Journées nationales APMEP
Caen 21 - 24 octobre 2005

Atelier ADa14
Michel HENRY

Ajustement de données expérimentales à une loi de probabilités

II - Ajustement de données expérimentales à une loi continue[1]

Principe des tests d'ajustement

On cherche à préciser le comportement d'un caractère *C* défini sur une population statistique *P*. Pour cela, on construit un modèle probabiliste dont l'univers Ω représente les choix aléatoires des éléments de *P*, sur lequel est défini une variable aléatoire X représentant les valeurs prises par *C* sur *P*. La loi P_X de X modélise la distribution des fréquences des valeurs de *C* sur *P*. On cherche donc à connaître cette loi avec assez de précision. Le problème est de proposer pour P_X une loi suffisamment adéquate pour représenter les variations attendues de *C* sur *P* et de contrôler cette adéquation dans le cadre d'une approximation acceptable.

La technique des tests d'hypothèse en statistique inférentielle donne une réponse adaptée aux situations d'échantillonnage. Elle consiste à tester si une loi théorique de probabilité peut représenter au mieux la distribution des fréquences des valeurs prises par le caractère *C* dans un échantillon *E* d'éléments prélevés au hasard dans *P* (ajustement de ces données expérimentales à la loi théorique). Si la taille de l'échantillon est assez grande, les théorèmes limite du calcul des probabilités (lois des grands nombres, théorème-limite central, théorème du Khi-deux…) permettent d'apprécier la qualité de cet ajustement pour toute la population.

Dans la pratique, les valeurs possibles d'un caractère *C* peuvent être regroupées dans des modalités qui s'imposent naturellement, soit que *C* soit un caractère qualitatif, soit qu'il soit quantitatif discret. Dans d'autre cas, ces valeurs peuvent être considérées comme appartenant à un ensemble numérique continu (unidimentionel), bien qu'on ne puisse en observer qu'un nombre fini et que la précision accessible des mesures les rendent nécessairement discrètes. C'est par exemple le cas de temps aléatoires d'attente d'un phénomène, ou de mesures physiques aléatoirement réparties autour d'une moyenne. Il existe divers tests d'ajustement à une loi continue. On va s'intéresser aux tests dits du Khi-deux.

Tests d'ajustement du Khi-deux

Le principe d'un test d'ajustement du Khi-deux consiste à répartir les valeurs possibles de *C* en un certain nombre k de modalités M_i. L'étude de la répartition de ces valeurs se limite alors à leur distribution entre les M_i dont le choix reste à la charge du statisticien. Il va de soi que les choix de k et des M_i peuvent être plus ou moins judicieux et peuvent induire des tests produisant des conclusions opposées. Des théorèmes donnent des valeurs optimales pour k en fonction de la taille de l'échantillon considéré. Remarquons que les fluctuations de *C* au sein de chacune des modalités M_i ne sont pas prises en compte dans une telle démarche qui peut donc s'avérer assez grossière. Dans la pratique, on considère la v. a. X représentative du caractère *C*, prenant ses valeurs dans un intervalle réel. Une partition de cet intervalle en sous-intervalles I_i détermine alors les modalités M_i.

En général, la nature du phénomène étudié suggère pour la loi inconnue P_X de X, un type de loi. Notons θ le ou les paramètres qui déterminent entièrement cette loi : λ pour une loi exponentielle, (μ, σ) pour une loi normale… Ces paramètres peuvent être connus a priori ou estimés à partir de l'échantillon considéré, par exemple par la moyenne et l'écart-type des valeurs observées de X dans

[1] Cet atelier faisait suite à l'atelier ASm12 : Ajustement de données expérimentales à une loi équirépartie. On trouvera dans son compte rendu une petite introduction aux tests d'hypothèses éclairant ces propos, ainsi que les éléments de bibliographie valables pour les deux ateliers. Une présentation plus étoffée de la problématique des tests d'hypothèses est donnée dans la brochure APMEP n° 156 : *Statistique au lycée*, vol. 1, p. 247-260.

l'échantillon. Désignons par P_θ cette loi théorique, ainsi choisie pour représenter les variations de *C* dans la population *P*. Le problème est donc de vérifier si la loi P_θ modélise bien ces variations, c'est-à-dire si elle est suffisamment « proche » de la loi inconnue P_X.

