Chapitre 2: Tests d'ajustement d'un échantillon à une loi théorique

*

1 Introduction

Le problème que nous allons examiner dans ce chapitre est d'une grande importance pratique : Etant donné un échantillon observé $x_1, x_2, \ldots, x_n$ constitué d'une suite numérique de mesures indépendantes d'un phénomène aléatoire dont la loi de probabilité n'est pas connue précisément, on veut tester si cet échantillon provient d'une loi F donnée par exemple de la loi $\mathcal{N}(0,1)$. Les méthodes qu'on va utiliser s'appellent **méthodes d'ajustement** de l'échantillon observé à la loi théorique F. **Le principe de ces méthodes est le suivant** : On fait l'hypothèse $\mathbf{H}_0$ que l'échantillon observé est issu de la loi F. La méthode consiste à transformer les valeurs observé d'une certaine façon (soit en un nombre pour l'ajustement du χ^2, soit en une fonction pour l'ajustement de Kolmogorov-Smirnov) de sorte que suivant le résultat obtenu, on puisse décider avec un **niveau de confiance** $1-\alpha \in]0,1[$ donné, :

soit de rejeter l'hypothèse $\mathbf{H}_0$, soit de l'accepter

Le nombre α généralement petit ($0,05$ ou $0,01$) est la probabilité d'accepter l'hypothèse alors qu'elle est fausse ; c'est **le risque d'erreur** (de première espèce) dont nous reparlerons plus en détail ci-dessous mais il est important de comprendre que c'est l'expérimentateur qui fixe le risque α qu'il accepte de prendre, la méthode tient compte de ce risque et donne un résultat qui se traduit par la décision d'accepter ou de rejetter l'hypothèse.

2 Le test du Chi^2

2.1 La loi du χ^2

Définition 2.1 *: Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi normale $\mathcal{N}(0,1)$. On appelle loi du χ^2 à n degrés de liberté, la loi de la variable aléatoire*

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2.$$

Remarque : Si on considère $(X_1, \ldots, X_n)$ comme un vecteur aléatoire de $\mathbb{R}^n$, χ_n^2 est le carré du module de ce vecteur.

*cours de Statistiques de M. Léonard Gallardo, Master 1, Université de Tours, année 2007-2008, Laboratoire de Mathématiques et Physique Théorique-UMR 6083 du CNRS, Parc de Grandmont, 37200 TOURS, FRANCE. email : gallardo@univ-tours.fr

Proposition 2.2 *: La loi du χ^2_n a une densité de probabilité de la forme*

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} \mathbf{1}_{[0,+\infty[}(x),$$

et une espérance et une variance égales à :

$$\mathbb{E}(\chi^2_n) = n, \quad Var(\chi^2_n) = 2n.$$

démonstration : C'est un exercice de calcul vu en TD de probabilités.

Remarque : Dans le cours de probabilités on a étudié la loi $\Gamma(\alpha, b)$ de paramètres $\alpha > 0$ et $b > 0$ qui est la loi d'une variable aléatoire de densité

$$f_{\alpha,b}(x) = \frac{b^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-bx} \mathbf{1}_{[0,+\infty[}(x).$$

La loi χ^2_n est donc une loi $\Gamma(n/2, 1/2)$.

2.2 La loi multinomiale

Notation : Toute mesure de probabilité concentrée sur l'ensemble $\{1, 2, \ldots, k\}$ des entiers entre 1 et k, s'écrit $\sum_{i=1}^k p_i \delta_i$ où δ_i est la mesure de Dirac en i, $p_i \geq 0$ et $\sum_{i=1}^k p_i = 1$. Une telle loi sera représentée par un vecteur ligne

$$p = (p_1, \ldots, p_k). \tag{1}$$

Soit $n \geq 1$ un entier fixé et $(X_1, \ldots, X_n)$ un n-échantillon d'une loi $p = (p_1, \ldots, p_k)$ portée par $\{1, 2, \ldots, k\}$. Pour chaque $1 \leq i \leq k$, on considère la variable aléatoire

$$N_i = \sum_{j=1}^n \mathbf{1}_{[X_j = i]}, \tag{2}$$

qui représente le nombre de variables de l'échantillon qui prennent la valeur i.

Définition 2.3 *: La loi du vecteur aléatoire $N = (N_1, \ldots, N_k)$ s'appelle loi multinomiale de paramètres $(n; p_1, \ldots, p_k)$ et notée $\mathcal{M}(n, p)$. Elle est telle que*

$$\mathbb{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}, \tag{3}$$

où $n_1, \ldots, n_k$ sont des entiers tels que $n_1 + \cdots + n_k = n$.

