ECOLE DES MINES DE NANCY

Décision et Prévision Statistiques

Thierry Verdel (2007)

Chapitre 6

Faits et modèles

Nous avons envisagé certaines lois de probabilité susceptibles de constituer des modèles pour les populations de références. Il s'agit maintenant, en présence d'observations, de choisir le modèle adapté et de vérifier que les observations disponibles s'y raccordent bien.

1. Distributions statistiques

1.1. Mise en ordre des observations

Ayant effectué des observations sur les n individus constituant un échantillon, la mise en ordre consiste à grouper ensemble les résultats identiques, c'est-à-dire à faire correspondre, aux valeurs observées de la variable prise en considération, les nombres d'individus ayant présenté ces valeurs. Le tableau obtenu définit ce qu'on appelle une *distribution statistique***.**

Dans le cas d'une variable susceptible de prendre les valeurs discrètes x_1, ..., x_k, ..., x_r, les résultats se présentent sous la forme du tableau ci-dessous.

Valeurs	Effectifs
x_1	n_1
⋮	⋮
x_k	n_k
⋮	⋮
x_r	n_r

Lorsque la variable est continue, il est commode de procéder à des groupages en classes. Cela consiste à diviser l'intervalle de variation de la variable en classes :

$$[x_1 - \tfrac{h}{2}, x_1 + \tfrac{h}{2}[, [x_1 + \tfrac{h}{2}, x_2 + \tfrac{h}{2}[, \ \ldots, [x_{r-1} - \tfrac{h}{2}, x_r + \tfrac{h}{2}[$$

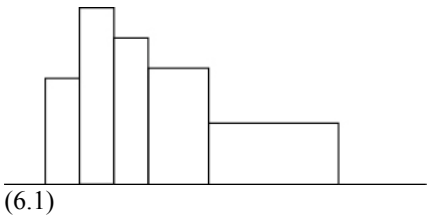
puis à grouper ensemble les valeurs observées qui tombent dans une même classe. Il est évident qu'une telle opération fait perdre de l'information ; on peut montrer toutefois que la perte d'information est négligeable si l'on choisit l'intervalle de classe de façon à obtenir 10 à 15 classes.

1.2. Représentations graphiques des distributions

Nous allons décrire les trois représentations graphiques les plus utilisées, dans le cas d'une variable continue. La transposition au cas d'une variable discrète ne pose pas de problèmes.

1.2.1. Histogramme

Ayant gradué l'axe des abcisses suivant les intervalles retenus, on construit sur chaque intervalle un rectangle de *surface* proportionnelle à la fréquence (absolue ou relative) des observations qui lui correspondent.

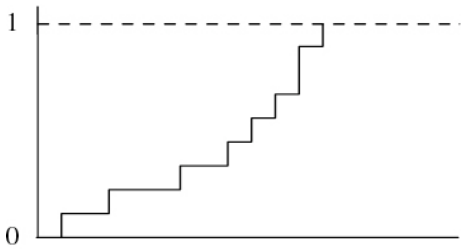


(6.1)

La surface est préférée à la hauteur pour éliminer, quand les classes sont inégales, l'influence de cette inégalité.

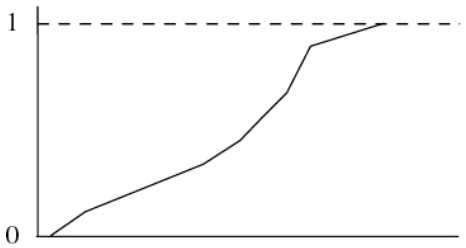
1.2.2. Diagramme des fréquences cumulées

C'est un graphe en escalier qui fait correspondre à chaque observation x, en abcisse, la fréquence, en ordonnée, des observations inférieures ou égales à x. On envisage le plus souvent les fréquences relatives, si bien que les fréquences cumulées sont comprises entre 0 et 1.



(6.2)

Une autre façon, plus simple, de le construire est de définir des classes, de faire correspondre à chaque frontière supérieure la fréquence cumulée correspondante et de joindre les points par des segments de droite. Cela revient à admettre une distribution uniforme à l'intérieur de chacune des classes

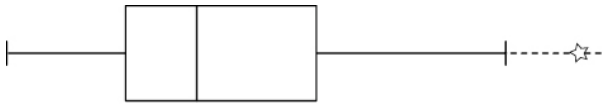


(6.3)

1.2.3. Boîte à moustaches

Cette représentation a été imaginée récemment, dans le même courant d'idées que celui des statistiques *robustes*. Elle consiste en une « boîte » dont les côtés verticaux correspondent aux *quartiles* de la distribution et qui est traversée par un segment correspondant à la *médiane*.

De part et d'autre de la boîte, on définit deux « moustaches » de longueur égale à 1,5 fois l'*étendue interquartile*. Si une observation dépasse une des moustaches, elle est considérée comme " aberrante " et individualisée. S'il n'y a pas d'observation aberrante, la moustache s'arrête à l' observation immédiatement supérieure (ou inférieure). Pour certains auteurs, les moustaches s'étendent jusqu'aux valeurs extrêmes même si celles-ci dépassent l'intervalle défini plus haut.


(6.4)

1.3. Caractéristiques des distributions

Rappelons que les caractéristiques essentielles sont :

- pour la tendance centrale, la moyenne m,

- pour la dispersion, la variance s^2.

Appelant f_k la fréquence relative $\frac{n_k}{n}$ pour la valeur x_k (ou pour la classe de centre x_k), on peut écrire que $m = \sum f_k\, x_k$ et $s^2 = \sum f_k (x_k - m)^2$.