Dans un test du Khi-deux, on se contente seulement de vérifier que la distribution des fréquences observées $(f_i)_{1\le i\le k}$ (notée (f)), des différentes modalités M_i du caractère *C* dans l'échantillon *E* n'est pas trop éloignée de la loi discrète finie de probabilité $(p_i)_{1\le i\le k}$ (notée (p)), où par définition $p_i = P_\theta(X \in I_i)$ pour tout i = 1… k. Les f_i sont considérées comme les réalisations sur *E* des variables F_i définies sur l'ensemble des échantillons de taille n que l'on peut extraire de la population *P*.

Un test d'ajustement du Khi-deux, consiste donc à vérifier si l'écart observé entre la distribution de fréquences (f) des modalités M_i dans l'échantillon observé et la loi discrète (p) induite par P_θ est seulement dû au hasard du prélèvement de cet échantillon ou s'il s'explique par l'inadéquation de P_θ pour décrire les variations de *C*.

Mise en œuvre d'un test d'ajustement du Khi-deux

Une loi modèle P_θ de la v. a. X étant donnée ainsi qu'une partition $(I_i)_{i=1\ldots k}$ de l'ensemble continu des valeurs prises par X, on considère la loi discrète (p) définie par les probabilités $p_i = P_\theta(X \in I_i)$ et on cherche à estimer la proximité de (p) avec la distribution (f).

Pour un test d'ajustement du Khi-deux, de même que pour un test d'adéquation à une loi équirépartie, on considère la pseudo-distance dans $\mathbb{R}^k$, d_n donnée par :

$$d_n^2((f),(p)) = n\sum_{i=1}^{k}\frac{(f_i - p_i)^2}{p_i} = \left(\sum_{i=1}^{k}\frac{n_i^2}{np_i}\right) - n$$

où $n_i = nf_i$ est l'effectif de la modalité M_i dans l'échantillon.

Le test d'ajustement consiste alors à regarder si la « distance d_n » est inférieure ou supérieure à une certaine *valeur critique* d_c définie par la condition que, *sous l'hypothèse* H_0 que P_θ modélise bien les variations du caractère *C* dans la population *P,* la variable aléatoire $D_n = d_n((F),(p))$, fonction des variables F_i, ne devrait dépasser d_c qu'avec une probabilité inférieure ou égale à un *seuil de signification* α. On voit que le calcul de cette probabilité de contrôle nécessite la connaissance, au moins approximativement, de la loi de D_n sous cette hypothèse H_0. C'est le cas, asymptotiquement, de la distance des tests du Khi-deux. Cette propriété fait l'objet d'un Théorème du Khi-deux adapté à ce contexte.

Théorème du Khi-deux

Soit P_θ *la loi d'une variable aléatoire* X *représentant les variations d'un caractère C défini sur une population statistique P et soit* $(I_i)_{i=1\ldots k}$ *une partition du domaine des valeurs de* X *correspondant aux modalités* M_i *de C.*

Si, dans un échantillon aléatoire de taille n, F_i *désigne la fréquence de la modalité* M_i*, et si* $p_i = P_\theta(X \in I_i)$ *est la probabilité qu'un élément de P pris au hasard soit de modalité* M_i*, la suite des variables aléatoires* $D_n^2 = n\sum_{i=1}^{k}\frac{(F_i - p_i)^2}{p_i}$ *converge en loi vers la loi du* χ^2_ν *à* ν *degrés de liberté, où*

- $\nu = k-1$ *quand* θ *est connu,*
- $\nu = k-r-1$ *quand on a estimé* r *paramètres à partir de l'échantillon.*

On admet en général que pour des n tels que pour tout i de 1 à k, $np_i(1-p_i) \ge 5$, sous les hypothèses de ce théorème, les probabilités $P(D_n^2 > Q)$ et $P(\chi^2_\nu > Q)$ sont assez proches, pour tout Q > 0. La valeur critique d_c est alors donnée par la relation $P(\chi^2_\nu > d_c^2) = \alpha$.