L'importance de cette loi est due au fait que le vecteur aléatoire

$$\bar{p}_n = \frac{N}{n} = (\frac{N_1}{n}, \ldots, \frac{N_k}{n}) \tag{4}$$

est tel que

$$\mathbb{E}(\frac{N_i}{n}) = p_i \quad (i = 1, \ldots, k) \tag{5}$$

et

$$\lim_{n\to\infty} \bar{p}_n = p \quad \text{presque surement.} \tag{6}$$

En effet, pour tout $i = 1, \ldots, k$ (fixé), les variables aléatoires $(\mathbf{1}_{[X_j=i]})_{1\leq j\leq n}$ sont i.i.d. et d'après la loi forte des grands nombres, on a

$$\frac{N_i}{n} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{[X_j=i]} \overset{p.s.}{\to} \mathbb{E}(\mathbf{1}_{[X_j=i]}) = \mathbb{P}(X_j = i) = p_i.$$

On traduit la propriété (6) en disant que $\bar{p}$ est **l'estimateur empirique** de la loi $p = (p_1, \ldots, p_k)$. La propriété (5) dit que l'estimateur $\bar{p}$ est **non biaisé** (ou sans biais).
Considérons maintenant le vecteur centré et quasi-réduit associé au vecteur multinomial :

$$\left(\frac{N_1 - np_1}{\sqrt{np_1}}, \ldots, \frac{N_k - np_k}{\sqrt{np_k}}\right) \tag{7}$$

Pour la somme des carrés de ses composantes, on a le résultat fondamental suivant

Théorème 2.4 *:*

$$\sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i} \overset{\mathcal{L}}{\to} \chi^2_{k-1} \quad (n \to \infty). \tag{8}$$

La démonstration de ce résultat sera vue ultérieurement. Notons que nous avons utilisé le terme quasi-réduit pour le vecteur aléatoire (7) car la véritable variable centrée réduite associée à N_i est $\widetilde{N_i} = \frac{N_i - np_i}{\sqrt{np_i(1-p_i)}}$. Notons aussi que le théorème limite central implique que la i-ième composante du vecteur centré et quasi-réduit, converge en loi vers une variable aléatoire normale $\mathcal{N}(0, 1-p_i)$ quand $n \to +\infty$. Le fait qu'on trouve une loi du χ^2_{k-1} comme limite dans le théorème, est lié à ce fait et à la non-indépendance des composantes (car $N_1 + \cdots + N_n = n$).

2.3 Le test du χ^2 d'ajustement

Soit $p = (p_1, \ldots, p_k)$ suivant la notation (1) une loi de probabilité de référence sur l'ensemble $\{1, \ldots, k\}$. Si $q = (q_1, \ldots, q_k)$ est une autre loi de probabilité, on définit *la distance du χ^2 de q à p* par :

$$\chi^2(p, q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{p_i}. \tag{9}$$

On prendra garde à la terminologie[1] car $\chi^2(p, q)$ n'est pas le carré d'une vraie distance[2] et que les termes $(p_i - q_i)^2$ ont plus d'importance dans la somme (9) lorsque la valeur de p_i est faible. Si l'on prend pour q la répartition empirique $\bar{p}_n$ définie en (4) d'un n-échantillon de la loi p, on mesure plutôt l'écart entre p et $\bar{p}_n$ de la manière suivante :

Définition 2.5 *: Le χ^2 d'ajustement entre la loi p et la loi empirique $\bar{p}_n$ est la variable aléatoire*

$$\chi^2_n(p, \bar{p}_n) = n\chi^2(p, \bar{p}_n) = \sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i}, \tag{10}$$

[1] on verra que le terme χ^2 est utilisé en statistiques pour désigner des quantités diverses ayant un lien avec la loi du χ^2.

[2] noter que $\chi^2(p, q) \neq \chi^2(q, p)$

qu'on écrit plus traditionellement sous la forme

$$\chi_n^2(p,\bar{p}_n) = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}, \tag{11}$$

où $e_i = np_i = \mathbb{E}(N_i)$ est la valeur "espérée" de N_i.

Remarque : La loi exacte de $\chi_n^2(p,\bar{p}_n)$ n'est pas connue mais d'après le résultat du Théorème 2.4, **si n est grand, on peut considérer que $\chi_n^2(p,\bar{p}_n)$ suit la loi $\chi^2(k-1)$.**

Observation fondamentale : La convergence en loi de $\chi_n^2(p,\bar{p}_n)$ vers $\chi^2(k-1)$ est très sensible au fait que $\bar{p}_n$ est la loi empirique de p. En effet supposons qu'on se soit trompé et que $\bar{p}_n$ soit en réalité la loi empirique d'une loi $q \neq p$. Alors d'après la loi forte des grands nombres, $\frac{N_i}{n} \to q_i$ p.s. et donc

$$\frac{1}{n}\chi_n^2(p,\bar{p}_n) = \chi^2(p,\bar{p}_n) = \sum_{i=1}^{k} \frac{(p_i - N_i/n)^2}{p_i} \to \chi^2(p,q) > 0 \quad p.s. \tag{12}$$