S'il n'y a pas eu regroupement en classes, la fréquence de chaque valeur x_i est $\frac{1}{n}$ et on retrouve les formules habituelles $m = \frac{1}{n} \sum x_i$ et $s^2 = \frac{1}{n} \sum (x_i - m)^2$.

Il convient de noter la parenté évidente entre concepts statistiques et concepts probabilistes :

- à la notion de fréquence f_k, pour une distribution statistique, correspond celle de probabilité p_k, pour une loi de probabilité.

- à la notion de moyenne d'une distribution $m = \sum f_k\, x_k$ correspond la notion d'espérance mathématique $E(X) = \sum p_k\, x_k$.

- enfin, à la notion de variance d'une distribution $s^2 = \sum f_k (x_k - m)^2$ correspond celle de variance d'une variable aléatoire $\sigma^2 = \sum p_k (x_k - \mu)^2$.

2. Fréquences et probabilités

2.1. Retour sur la loi des grands nombres

Considérons un ensemble de possibilités E et deux évènements *complémentaires* A et $\overline{A}$. A est, par exemple, l'évènement : la variable X prend sa valeur dans un certain intervalle $[x,\ x'[$. Soit ϖ la probabilité de A ; celle de $\overline{A}$ est donc égale à $(1 - \varpi)$.

Faisons successivement n épreuves (expériences) identiques, et supposons qu'au cours de ces n épreuves, A se produise k fois et $\overline{A}$, par conséquent, $(n - k)$ fois. La fréquence relative de A est $f_n = \frac{k}{n}$. L'expérience montre que, si n est assez grand, f_n est voisin de ϖ. C'est la fameuse loi des grands nombres due à Bernouilli en 1713, et qui jette un pont entre fréquences et probabilités

Nous avons démontré, au chapitre 4 en partant de l'inégalité de Bienaymé-Tchebichef, qu'étant donnée une suite de variables aléatoires indépendantes et suivant la même loi de probabilité, leur moyenne convergeait *en probabilité* vers la moyenne de la loi. Appliquée à la moyenne de n variables de Bernouilli $X_1,\ \ldots,\ X_n$, cela va permettre de démontrer ce qui précède.

En effet, soit $F_n = \frac{X_1 + \ldots + X_n}{n}$ cette moyenne. Elle a pour espérance ϖ et pour variance $\frac{\varpi(1-\varpi)}{n}$ et l'inégalité de Bienaymé-Tchebichef pemet d'écrire que :

$$\text{Prob}\{\ |\ F_n - \varpi\ |\ > \varepsilon\} \leq \frac{\varpi(1-\varpi)}{n\,\varepsilon^2}$$

Ce résultat s'énonce ainsi : ε étant un nombre positif arbitraire, aussi petit que l'on veut, la probabilité pour que la fréquence relative F_n s'écarte de la probabilité ϖ d'une quantité supérieure à ε, tend vers 0 lorsque le nombre d'épreuves augmente indéfiniment.

D'un point de vue pratique, cela exprime qu'en faisant un nombre suffisant d'épreuves, il est possible d'avoir une « idée » aussi précise qu'on le veut de la probabilité ϖ qu'on ne connaît pas.

Il suffit, par exemple, de faire un nombre assez grand de tirages dans une urne de composition inconnue (proportion ϖ de boules noires) pour que la fréquence observée des boules noires soit *presque sûrement* très voisine de la proportion ϖ. La loi des grands nombres constitue, à ce titre, la base de la statistique mathématique.

2.2. Nombre de mesures à effectuer pour une précision donnée

L'inégalité de Bienaymé-Tchebichef majore beaucoup la probabilité cherchée. Il en résulte que la valeur de n qu'il faut dépasser pour que cette probabilité n'excède pas un seuil fixé, est inutilement grande. Souhaitant, par exemple, estimer une proportion (ou une probabilité) ϖ voisine de 0.2 avec une précision égale à ± 0.01 et un risque de 5%, l'application de l'inégalité de Bienaymé-Tchebichef conduit à un nombre n très grand puisque :

$$0.05 = \text{Prob}\{\mid F_n - \varpi \mid > 0.01\} \leq \frac{0.2 \times 0.8}{n(0.01)^2} \Longrightarrow n \geq 32000.$$

Il est préférable d'utiliser le théorème central limite qui établit que, si n est suffisamment grand, la fréquence relative obéit à une loi qui s'approche d'une loi normale de moyenne ϖ et de variance $\frac{\varpi(1-\varpi)}{n}$. D'où il résulte que la variable aléatoire

$U_n = \frac{F_n - \varpi}{\sqrt{\frac{\varpi(1-\varpi)}{n}}}$ suit approximativement une loi normale réduite.

Il s'ensuit que :

$$\text{Prob}\{\mid F_n - \varpi \mid > \varepsilon\} \simeq 2\,\text{Prob}\left\{U > \frac{\varepsilon}{\sqrt{\frac{\varpi(1-\varpi)}{n}}}\right\}$$

qui conduit, avec les mêmes données que ci-dessus, à :

$$0.05 \simeq 2\,\text{Prob}\left\{U > \frac{0.01}{\sqrt{\frac{0.2 \times 0.8}{n}}}\right\}$$

et, après lecture dans la table de la loi normale réduite, à $n \geq 984$.

On peut constater que c'est généralement l'ordre de grandeur de la taille des échantillons constitués pour les sondages d'opinion.

2.3. Estimation d'une proportion et intervalle de confiance

Puisque $E(F_n) = \varpi$ et que $\sigma^2(F_n) = \frac{\varpi(1-\varpi)}{n} \longrightarrow 0$, F_n est un estimateur sans biais de ϖ.