Si ces conditions ne sont pas vérifiées, il est nécessaire de regrouper des intervalles I_i contigus pour pouvoir appliquer ce théorème.

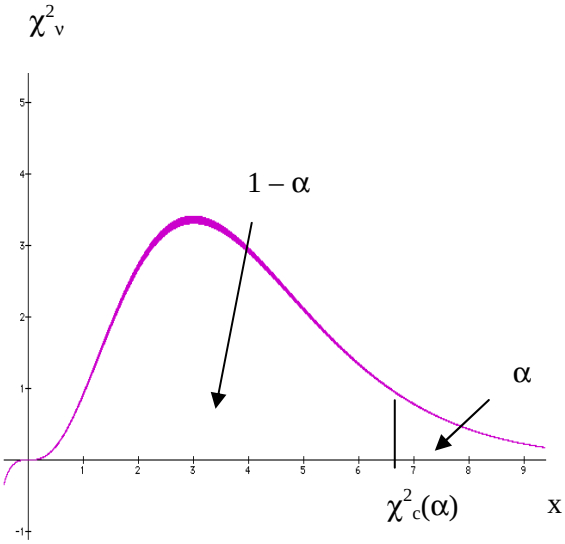
La loi de χ^2_ν est bien connue des statisticiens.

Densité : $f_{\chi^2_\nu}(x) = \dfrac{x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\,\Gamma(\frac{\nu}{2})}\, e^{-\frac{x}{2}}$

On a $E(\chi^2_\nu) = \nu$ et $Var(\chi^2_\nu) = 2\nu$.

Quantile $\chi^2_c(\alpha)$ d'ordre α :

$$P(\chi^2_\nu > \chi^2_c(\alpha)) = \alpha$$



Courbe de densité de la loi du Khi-deux et point critique pour un seuil α donné.

Pratique d'un test d'ajustement du Khi-deux

Décrivons maintenant les différentes étapes à réaliser pour ajuster des données expérimentales à une loi continue de probabilité par la méthode du Khi-deux.

On suppose que ces données sont les valeurs prises par un caractère *C* continu sur un échantillon aléatoire de taille n prélevé dans une population *P*.

- Mise en place d'un modèle probabiliste

1) Considérer une variable aléatoire X à valeurs dans un intervalle réel I, modélisant les variations de *C* dans la population *P*. Une interprétation probabiliste du caractère *C* permet de conjecturer le type de la loi inconnue P_X de X.

2) Expliciter le paramètre θ (éventuellement multidimentionel) déterminant la loi P_X. Les valeurs numériques fixant θ et déterminant une loi P_θ sont :

- soit obtenues par r estimations à partir de l'échantillon donné,
- soit choisies a priori (r = 0).

3) Partager l'intervalle I en k intervalles I_i :

- soit d'amplitudes égales (sauf éventuellement pour les extrémités) et calculer les probabilités $p_i = P_\theta(X \in I_i)$,
- soit de probabilités $p_i = 1/k$ égales et déterminer les quantiles a_i successifs ($P_\theta(X < a_1) = 1/k$, $P_\theta(a_1 < X < a_2) = 1/k$, …) de la loi P_θ délimitant les intervalles I_i.

4) Vérifier que pour tous les i de 1 à k, on a $np_i(1-p_i) \geq 5$ (éventuellement regrouper des intervalles I_i contigus).

Les p_i constituent alors la loi discrète finie (p), sous-modèle probabiliste pour représenter la distribution de fréquences des valeurs de *C* dans les différentes modalités M_i définies dans la population *P* par les intervalles I_i.

- Hypothèses à tester et mise en œuvre du test

5) H_0 : « La loi P_θ est un modèle satisfaisant et la distance entre (p) et la distribution (f) des fréquences des modalités M_i au sein de l'échantillon observé, n'est dû qu'aux fluctuations d'échantillonnage ».

H_1 : « Le hasard du prélèvement de l'échantillon ne peut expliquer à lui seul l'écart observé entre (p) et (f) ».