Il en résulte que $\chi_n^2(p,\bar{p}_n) \to +\infty$ p.s. Ainsi si n est grand, **les valeurs observées de $\chi_n^2(p,\bar{p}_n)$ seront très grandes**. Cette observation est à la base du test suivant :

LA PROCÉDURE DU TEST DU CHI-2 : A partir d'un n-échantillon d'une loi discrète sur l'ensemble $\{1,\ldots,k\}$, on veut tester l'hypothèse

$$H_0 = \text{"la loi de l'échantillon est égale à } p\text{"}$$

contre l'hypothèse alternative H_1="la loi de l'échantillon est différente de p". Compte tenu de ce qui précède, on forme la quantité $\chi_n^2(p,\bar{p}_n)$ (11) à partir des valeurs observées N_i et des valeurs espérées e_i résultant de l'hypothèse H_0. Suivant l'idée expliquée ci-dessus (c'est à dire que **si la valeur observée de $\chi_n^2(p,\bar{p}_n)$ est trop grande, l'hypothèse H_0 est peu crédible** mais il faut préciser ce qu'on entend par là. Le plus simple est de pouvoir définir une borne au delà de laquelle on rejettera H_0. Pour cela on procède de la manière suivante :

1) On choisit un risque (de première espèce) α.
2) Avec la table de la loi du $\chi^2(k-1)$, on détermine la borne $b_\alpha := \chi^2_{k-1,\alpha}$ de la queue d'ordre α de la loi de la variable X du $\chi^2(k-1)$ (i.e. $\mathbb{P}(X > b_\alpha) = \alpha$).
3) on rejette l'hypothèse H_0 au profit de H_1 si :

$$\chi_n^2(p,\bar{p}_n) > b_\alpha. \tag{13}$$

Justification du test : Sous l'hypothèse H_0, et si n est assez grand, on sait d'après le Théorème 2.4 que la variable aléatoire $\chi_n^2(p,\bar{p}_n)$ suit la loi $\chi^2(k-1)$, donc

$$\mathbb{P}_{H_0}(rejeter H_0) = \mathbb{P}_{H_0}\left(\chi_n^2(p,\bar{p}_n) > \chi^2_{k-1,\alpha}\right) \approx \alpha,$$

ce qui montre que le test est justifié.

Remarque : Le test du χ^2 est seulement un test asymptotique et il faut que n soit assez grand. Disons seulement qu'il vaut mieux éviter d'utiliser le test du χ^2 si la taille de l'échantillon est inférieure à 50.

Exemple : On veut tester si un dé n'est pas truqué au risque $\alpha = 0,05$. Pour cela on lance le dé 60 fois et on obtient les résultats suivants

face	1	2	3	4	5	6
N_i	15	7	4	11	6	17
e_i	10	10	10	10	10	10

On a fait figurer dans le tableau la valeur espérée e_i du nombre d'apparitions de la face i dans l'l'hypothèse où le dé n'est pas truqué, ceci afin de faciliter le calcul de la quantité $\chi^2_n(p, \bar{p}_n)$ qui est donc ici égale à

$$\sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i} = \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} = 13,6.$$

Sous l'hypothèse H_0 :"$p_1 = \cdots = p_6 = \frac{1}{6}$", la variable aléatoire $\chi^2_n(p, \bar{p}_n)$ a donc pris la valeur $13,6$. Or le seuil de rejet lu dans la table de la loi du $\chi^2(k-1) = \chi^2(5)$ est $\chi^2_{5,0,05} = 11,07$. La valeur observée dépassant cette valeur, on est amené à rejeter l'hypothèse H_0 au risque $\alpha = 0,05$. On notera qu'au risque $\alpha = 0,025$, on rejette aussi H_0. Mais au risque $\alpha = 0,01$, on ne peut plus rejeter l'hypothèse H_0 malgré la mauvaise impression donnée par les résultats. Si on persiste à vouloir le risque $0,01$, il est plus raisonnable de recommencer l'expérience avec un échantillon de taille beaucoup plus grande.

2.4 Le test du χ^2 d'indépendance

On considère ici un couple (X, Y) de variables aléatoires. On suppose que X (resp. Y) prend ses valeurs dans l'ensemble $\{1, \ldots, k\}$ (resp. $\{1, \ldots, l\}$). Si $p_{ij} = \mathbb{P}(X = i, Y = j)$, on représentera la loi du couple (X, Y) par la matrice $p = (p_{ij})$ à k lignes et l colonnes.
Le problème qui nous intéresse dans ce paragraphe est de tester l'indépendance des variables X et Y. Rappelons[3] que si la loi p est connue, les variables X et Y sont indépendantes si et seulement si la loi p est le produit (tensoriel) de ses lois marginales i.e.

$$\forall i = 1, \ldots, k, \quad \forall j = 1, \ldots, l, \quad p_{ij} = p_{i.} p_{.j}, \tag{14}$$

où $p_{i.} = \sum_{j=1}^{l} p_{ij}$ (resp. $p_{.j} = \sum_{i=1}^{k} p_{ij}$).