D'autre part, l'approximation de la loi de F_n par une loi normale permet d'écrire, au risque α près, que :

$\left| f_n - \varpi \right| \leq u_{\alpha/2} \sqrt{\frac{\varpi(1-\varpi)}{n}}$, où $u_{\alpha/2}$ est lu dans une table de la loi normale réduite.

Si n est suffisamment grand, on peut approximer le deuxième membre de l'inégalité en remplaçant ϖ par son estimation f_n . D'où l'intervalle de confiance, au risque α :

$$f_n - u_{\alpha/2} \sqrt{\frac{f_n(1-f_n)}{n}} < \varpi < f_n + u_{\alpha/2} \sqrt{\frac{f_n(1-f_n)}{n}} \ .$$

2.4. Comparaison de deux proportions

Soit ϖ_1 et ϖ_2 les proportions caractérisant deux populations, et soit f_1 et f_2 les fréquences observées sur deux échantillons, de tailles respectives n_1 et n_2, prélevés au hasard dans chacune de ces populations. Faisant l'hypothèse que $\varpi_1 = \varpi_2 = \varpi$, une démarche calquée sur celle mise en oeuvre pour la comparaison de deux moyennes, permet d'établir que, en estimant ϖ par la quantité $\varpi^* = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$, le quotient :

$$u = \frac{f_1 - f_2}{\sqrt{\varpi^*(1-\varpi^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

est approximativement une réalisation d'une variable normale réduite, si l'hypothèse est vraie. Il suffit, pour tester l'hypothèse, de placer u par rapport à l'intervalle correspondant au risque choisi.

2.5. Métrique du χ^2

Soit $p(x)$ la densité de probabilité d'une variable aléatoire X. La probabilité pour que X prenne une valeur dans l'intervalle $[x_k - \frac{h}{2}, x_k + \frac{h}{2}[$ est égale à $\varpi_k = \int_{x_k - \frac{h}{2}}^{x_k + \frac{h}{2}} p(x)\, dx$, et la probabilité pour que X prenne une valeur en dehors de cet intervalle est $(1 - \varpi_k)$. Pour un échantillon de n observations de la variable X, le nombre de valeurs n_k tombant dans l'intervalle $[x_k - \frac{h}{2}, x_k + \frac{h}{2}[$ est une réalisation d'une variable N_k qui suit une loi binomiale de moyenne $n\,\varpi_k$ et de variance $n\,\varpi_k(1 - \varpi_k)$, et lorsque n augmente N_k converge en probabilité vers $n\,\varpi_k$.

Pour tenir compte de toutes les classes, soit $n_1, \ldots, n_k, \ldots, n_r$ les *effectifs observés* dans les classes et soit $\varpi_1, \ldots, \varpi_k, \ldots, \varpi_r$, les probabilités théoriques correspondant à la loi de référence. On définit ce qu'on appelle les *effectifs théoriques* dans les classes, qui sont les quantités $n\,\varpi_1, \ldots, n\,\varpi_k, \ldots, n\,\varpi_r$. Notons qu'effectifs observés et théoriques ont même somme n.

Effectifs observés	Effectifs théoriques
n_1	$n\,\varpi_1$
$\vdots$	$\vdots$
n_k	$n\,\varpi_k$
$\vdots$	$\vdots$
n_r	$n\,\varpi_r$
n	n

Pour mesurer la *distance* entre distribution observée des n_k et distribution théorique des $n\,\varpi_k$, on peut envisager plusieurs quantités.

L'une d'elle est la distance de Kolmogorov qui est la quantité :

$$D = \sup\{\,|\, n_k - n\,\varpi_k \,|\}.$$

Mais on retient le plus souvent ce qu'on appelle la distance du χ^2 qui est la quantité :

$$\chi^2 = \sum_{k=1}^{r} \frac{(n_k - n\,\varpi_k)^2}{n\,\varpi_k}.$$

Nous allons montrer que, *si l'échantillon provient bien d'une population définie par la loi de probabilité envisagée*, cette quantité est une réalisation d'une loi du χ^2 à $(r - 1)$ degrés de liberté : nombre de classes moins 1.

La démonstration qui suit n'est pas essentielle. Pour simplifier les notations, nous la ferons dans le cas de trois classes, mais elle est facilement transposable au cas général.

Soit les variables : $X_1 = \frac{N_1 - n\,\varpi_1}{\sqrt{n\,\varpi_1}}$, $X_2 = \frac{N_2 - n\,\varpi_2}{\sqrt{n\,\varpi_2}}$, $X_3 = \frac{N_3 - n\,\varpi_3}{\sqrt{n\,\varpi_3}}$.

La variable dont nous cherchons la loi est $\chi^2 = {X_1}^2 + {X_2}^2 + {X_3}^2$. *Les variables* X_1, X_2, X_3 *suivent à la limite, quand n augmente, des lois normales centrées et de variances respectives :* $(1 - \varpi_1)$, $(1 - \varpi_2)$, $(1 - \varpi_3)$, *mais elles ne sont* ***pas indépendantes*** *puisque* : $N_1 + N_2 + N_3 = n$ *et, par conséquent* : $\sqrt{\varpi_1}\,X_1 + \sqrt{\varpi_2}\,X_2 + \sqrt{\varpi_3}\,X_3 = 0$.

