij}}{n} = \frac{O_{i+}}{n} \times \frac{O_{+j}}{n}$. On note ainsi $E_{ij} = \frac{O_{i+}O_{+j}}{n}$ ces effectifs théoriques.

On calcule alors $T = \sum_{i=1}^{r} \sum_{j=1^s} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. Ce nombre mesure encore un écart entre les effectifs théoriques et empiriques. Si l'hypothèse H_0 est vraie, alors pour r et s assez grand, on peut considérer que la variable T suit la loi du χ^2 à $(r-1)(s-1)$ degrés de liberté.. On peut ainsi établir un test : pour un risque α imposé, on cherche t_α tel que $\mathbf{P}([0, t_\alpha]) = 1 - \alpha$ pour la loi $\chi^2_{(r-1)(s-1)}$.
Alors

$$\begin{cases} \text{si } T \in [0, t_\alpha] & \text{on accepte } H_0 \\ \text{si } T \notin [0, t_\alpha] & \text{on rejette } H_0 \end{cases}$$