Prenons un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de la loi p et pour toute valeur (i, j) considérons la variable aléatoire

$$N_{ij} = \sum_{m=1}^{n} \mathbf{1}_{[(X_m, Y_m) = (i,j)]}, \tag{15}$$

qui compte le nombre de fois que la valeur (i, j) apparait dans l'échantillon. La variable aléatoire

$$\bar{p}_n(i, j) = \frac{N_{ij}}{n}, \tag{16}$$

[3]voir le cours de probabilités

est l'estimateur empirique de la probabilité p_{ij} et la loi empirique du n-échantillon est la matrice (aléatoire) $\bar{p}_n$ à k lignes et l colonnes donnée par

$$\bar{p}_n = (\bar{p}_n(i,j)) \,. \tag{17}$$

Comme dans le paragraphe précédent, on peut considérer le χ^2 d'ajustement entre la loi p et la loi empirique $\bar{p}_n$:

$$\chi^2_n(p, \bar{p}_n) = n \sum_{i,j} \frac{(p_{ij} - N_{ij}/n)^2}{p_{ij}}, \tag{18}$$

où la somme précédente porte sur tous les couples (i,j) avec $i = 1, \ldots, k$ et $j = 1, \ldots, l$. Le raisonnement du corollaire 4.2 montre que

$$\chi^2_n(p, \bar{p}_n) \xrightarrow{\mathcal{L}} \chi^2(kl-1) \quad (n \to \infty), \tag{19}$$

et on peut alors pratiquer le test d'ajustement du χ^2 à une loi p comme on l'a expliqué au paragraphe 4.2. Mais si l'on ne veut pas faire d'hypothèse sur la loi p et qu'on veut seulement savoir si la loi p est une loi produit (i.e. si les composantes du couple sont indépendantes), on considére les estimateurs empiriques $\hat{p}_n(i,.)$ (resp $\hat{p}_n(.,j)$) des lois marginales $p_{i.}$ (resp $p_{.j}$) définis par

$$\hat{p}_n(i,.) = \frac{N_{i.}}{n} \quad (\text{resp } \hat{p}_n(.,j) = \frac{N_{.j}}{n}), \tag{20}$$

où

$$N_{i.} = \sum_{m=1}^{n} \mathbf{1}_{[X_m = i]}, \quad (\text{resp } N_{.j} = \sum_{m=1}^{n} \mathbf{1}_{[Y_m = j]})$$

est le nombre de fois que la valeur i (resp. la valeur j) apparait en première coordonnée (resp. en deuxième coordonnée) dans l'échantillon. On définit la loi empirique produit (tensoriel) des deux lois empiriques marginales comme étant la matrice $\hat{p}_n$ donnée par

$$\hat{p}_n = (\hat{p}_n(i,j)) = \left(\frac{N_{i.} N_{.j}}{n^2} \right). \tag{21}$$

Si on remplace dans (18) p par $\hat{p}_n$, on peut espérer mesurer l'indépendance des composantes X_m et Y_m de l'échantillon. Précisément

Définition 2.6 *: On appelle χ^2 d'indépendance la variable aléatoire*

$$\begin{aligned} \chi^2_n(\hat{p}_n, \bar{p}_n) &= n \sum_{i,j} \frac{(\hat{p}_n(i,j) - \bar{p}_n(i,j))^2}{\hat{p}_n(i,j)} \\ &= n \sum_{i,j} \frac{(\frac{N_{i.}N_{.j}}{n^2} - \frac{N_{ij}}{n})^2}{\frac{N_{i.}N_{.j}}{n^2}} \end{aligned} \tag{22}$$

On a alors le résultat suivant que nous admettrons

Théorème 2.7 *:*

$$\chi^2_n(\hat{p}_n, \bar{p}_n) \xrightarrow{\mathcal{L}} \chi^2\left((k-1)(l-1)\right) \quad (n \to \infty). \tag{23}$$

Le test du χ^2 d'indépendance : On déduit du résultat précédant que pour tester l'hypothèse

$$H_0 : \text{"}X \text{ et } Y \text{ sont des variables aléatoires indépendantes"}$$

contre l'hypothèse alternative H_1 :"X et Y ne sont pas indépendantes", on calcule à partir de l'échantillon la quantité $\chi^2_n(\hat{p}_n, \bar{p}_n)$ définie en (22) et au risque α, on rejette H_0 au profit de H_1 si

$$\chi^2_n(\hat{p}_n, \bar{p}_n) > \chi^2_{(k-1)(l-1),\alpha},$$

où $\chi^2_{(k-1)(l-1),\alpha}$ est la borne d'ordre α de la queue de la loi $\chi^2((k-1)(l-1))$. Ceci suppose encore que la valeur de n est suffisamment grande car sous l'hypothèse H_0, la loi de $\chi^2_n(\hat{p}_n, \bar{p}_n)$ n'est qu'approximativement égale à $\chi^2((k-1)(l-1))$.