Calculons les covariances $E(X_1\,X_2)$, $E(X_2\,X_3)$, $E(X_3\,X_1)$. *Pour cela, multiplions l'égalité précédente par* X_1, *puis* X_2, *puis* X_3. *On obtient* :

$$\sqrt{\varpi_1}\,E({X_1}^2) + \sqrt{\varpi_2}\,E(X_1\,X_2) + \sqrt{\varpi_3}\,E(X_1\,X_3) = 0$$
$$\sqrt{\varpi_1}\,E(X_1\,X_2) + \sqrt{\varpi_2}\,E({X_2}^2) + \sqrt{\varpi_3}\,E(X_2\,X_3) = 0$$
$$\sqrt{\varpi_1}\,E(X_1\,X_3) + \sqrt{\varpi_2}\,E(X_2\,X_3) + \sqrt{\varpi_3}\,E({X_3}^2) = 0$$

Tenant compte de ce que : $E({X_1}^2) = (1 - \varpi_1)$, $E({X_2}^2) = (1 - \varpi_2)$, $E({X_3}^2) = (1 - \varpi_3)$, *et de ce que* : $\varpi_1 + \varpi_2 + \varpi_3 = 1$, *on trouve, après résolution du système linéaire* :

$$E(X_1\,X_2) = -\sqrt{\varpi_1\,\varpi_2}\,,\ E(X_2\,X_3) = -\sqrt{\varpi_2\,\varpi_3}\,,\ E(X_1\,X_3) = -\sqrt{\varpi_1\,\varpi_3}\,.$$

Cela étant, le vecteur $\overrightarrow{OP}(X_1, X_2, X_3)$ *est orthogonal au vecteur unitaire* $\vec{u}\left(\sqrt{\varpi_1}\,,\ \sqrt{\varpi_2}\,,\ \sqrt{\varpi_2}\right)$.

Faisons un changement de coordonnées orthonormales en prenant l'un des axes passant par le vecteur $\vec{u}$. *Les nouvelles coordonnées de* $\overrightarrow{OP}$ *seront* :

$$Y_1 = a_1\,X_1 + a_2\,X_2 + a_3\,X_3$$
$$Y_2 = b_1\,X_1 + b_2\,X_2 + b_3\,X_3$$
$$Y_3 = \sqrt{\varpi_1}\,X_1 + \sqrt{\varpi_2}\,X_2 + \sqrt{\varpi_3}\,X_3$$

Et, dans ce nouveau système, $\chi^2 = {Y_1}^2 + {Y_2}^2$.

Les six valeurs des coefficients a et b sont déterminées d'une infinité de façons par les cinq relations :

$${a_1}^2 + {a_2}^2 + {a_3}^2 = 1$$
$${b_1}^2 + {b_2}^2 + {b_3}^2 = 1$$
$$a_1\,b_1 + a_2\,b_2 + a_3\,b_3 = 0$$
$$\sqrt{\varpi_1}\,a_1 + \sqrt{\varpi_2}\,a_2 + \sqrt{\varpi_3}\,a_3 = 0$$
$$\sqrt{\varpi_1}\,b_1 + \sqrt{\varpi_2}\,b_2 + \sqrt{\varpi_3}\,b_3 = 0$$

Chacune des variables Y suit une loi normale puisque toute fonction linéaire de variables normales suit une loi normale. Calculons leurs variances $E(Y_1{}^2)$, $E(Y_2{}^2)$ *et leur covariance* $E(Y_1\ Y_2)$. *On a, pour une fonction linéaire* :

$$E(Y_1{}^2) = a_1{}^2(1-\varpi_1) + a_2{}^2(1-\varpi_2) + a_3{}^2(1-\varpi_3) - 2\left(a_1\ a_2\ \sqrt{\varpi_1\ \varpi_2} + a_1\ a_2\ \sqrt{\varpi_2\ \varpi_3} + a_1\ a_3\ \sqrt{\varpi_1\ \varpi_3}\right)$$

$$E(Y_1{}^2) = a_1{}^2 + a_2{}^2 + a_3{}^2 - \left(a_1\ \sqrt{\varpi_1} + a_2\ \sqrt{\varpi_2} + a_3\ \sqrt{\varpi_3}\right)^2$$

et finalement :

$$E(Y_1{}^2) = 1.$$

On peut montrer, de la même façon, que $E(Y_2{}^2) = 1$ *et que* $E(Y_1\ Y_2) = 0$. *Par suite* Y_1 *et* Y_2 *sont des variables normales centrées, réduites et indépendantes, la condition* $E(Y_1\ Y_2) = 0$ *étant nécessaire et* ***suffisante*** *pour l'indépendance de variables* ***normales.***

Donc la variable $\chi^2 = Y_1{}^2 + Y_2{}^2$ *est la somme des carrés de deux variables normales, réduites, indépendantes et suit, par conséquent, une loi du* χ^2 *à deux degrés de liberté* : *nombre de classes moins un*.

3. Techniques de raccordement entre distributions statistiques et lois de probabilité

3.1. Loi de référence

Le problème, sous la forme la plus générale, consiste à caractériser à partir des données le type de la loi de référence, puis à préciser cette loi par estimation des paramètres qui la définissent complètement. En pratique, cependant, on n'opère pas exactement ainsi. Les lois de référence s'identifiant le plus souvent aux lois de probabilité fondamentales (loi binomiale, normale, log-normale), il s'avère plus simple :

- de rapprocher la distribution examinée de la loi de probabilité à laquelle il semble intuitivement (ou pour des raisons théoriques) qu'elle doive se raccorder ;

- de vérifier ensuite la validité du rapprochement ainsi opéré.

Lorsque le raccordement à l'une des lois fondamentales s'avère injustifié, il y a lieu de faire appel à d'autres lois de référence, et il en existe un nombre considérable (loi gamma, loi beta, loi de Pareto, loi de Gumbel, loi de Weibull, ...), ou d'en créer éventuellement pour la circonstance.

3.2. Détermination du type de la loi de référence

Il n'y a pas de recette particulière à mettre en jeu pour déterminer le type de la loi de référence à laquelle on soupçonne la distribution observée de se rattacher. En général, on se laisse guider par des considérations logiques ou bien, plus simplement, on tente des rapprochements qui semblent résulter de la forme même des distributions observées.

Dans le cas de distributions relatives à des variables discrètes, le raccordement à des lois de référence binomiale ou de Poisson s'impose de prime abord.

Dans le cas de variables continues, le raccordement à des lois de référence normale ou log-normale s'avère très souvent, mais pas toujours, légitime. En vue de vérifier, avant tout calcul compliqué, que l'hypothèse de tels raccordements n'est pas a priori absurde, on dispose d'ailleurs de moyens simples et rapides.

3.2.1. Raccordement à une loi normale

La loi normale est une loi symétrique. De plus, on a vu que l'intervalle $[\mu \pm u\sigma[$ comprend approximativement la probabilité : 50% pour $u = 2/3$, 68% pour $u = 1$, 95% pour $u = 2$ et presque 100% pour $u = 3$. Si donc une distribution observée en pratique est telle que les fréquences des observations comprises à l'intérieur de ces intervalles sont voisines de ces probabilités, il y a présomption de normalité.

On peut également vérifier cette présomption à l'aide d'une transformation connue sous l'appellation de *droite de Henry*. Soit $P(x)$ le graphe de la fonction de répartition d'une loi normale ; il a une forme en S. Il existe dans le commerce un papier dit « gausso-arithmétique » qui, par un changement d'échelle de l'axe des ordonnées, permet de réaliser une anamorphose de $P(x)$ qui le transforme en une droite. Dès lors, si l'on trace sur ce papier le diagramme des fréquences cumulées de la distribution observée et que ses points sont sensiblement alignés, on peut penser que le raccordement à une loi normale est légitime. Un tel graphique est réalisé par tous les logiciels de calculs statistiques.

3.2.2. Raccordement à une loi log-normale

Pour reconnaitre sommairement si une distribution observée est du type log-normal, il est également commode d'utiliser un graphique de Henry avec échelle des abcisses logarithmique (papier gausso-logarithmique). On procède alors à la détermination du paramètre θ_0 par tâtonnement ou, ce qui est mieux, par remarques d'ordre technique.

3.3. Estimation des paramètres de la loi de référence

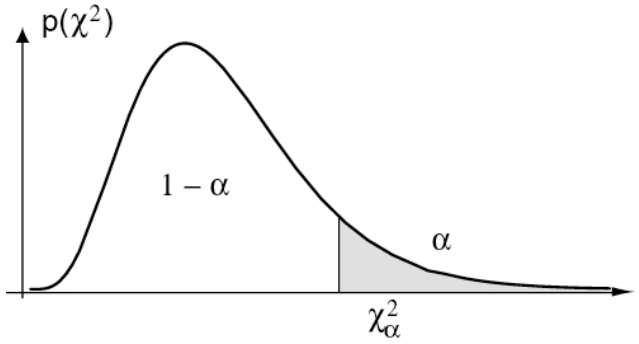
La loi de référence dépend le plus souvent d'un certain nombre de paramètres qu'il est nécessaire d'estimer pour la définir complètement. Une loi binomiale ou de Poisson est entièrement définie par la proportion ϖ à laquelle elle correspond (n étant connu). Une loi normale est entièrement définie par sa moyenne μ et son écart-type σ. Il convient donc, à partir des données disponibles, d'estimer soit la proportion ϖ, soit la moyenne μ et l'écart-type σ de la loi de référence binomiale, de Poisson ou normale, pour ne considérer que ces trois lois là.

3.4. Vérification de la légitimité d'un raccordement effectué

La comparaison des n_k observés et des $n\,\varpi_k$ théoriques met en évidence des différences plus ou moins fortes. Cela n'a rien d'étonnant puisque, dans l'hypothèse où le raccordement opéré est justifié, la distribution des $n\,\varpi_k$ n'est que la loi limite de la distribution des n_k. Il reste toutefois à savoir si les différences ainsi mises en évidence sont compatibles avec les seuls aléas de l' échantillonnage. Ce n'est en effet qu'à cette condition qu'on peut considérer le raccordement opéré comme légitime.

La vérification consiste à déterminer la loi d'une certaine fonction de l'ensemble des fluctuations entre effectifs observés et théoriques, *dans l'hypothèse où ces fluctuations ne sont effectivement dues qu'aux aléas de l'échantillonnage*.

Retenant la fonction $\chi^2 = \sum_{k=1}^{r} \frac{(n_k - n\,\varpi_k)^2}{n\,\varpi_k}$, nous avons montré que, dans ces conditions, elle était une réalisation d'une loi du χ^2. A un seuil de probabilité α faible pouvant être considéré comme négligeable correspond une valeur ${\chi_\alpha}^2$ telle que la probabilité d'observer $\chi^2 > {\chi_\alpha}^2$ soit justement égale à α.



(6.5)

Si la valeur χ^2 observée est supérieure à ${\chi_\alpha}^2$, il parait préférable de mettre en doute l'hypothèse de la légitimité du raccordement. Si, au contraire, χ^2 est inférieur à ${\chi_\alpha}^2$, il n'y a pas de raison de mettre en doute cette hypothèse. Comme on s'y est déjà habitué, cela ne signifie malheureusement pas qu'elle soit vraie. Or ce que l'on souhaiterait généralement c'est confirmer la validité du modèle envisagé. L'aspect négatif du test statistique, dans le sens où il ne prend pas en compte le risque de conserver à tort l'hypothèse faire, est gênant dans ce cas précis.

Notez que l'on a effectué un test à droite, puisque ce sont des écarts importants entre effectifs observés et théoriques que l'on veut éventuellement détecter, donc une valeur grande du χ^2. Ajoutons enfin deux remarques sur la mise en oeuvre du test.