Remarque : Sous l'hypothèse H_0 et si les lois marginales $p_{i.}$ et $p_{.j}$ étaient connues, on a vu en (18) que la variable aléatoire

$$n \sum_{i,j} \frac{(p_{i.}p_{.j} - N_{ij}/n)^2}{p_{i.}p_{.j}},$$

suivrait approximativement la loi $\chi^2(kl-1)$. Mais on estime les $p_{i.}$ (resp. les $p_{.j}$) qui sont au nombre de $k-1$ (resp. $l-1$). On constate que la règle suivante est satisfaite : on doit diminuer le nombre de degrés de liberté a priori (i.e. $kl-1$) par le nombre total de paramètres estimés (soit $(k-1)+(l-1)$). En effet, on a bien

$$kl - 1 - (k-1) - (l-1) = (k-1)(l-1).$$

Cette règle s'applique dans d'autres situations qui ne sont pas au programme de ce cours.

Exemple : Un échantillon de 1000 personnes ont été interrogées sur leur opinion à propos d'une question qui sera posée à un référendum. On a demandé à ces personnes de préciser leur appartenance politique. Les résultats sont donnés par le tableau suivant[4] :

Appartenance	Réponse		
	Favorable	Défavorable	Indécis
Gauche	210	194	91
Droite	292	151	62

On veut savoir la réponse au référendum est indépendante de l'opinion politique. Pour cela associons les indices de ligne $i = 1$ et 2 à gauche et droite respectivementet les indices de colonne $j = 1, 2, 3$ aux réponses favorable, défavorable et indécis respectivement. On calcule alors les valeurs $\frac{N_{ij}}{n}$ (ici $n = 1000.$) qu'on dispose dans un tableau ainsi que les valeurs $\frac{N_{i.}N_{.j}}{n^2}$ (dans le même tableau entre parenthèses), ce qui donne

i	j		
	1	2	3
1	0,21(0,248)	0,194(0,170)	0,091(0,076)
2	0,292(0,254)	0,151(0,174)	0,062(0,077)

[4]appelé souvent tableau de contingence dans la littérature.

La quantité $\chi^2_n(\hat{p}_n, \bar{p}_n) = n \sum_{i,j} \frac{(\frac{N_{i.}N_{.j}}{n^2} - \frac{N_{ij}}{n})^2}{\frac{N_{i.}N_{.j}}{n^2}}$ est alors égale à :

$$1000\Big(\frac{(0,248-0,210)^2}{0,248} + \frac{(0,170-0,194)^2}{0,170} + \frac{(0,076-0,091)^2}{0,076} + \frac{(0,254-0,292)^2}{0,254} + \frac{(0,174-0,151)^2}{0,174} + \frac{(0,077-0,062)^2}{0,077}\Big) = 23,82.$$

Dans ce cas on a $k-1=1$ et $l-1=2$ i.e. on utilise la loi $\chi^2(2)$ pour laquelle le seuil de rejet au risque $\alpha = 0,05$ est égal à 5,99. On doit donc rejeter l'hypothèse H_0 d'indépendance de la réponse et de l'opinion politique. On constate qu'on rejette aussi H_0 au risque $0,01$.

3 Le test de Kolmogorov-Smirnov

On examine maintenant l'ajustement d'un échantillon observé $x_1, \ldots, x_n$ à une loi continue F. La méthode est complètement différente de celle étudiée dans le premier paragraphe. Elle est basée sur la construction d'une fonction étagée appelée **fonction de répartition empirique** de l'échantillon observé.

3.1 Loi empirique d'un n-échantillon

Définition 3.1 *: Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi F sur $\mathbb{R}^d$. On appelle loi empirique de cet échantillon la mesure aléatoire*

$$dF_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{24}$$

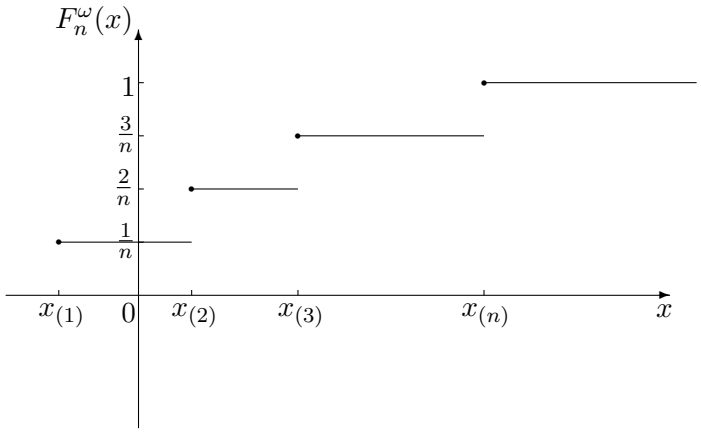
qui pour $\omega \in \Omega$ est la mesure[5]

$$dF_n^\omega = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}, \tag{25}$$

qu'on appelle mesure empirique du n-échantillon observé $(X_1(\omega), \ldots, X_1(\omega))$.