La première est que, pour que la loi de la quantité χ^2 soit suffisamment voisine d'une loi du χ^2, il faut non seulement que n soit assez grand, mais encore que les nombres $n\,\varpi_k$ ne soient pas trop petits : en pratique ils ne doivent pas être inférieurs à 5. Si certains d'entre eux sont trop petits, il est nécessaire de procéder à des *groupages de classes.*

La seconde remarque est que, le plus souvent, la loi de référence dépend d'un ou plusieurs paramètres inconnus. A ce moment là, les $n\,\varpi_k$ sont calculés non pas à partir des paramètres véritables de la loi, mais à partir des paramètres estimés. Ils sont donc eux-mêmes aléatoires. On démontre alors que le nombre de degrés de liberté de la loi du χ^2 à laquelle il faut se référer est égal à $(r - 1 - p)$, où p est le *nombre de paramètres estimés.*

4. Tests non paramétriques

Fort souvent on est amené à prendre en considération des variables dont on ignore la distribution. Il n'est alors plus possible de se référer aux tests de comparaison décrits dans le chapitre 5. Pour lever cette difficulté, on s'est donc préoccupé de définir des tests, dits *non paramétriques*, ne faisant aucune hypothèse sur la nature des populations mises en jeu.

Il existe une très grande variété de tels tests non paramétriques, mais nous nous limiterons à la présentation de ceux qui sont les plus utilisés et qui se trouvent reposer sur la prise en compte d'une même quantité χ^2 que le test ci-dessus du raccordement entre une distribution observée et une distribution théorique.

4.1. Test de comparaison de plusieurs populations qualitatives

Soit p populations $\mathcal{P}_1$, ..., $\mathcal{P}_i$, ..., $\mathcal{P}_p$ dont les individus sont distingués suivant r *catégories* C_1, ..., C_k, ..., C_r qui peuvent être les modalités d'une variable qualitative (ou les classes d'une variable quantitative). Pour deux lots de pièces, par exemple, classées en bonnes ou mauvaises, on a $p = 2$ et $r = 2$.

On a prélevé un échantillon dans chacune de ces populations. Soient n_1, ..., n_j, ..., n_p, leurs tailles et soit n_{jk} le nombre d'individus qui proviennent de la population $\mathcal{P}_j$ et qui appartiennent à la catégorie C_k.

	$\mathcal{P}_1$	...	$\mathcal{P}_j$	...	$\mathcal{P}_p$
C_1	n_{11}	...	n_{j1}	...	n_{p1}
⋮	⋮		⋮		⋮
C_k	n_{1k}	...	n_{jk}	...	n_{pk}
⋮	⋮		⋮		⋮
C_r	n_{1r}	...	n_{jr}	...	n_{pr}
	n_1	...	n_j	...	n_p

Si l'on fait l'hypothèse que les populations sont *identiques*, alors les probabilités d'appartenir à chacune des classes sont les mêmes pour toutes les populations, soit ϖ_1, ..., ϖ_k, ..., ϖ_r, et l'on peut définir des effectifs théoriques dans chaque classe et pour chaque population : $n_j \varpi_k$ pour la classe C_k de la population $\mathcal{P}_j$. Il semble naturel d'estimer la probabilité dans la classe C_k par :

$$\varpi_k^* = \frac{\sum_{j=1}^{p} n_{jk}}{\sum_{j=1}^{p} n_j},$$

et d'envisager la quantité :

$$\chi^2 = \sum_{j=1}^{p} \sum_{k=1}^{r} \frac{(n_{jk} - n_j \varpi_k^*)^2}{n_j \varpi_k^*}$$

pour tester l'hypothèse d'identité des populations. On montre que, sous cette hypothèse, elle obéit à une loi du χ^2 à $(p(r-1)-(r-1)) = (p-1)(r-1)$ degrés de liberté.

4.2. Test de la médiane

Etant donnés les résultats fournis par deux échantillons de taille n_1 et n_2 :

échantillon 1 : $x_1, x_2, \ldots, x_{n_1}$

échantillon 2 : $y_1, y_2, \ldots, y_{n_2}$

Arrangeons l'ensemble de ces résultats selon une même suite croissante : $x_1, y_1, y_2, x_2, y_3, x_3, \ldots$ et désignons la médiane de cette suite par M.

Après décompte des observations au dessus et en dessous de M, le tableau des données peut être résumé ainsi.

Observés	$> M$	$< M$	Total
Echantillon 1	n_{11}	n_{12}	n_1
Echantillon 2	n_{21}	n_{22}	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	$n_1 + n_2$

Dans l'hypothèse où les deux populations sont identiques, la proportion théorique des observations au dessus et en dessous de la médiane est dans tous les cas $1/2$. Au tableau précédent correspond le tableau théorique ci-contre, et on est en définitive ramené à un test du χ^2 avec un degré de liberté.

Théoriques	$> M$	$< M$	Total
Echantillon 1	$\frac{n_1}{2}$	$\frac{n_1}{2}$	n_1
Echantillon 2	$\frac{n_2}{2}$	$\frac{n_2}{2}$	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	$n_1 + n_2$

4.3. Test des signes

Ce test s'applique à des observations appariées. Sur un même individu i on a effectué deux mesures y_i et x_i et on s'intéresse aux différences $d_i = y_i - x_i$. Dans le test classique on prenait en compte les valeurs de ces différences, mais dans le test des signes on ne retiendra que les signes, plus ou moins, de ces différences. Il y a donc perte d'information.