Dans le cas $d = 1$, la fonction de répartition F_n de la mesure dF_n est appelée **fonction de répartition empirique** du n-échantillon $(X_1, \ldots, X_n)$. C'est une fonction aléatoire qui pour $\omega \in \Omega$ prend pour valeur F_n^ω, la fonction de répartion de la mesure empirique observée (25). Si $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ sont les valeurs du n-échantillon observé rangées par ordre croissant et supposées distinctes deux à deux, les sauts de la fonction F_n^ω sont tous égaux à $\frac{1}{n}$ (voir la figure ci-dessous)

[5] δ_a désigne la mesure de Dirac au point $a \in \mathbb{R}^d$.



Lorsque les valeurs $x_{(i)}$ ne sont pas toutes distinctes, par exemple s'il y a k fois la valeur $x_{(i)}$ dans l'échantillon, le saut de F_n^ω au point $x_{(i)}$ est égal à $\frac{k}{n}$.
L'intérêt de la loi empirique F_n est d'approcher la loi F lorsque la taille n de l'échantillon est suffisamment grande.

3.2 Le théorème fondamental de la Statistique

On continue l'étude précédente. Soit F_n la fonction de répartition empirique d'un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires réelles de fonction de répartition F. On veut des précisions sur l'écart uniforme

$$D_n(\omega) = \sup_{x\in\mathbb{R}} |F_n^\omega(x) - F(x)|, \tag{26}$$

entre F_n et F. Bien entendu, D_n est une variable aléatoire puisqu'elle dépend de chaque réalisation du n-échantillon. A priori la loi de D_n semble dépendre de F mais, (et c'est particulièrement remarquable), ce n'est pas le cas si on se restreint à des lois F continues.

Théorème 3.2 *: Si F est continue, la loi de probabilité de la variable aléatoire D_n définie en (26) est une loi intrensèque i.e. elle ne dépend pas de F.*

démonstration : Posons $Y_i = F(X_i)$ $(i = 1, \ldots, n)$. D'après un résultat du chapitre 1, $(Y_1, \ldots, Y_n)$ est un n-échantillon de la loi uniforme sur $[0, 1]$ dont nous noterons U_n^ω la fonction de répartition empirique :

$$U_n^\omega(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{F(X_i(\omega))}(]-\infty, t]) \tag{27}$$

Mais d'après le 2) de la Proposition 2.4, $\delta_{F(X_i(\omega))}(]-\infty, t]) = \delta_{X_i(\omega)}(]-\infty, F^{-1}t])$ où F^{-1} est l'inverse généralisée de F. Il résulte alors de (27) que $U_n^\omega(t) = F_n^\omega(F^{-1}(t))$. Ainsi, en notant $U(t)$ la fonction de répartition de la loi uniforme, on a

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |U_n^\omega(t) - U(t)| &= \sup_{t\in]0,1[} |U_n^\omega(t) - t| \\
&= \sup_{t\in]0,1[} |F_n^\omega(F^{-1}(t)) - F(F^{-1}(t))| \\
&= \sup_{t\in\mathbb{R}} |F_n^\omega(t) - F(t)| = D_n(\omega),
\end{aligned}$$

ce qui prouve que $D_n(\omega)$ a la même valeur que pour la loi uniforme. Q.E.D. $\square$

Remarque : La loi de probabilité de D_n a été tabulée pour différentes valeurs de l'entier n et on connait son comportement asymptotique pour "n grand" (voir le paragraphe suivant). Voyons maintenant le résultat fondamental obtenu par Glivenko et Cantelli :

Théorème 3.3 *(Théorème fondamental de la Statistique) : Pour $\mathbb{P}$-presque tout ω, on a*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n^\omega(x) - F(x)| = 0,$$

i.e. $D_n(\omega) \to 0, \quad \mathbb{P} - p.s.$ Autrement dit, la fonction de répartition empirique converge uniformément vers F $\mathbb{P}$-presque surement.

Remarque : La fonction de répartition empirique est donc un **estimateur** de la fonction de répartition. Le résultat du théorème justifie **le principe des méthodes statistiques** : Ce principe dit qu'on peut déterminer la loi de probabilité (inconnue) d'une variable aléatoire à partir d'un échantillon (assez grand) de tirages effectués suivant cette loi.