S'il n'y a pas de différence entre les mesures, la probabilité d'un signe plus est égale à celle d'un signe moins et égale à 0.5. S'il y a n individus dans l'échantillon, les effectifs théoriques sont égaux à $0.5\,n$ et on est encore ramené à un test du χ^2 avec un degré de liberté sur la quantité :

$$\frac{(n_+ - 0.5\,n)^2}{0.5\,n} + \frac{(n_- - 0.5\,n)^2}{0.5\,n}.$$

4.4. Test d'indépendance entre deux variables qualitatives

Soit $x_1, \ ..., \ x_i, \ ..., \ x_p$ et soit $y_1, \ ..., \ y_j, \ ..., \ y_q$ les modalités de deux variables qualitatives X et Y. Un échantillon de n individus sur lesquels ont été repérées les valeurs prises simultanément par les deux variables a donné les résultats ci-contre : n_{ij} est le nombre d'individus ayant présenté à la fois la valeur x_i de X et la valeur y_j de Y. $n_{i.}$ et $n_{.j}$ représentent respectivement le total de la ligne x_i et celui de la colonne y_j.

$X \setminus Y$	y_1	...	y_j	...	y_q	Total
x_1	n_{11}	...	$n_{1\,j}$	...	$n_{1\,q}$	n_1
⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i.}$
⋮	⋮		⋮		⋮	⋮
x_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.q}$	n

Soit les probabilités suivantes :

$$\varpi_{i.} = \text{Prob}\{X = x_i\}$$

$$\varpi_{.j} = \text{Prob}\{Y = y_i\}$$

$$\varpi_{ij} = \text{Prob}\{X = x_i \text{ et } Y = y_j\}$$

Faisons l'hypothèse que les deux variables sont indépendantes. Il s'ensuit, d'après le théorème des probabilités composées, que :

$$\varpi_{ij} = \varpi_{i.}\ \varpi_{.j}$$

Estimons $\varpi_{i.}$ et $\varpi_{.j}$ respectivement par $\varpi_{i.}^* = \frac{n_{i.}}{n}$ et $\varpi_{.j}^* = \frac{n_{.j}}{n}$, donc ϖ_{ij} par $\varpi_{ij}^* = \frac{n_{i.}\ n_{.j}}{n^2}$.

Sous l'hypothèse d'indépendance, l'effectif théorique correspondant à l'effectif observé n_{ij} est égal à $n\,\varpi_{ij}^* = \frac{n_{i.}\ n_{.j}}{n}$ et la quantité :

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.}\ n_{.j}}{n}\right)^2}{\frac{n_{i.}\ n_{.j}}{n}}$$

obéit à une loi du χ^2 à $(p - 1)\,(q - 1)$ degrés de liberté.

Exercices du chapitre 6

Exercice 1

Il y a dans un certain pays 25% d'illettrés. On considère les 400 prisonniers d'une prison de ce pays.

a) Calculer, en admettant que l'échantillon constitué par les prisonniers est tiré au hasard dans la population du pays, la probabilité pour que la fréquence des prisonniers illettrés dépasse 35% ?

b) Quelle valeur cette fréquence n'a-t-elle qu'une chance sur 100 de dépasser ?

Exercice 2

Un sondage d'opinion sur les intentions de vote d'un certain électorat a porté sur un échantillon de 2500 électeurs représentatifs de cet électorat et a fourni les résultats suivants : 1200 électeurs favorables au candidat A et 1300 favorables au candidat B.

a) En admettant que la proportion ϖ inconnue de l'électorat favorable au candidat B soit égale à 0.5 (ballotage), donner la fourchette dans laquelle la fréquence relative à un échantillon de 2500 électeurs aurait la probabilité 95% de tomber.

b) Que penser d'un journal qui annoncerait, au vu des résultats du sondage, l'élection du candidat B ?

c) Quelle devrait être la taille de l'échantillon pour qu'une différence de 4% entre les fréquences (48% favorables à A et 52% à B) permette de conclure, avec moins de une chance sur 100 de se tromper, à l'élection du candidat B ?

Exercice 3

Une enquête réalisée auprès de 4000 foyers d'une grande ville en 1990 a montré que 1944 d' entre eux possédaient une machine à laver la vaisselle. Une enquête réalisée dans la même ville en 1995 sur 5000 foyers a montré que 2587 possédaient un tel appareil. Peut-on admettre que le taux d'équipement a augmenté ?

Exercice 4

La direction du marketing d'une entreprise a fait procéder à une enquête auprès d'un échantillon de n consommateurs en leur soumettant deux modes de présentation A et B d'une même marchandise et en leur demandant de faire connaître leur préférence. Soient n_1 et n_2 les nombres de consommateurs ayant préféré respectivement les présentations A et B et soient ϖ_1 et ϖ_2 les proportions correspondantes dans la population. n_1 et n_2 sont des réalisations de variables aléatoires liées par la relation $N_1 + N_2 = n$. On est amené à se demander si la différence constatée $d = n_1 - n_2$ est significative d'une préférence réelle dans la population pour l'un des modes de présentation.

a) d est une réalisation d'une variable D que l'on peut écrire $D = 2N_1 - n$. Calculer sa moyenne et sa variance.

b) On se propose de tester l'hypothèse d'une différence nulle $\varpi_1 - \varpi_2 = 0$, dans la population échantillonnée. Quelle est l'estimation de la variance de D qui vous parait devoir être utilisée ? A quelle condidion doit satisfaire la différence d pour qu'elle puisse être considérée comme significative au niveau de signification 95% ?

Exercice 5

Sur un échantillon de 10000 bébés, 5136 sont des garçons et 4864 sont des filles. Peut-on admettre que les probabilités pour qu'un bébé soit un garçon ou une fille sont égales ?

Exercice 6

On effectue des croisements de poules blanches et noires. D'après les lois de Mendel, lorsqu'on effectue un tel croisement, on a 1 chance sur 4 d'obtenir une poule blanche, 1 chance sur 4 d' obtenir une poule noire et 1 chance sur 2 d'obtenir une poule bleue (hybride). Les 158 croisements effectués ont donné 43 poules blanches, 40 poules noires et 75 poules bleues. Ces résultats vous paraissent-ils compatibles avec les lois de Mendel ?

Exercice 7

On considère ci-dessous la distribution du dernier chiffre de 200 lectures de pesée. Peut-on craindre que celui qui effectuait les pesées a une préférence pour certains chiffres ? Peut-on justifier cette crainte ?

x	0	1	2	3	4	5	6	7	8	9	Total
n (x)	35	16	15	17	17	19	11	16	30	24	200

Exercice 8

La direction du personnel d'une usine veut déterminer si le nombre d'avis d'arrêt pour maladie dépend du jour de la semaine. Pendant la période étudiée, il y a eu 720 avis d'arrêt. Les cas de maladie du samedi après-midi où l'on ne travaille pas ne sont notifiés que le lundi. Chaque jour on notifie les arrêts survenus entre la veille 17 h et le jour 17 h, sauf le samedi où l'on s'arrête à 13 h et le lundi où sont notifiés les arrêts du samedi 13 h au lundi 17 h. Faisant l'hypothèse d'une répartition uniforme des arrêts, que peut-on conclure de cette enquête ?

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Total
Effectifs	189	100	127	115	102	87	720

Exercice 9

La fabrication de pièces dans un atelier de mécanique donne lieu à un certain pourcentage de pièces rebutées comme non utilisables. On a observé 100 lots différents de 100 pièces chacun qui ont donné les résultats suivants.

Rebuts par lot	0	1	2	3	4	5	6	7	8	9	10	11	Total
Nombre de lots	0	2	9	14	20	18	15	9	6	4	2	1	100

Quelle distribution théorique paraît devoir donner une bonne description de la distribution observée et pour quelle raison ? Tester l'accord entre les observations et le modèle.

Exercice 10

On a mesuré les duretés Rockwell sur 100 tôles minces en faisant 3 mesures au centre, dont on a pris la moyenne. Les résultats des mesures, une fois ordonnés et mis en classes, ont été rassemblés dans le tableau suivant.

Classes de dureté	53	54	55	56	57	58	59	60	61	62	63	64	Total
Nb d' observations	1	4	8	13	16	19	14	11	6	5	2	1	100

Que penser du raccordement à une loi normale ? Tracer la droite de Henry.

Exercice 11

On donne dans le tableau suivant la distribution des vitesses de passage de bobines de tôle lors de l'opération de décapage (décalaminage chimique et mécanique). Ces vitesses sont en m/mn.

Classes de vitesse	40	50	60	70	80	90	100	110	120	130	140	150	Total
Effectifs	3	4	19	38	62	72	77	56	46	16	5	2	400

Raccorder cette distribution à une loi normale. On donne $m \simeq 100$ et $s \simeq 20$

Exercice 12

On étudie conjointement la couleur des cheveux et celle des sourcils d'une certaine population. On les classe en deux catégories : clairs (blonds ou roux) et foncés (bruns ou noirs). On trouve les résultats suivants.

Cheveux / Sourcils	Clairs	Foncés	Total
Clairs	30472	3238	33710
Foncés	3364	9468	12832
Total	33836	12706	46542

Y-a-t-il une dépendance entre la couleur des cheveux et celle des sourcils ?

Exercice 13

Dans une usine, on a remplacé la commande manuelle de quelques presses par une commande automatique. On désire voir si cette modification a une influence sur les accidents du travail. On a relevé, pendant une période donnée, le nombre d'ouvriers qui ont eu ou non des accidents et on les a classés suivant qu'ils travaillaient sur des presses à commande manuelle ou à commande automatique. On a obtenu les résultats suivants.

	Manuelle	Automatique	Total
Accidentés	25	23	48
Non accidentés	183	112	295
Total	208	135	343

La modification du type de commande a-t-elle ou non une influence sur le nombre des accidents ?

Exercice 14

En vue d'étudier les effets du tabac sur l'artério-sclérose, on a classé 870 individus selon :

- leur degré d'artério-sclérose (grave, moyen, faible),

- leur consommation de tabac (non fumeurs, légers fumeurs, moyens fumeurs, grands fumeurs de cigarettes, fumeurs de pipe et cigares).

Analyser le tableau suivant qui a été obtenu.

Sclérose	Non fumeurs	Légers	Moyens	Grands	Pipes et cigares
Grave	16	29	76	50	7
Moyen	97	96	180	122	42
Faible	48	27	32	27	21

Exercice 15

Les conclusions d'une étude de la Gendarmerie sur 823 accidents graves (ayant provoqué la mort ou des blessures pour lesquelles une hospitalisation de plus de 8 jours a été nécessaire) étaient les suivantes : " Le risque d'être tué pour un passager placé à l'avant est le même que celui encouru par le conducteur, mais ce risque est presque deux fois plus élevé que pour un passager placé à l'arrière. On peut donc parler de " place du mort " pour la place occupée par le passager avant droit, mais uniquement si l'on compare le passager avant à l'ensemble des passagers, conducteur exclu. Du point de vue de la gravité des accidents, le passager avant est celui qui est le plus exposé. "

Justifier ces conclusions à partir du tableau des résultats.

Place	Tués	Blessés graves	Blessés légers	Indemnes	Total
Conducteur	46	153	95	38	332
Place avant	34	150	60	20	264
Place arrière	16	81	96	34	227
Total	96	384	251	92	823