4 Estimation d'une fonction de répartition et test de Kolmogorov-Smirnov

4.1 Estimation d'une fonction de répartition

On a vu, au paragraphe précédent que la fonction de répartion empirique F_n^ω d'un n-échantillon issu d'une loi F inconnue, fournit une très bonne approximation de la fonction de répartition F. On va voir comment on utilise ce résultat dans la pratique.
On travaille avec une valeur de n fixée. La loi de la variable aléatoire D_n définie en (26) est tabulée. On peut la trouver dans les recueils de tables statistiques. Ainsi pour tout $0 < \alpha < 1$, on peut déterminer la valeur[6] d_n^α telle que

$$\mathbb{P}(D_n \leq d_n^\alpha) = 1 - \alpha, \tag{28}$$

c'est à dire $\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq d_n^\alpha) = 1 - \alpha$, ce qui revient encore à dire que

$$\mathbb{P}\big(\forall x \in \mathbb{R}, \quad F_n(x) - d_n^\alpha \leq F(x) \leq F_n(x) + d_n^\alpha\big) = 1 - \alpha. \tag{29}$$

L'encadrement

$$F_n(x) - d_n^\alpha \leq F(x) \leq F_n(x) + d_n^\alpha \tag{30}$$

est donc vrai pour tout $x \in \mathbb{R}$, avec la probabilité $1 - \alpha$. La quantité $1 - \alpha$ est appelée **le niveau de confiance** de l'encadrement (30) et α est appelée le **risque d'erreur de première espèce**[7] car c'est la probabilité que (30) ne soit pas vraie.

Définition 4.1 *: Pour $0 < \alpha < 1$, l'ensemble aléatoire*

$$\big\{(x, y) \in \mathbb{R}^2; \quad F_n(x) - d_n^\alpha \leq y \leq F_n(x) + d_n^\alpha\big\} \tag{31}$$

[6] en fait les tables donnent les valeurs de d_n^α pour différentes valeurs de α et non la loi de D_n.
[7] il y a un risque de deuxième espèce dont nous parlerons plus tard.

est appelée la **bande de confiance** *de la fonction de répartition F au niveau de confiance $1-\alpha$.*
Estimer une fonction de répartition c'est par définition la donnée d'un niveau de confiance et de la bande de confiance correspondante qu'on obtient de la façon décrite ci-dessus.

Pour chaque réalisation $(X_1(\omega),\dots,X_n(\omega))$ du n-échantillon, la fonction de répartition empirique prend la valeur F_n^ω et **la bande de confiance observée** est égale à

$$\left\{(x,y)\in\mathbb{R}^2;\quad F_n^\omega(x)-d_n^\alpha\le y\le F_n^\omega(x)+d_n^\alpha\right\} \tag{32}$$

Remarque : La bande de confiance est évidemment sensible à la valeur de n choisie au départ. En effet, à un niveau de confiance donné $1-\alpha$, la largeur de la bande est déterminée par la valeur d_n^α qui décroit lorsque n augmente comme le montrent les tables. Ainsi si l'on veut de la précision sur l'estimation de F il faudra que la bande de confiance soit étroite donc que n soit grand. Pour avoir une bande de largeur donnée, on déterminera la taille n que devra avoir l'échantillon à partir des tables de la loi de D_n.

4.2 Le test de Kolmogorov-Smirnov

En statistiques toute méthode qui permet d'estimer un objet (paramètre, loi, etc...) est susceptible d'être présentée sous une forme équivalente qu'on appelle **un test**. Le test de Kolmogorov-Smirnov est basé sur les observations du paragraphe précédent :
On s'intéresse à une loi de probabilité inconnue F dont on pense qu'elle est égale à une certaine loi F_0. On veut au niveau de confiance $1-\alpha$,

tester l'hypothèse H_0 : "$F=F_0$" contre l'hypothèse alternative H_1 : "$F\neq F_0$".

Mise en oeuvre du test : A partir d'un n-échantillon observé de la loi F, on construit la bande de confiance au niveau de confiance $1-\alpha$. Si cette bande ne contient pas entièrement le graphe de F_0, on *rejette* l'hypothèse H_0. Dans le cas où la bande contient entièrement le graphe de F_0, on ne rejette pas H_0 au niveau de confiance $1-\alpha$. Ceci ne revient pas tout à fait à dire qu'on accepte H_0 mais nous n'entrerons pas dans les détails. La théorie des tests statistiques présente des subtilités que nous présenterons par la suite. En particulier nous ignorons ici un aspect très important du test qui est l'erreur de 2ième espèce i.e. la probabilité d'accepter l'hypothèse H_0 quand elle est fausse.

Remarque 1(et exercice) : En fait $D_n(\omega)=\sup_{x\in\mathbb{R}}|F_n^\omega(x)-F(x)|,=\max_{1\le i\le n}\Delta_i$, avec $\Delta_i=\max\{|F_n^\omega(x_{(i)}^-)-F(x_{(i)})|,\ |F_n^\omega(x_{(i)})-F(x_{(i)})|\}$, où $F_n^\omega(x_{(i)}^-)$ est la limite à gauche de la fonction F_n^ω au point $x=x_{(i)}$. On n'a donc pas besoin de la bande de confiance pour faire le test de Kolmogorov-Smirnov : il suffit de calculer la quantité $\max_{1\le i\le n}\Delta_i$ et de tester si elle dépasse la borne d_n^α pour rejeter H_0.

Remarque 2 : Le problème que nous venons de traiter concerne l'ajustement d'une distribution inconnue à une distribution théorique. Il existe un autre test beaucoup plus utilisé : le test du χ^2 vu au paragraphe 1 qui traite de la même question pour les lois discrètes et qui est également utilisé pour les lois continues mais il faut alors utiliser un procédé de discrétisation (assez arbitraire) de la loi continue. Ce test a le défaut d'être approximatif et tout à fait inadapté lorsqu'on utilise des petits échantillons. Par contre la méthode de Kolmogorov-Smirnov est exacte et elle est particulièrement bien adaptée aux petits échantillons. Il conviendrait donc de mieux faire connaitre ce test aux utilisateurs qui font des

études de produits sur des échantillons de petite ou moyenne taille (médecins, laboratoires industriels, etc...).

4.3 Annexe

Démonstration du théorème fondamental de la Statistique

On aura besoin du lemme suivant

Lemme 4.2 *: Pour tout $x \in \mathbb{R}$ fixé, $\lim_{n\to\infty} F_n^\omega(x) = F(x)$ (resp[8]. $\lim_{n\to\infty} F_n^\omega(x^-) = F(x^-)$) pour $\mathbb{P}$-presque tout ω.*

démonstration du lemme : Pour $\mathbb{P}$-presque tout ω, on a

$$\begin{aligned} F_n^\omega(x) &= \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}(]-\infty, x]) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{]-\infty,x]}(X_i(\omega)) \\ &= \frac{1}{n}\big(\sum_{i=1}^n \mathbf{1}_{]-\infty,x]}(X_i)\big)(\omega) \to \mathbb{E}\left(\mathbf{1}_{]-\infty,x]}(X_1)\right) = F(x), \quad (n\to\infty) \end{aligned}$$

d'après la loi forte des grands nombres appliquée à la suite des variables aléatoires i.i.d. $\mathbf{1}_{]-\infty,x]}(X_i)$. De même on montre que $F_n^\omega(x^-) \to F(x^-)$ $(n\to\infty)$ en remplaçant l'intervalle $]-\infty, x]$ par $]-\infty, x[$ dans le calcul précédent. □

démonstration du théorème : Soit $N \geq 1$ un entier. Pour $j \in \{0, 1, \ldots, N\}$ posons

$$x_{j,N} = \inf\{x; F(x) \geq \frac{j}{N}\} \quad \text{si } j \geq 1 \text{ et } x_{0,N} = -\infty.$$

Pour $x \in [x_{j,N}, x_{j+1,N}[$, d'après la croissance de F et F_n^ω, on a

$$F_n^\omega(x_{j,N}) - F(x_{j+1,N}^-) \leq F_n^\omega(x) - F(x) \leq F_n^\omega(x_{j+1,N}^-) - F(x_{j,N}). \tag{33}$$

Mais $F(x_{j,N}) \geq \frac{j}{N} \geq F(x_{j,N}^-)$, donc $F(x_{j,N}) + \frac{1}{N} \geq F(x_{j+1,N}^-)$ et les inégalités (33) impliquent qu'on a

$$F_n^\omega(x_{j,N}) - F(x_{j,N}) - \frac{1}{N} \leq F_n^\omega(x) - F(x) \leq F_n^\omega(x_{j+1,N}^-) - F(x_{j+1,N}^-) + \frac{1}{N}. \tag{34}$$

D'où

$$\sup_{x\in[x_{j,N},x_{j+1,N}[} |F_n^\omega(x) - F(x)| \leq \frac{1}{N} + |F_n^\omega(x_{j,N}) - F(x_{j,N})| + |F_n^\omega(x_{j+1,N}^-) - F(x_{j+1,N}^-)| \tag{35}$$

pour $j = 1, \ldots, N-1$ mais cette inégalité est vraie aussi pour $j = 0$ (resp. $j = N$) avec la convention $x_{0,N} = -\infty$ (resp. $x_{N+1,N} = +\infty$). En passant au Max en j dans les inégalités (35), on obtient

$$\begin{aligned} D_n(\omega) &= \sup_{x\in\mathbb{R}} |F_n^\omega(x) - F(x)| \\ &\leq \frac{1}{N} + \max_{0\leq j\leq N} \left\{|F_n^\omega(x_{j,N}) - F(x_{j,N})| + |F_n^\omega(x_{j+1,N}^-) - F(x_{j+1,N}^-)|\right\} \end{aligned} \tag{36}$$

[8]Si G est une fonction de répartition, $G(x^-)$ désigne la limite à gauche de G au point x.

En passant à la limsup quand $n \to \infty$ dans les deux membres de (36), on déduit immédiatement du Lemme 4.3 que pour $\mathbb{P}$-presque tout ω, on a :

$$\limsup_{n\to\infty} D_n(\omega) \leq \frac{1}{N}.$$

Or N est arbitraire donc $\limsup_{n\to\infty} D_n(\omega) = 0$. D'où le résultat du théorème. □