6) Un seuil α étant donné, trouver dans la table du χ^2_ν à $k-r-1$ degrés de liberté la valeur critique $\chi^2_c(\alpha)$ telle que $P(\chi^2_\nu > \chi^2_c(\alpha)) = \alpha$. La fonction KHIDEUX.INVERSE d'Excel donne aussi ces valeurs.

7) Expliciter la distribution (f) des fréquences f_i des modalités M_i dans l'échantillon observé. On note $n_i = nf_i$ les effectifs des M_i dans l'échantillon.

8) Calculer la *« distance »* du Khi-deux : $d_n^2 = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = (\sum_{i=1}^{k} \frac{n_i^2}{np_i}) - n.$

9) Si $d_n^2 > \chi^2_c(\alpha)$, conclure que la loi P_θ n'est pas un modèle suffisamment adéquat pour la loi P_X représentant les variations de *C* dans *P*.
Rejeter H_0 en prenant un risque de se tromper $P_{H0}(D_n^2 > \chi^2_c(\alpha))$ proche de α.

10) Si $d_n^2 \leq \chi^2_c(\alpha)$, décider de conserver l'hypothèse H_0 : admettre que la loi P_θ est un modèle suffisamment adéquat pour représenter les variations de *C* dans *P*.

Risques et conclusions d'un test

Dans un test d'ajustement du Khi-deux, l'événement peu probable sur lequel se fonde la décision (acceptation ou rejet de H_0) est que la valeur observée de D_n^2 sur l'échantillon considéré dépasse une valeur critique $\chi^2_c(\alpha)$. Dans l'hypothèse H_0, cet événement est de probabilité $P_{H0}(D_n^2 > \chi^2_c(\alpha))$ très voisine de $P(\chi^2_\nu > \chi^2_c(\alpha)) = \alpha$.

$P_{H0}(D_n^2 > \chi^2_c(\alpha))$ est donc le risque que l'on prend que cet événement se réalise effectivement dans l'échantillon observé alors que la loi P_θ est adéquate. C'est le risque de rejeter à tort H_0 appelé *risque de première espèce*. La valeur $1-\alpha$ est appelée *niveau de confiance du test*.

On prend aussi un risque en acceptant H_0. C'est le risque que D_n^2 soit encore inférieur à la valeur critique $\chi^2_c(\alpha)$ quand H_1 est vraie. Ce risque dépend donc de la loi P_X de la variable X qui réalise H_1, il est égal à $P_X(D_n^2 \leq \chi^2_c(\alpha))$. C'est le *risque de deuxième espèce.*

La valeur $1 - P_X(D_n^2 \leq \chi^2_c(\alpha)) = P_{H1}(D_n^2 > \chi^2_c(\alpha))$ est appelée *puissance du test.*

Le risque de première espèce est contrôlé a priori par la donnée du seuil α. La loi P_θ étant connue, la valeur critique $\chi^2_c(\alpha)$ qui en découle est bien calculable. Par contre le risque de deuxième espèce dépend de la loi inconnue P_X du phénomène qui régit les variations du caractère *C*. Sa détermination doit faire l'objet d'une hypothèse alternative testant une autre loi contre P_θ. Ce risque peut être déterminé par simulation, il peut être élevé.

On peut voir que les risques de première et seconde espèce varient en sens contraire. Il est donc erroné de chercher à minimiser le seuil α, il y a nécessairement un compromis à trouver. Pour améliorer la performance d'un test (c'est à dire sa puissance, un seuil α étant donné), il faut pouvoir augmenter la taille n de l'échantillon, paramètre qui n'est pas souvent à la disposition du statisticien.

Dans la pratique, on préfère limiter le risque de rejeter à tort H_0, hypothèse dans laquelle on a obtenu le modèle cherché. On prend souvent $\alpha = 0{,}05$, ce qui veut dire que quand H_0 doit être acceptée, 5% des échantillons conduiront néanmoins à la rejeter. Cela ne veut surtout pas dire que dans 95% des échantillons, on acceptera H_0 en ayant raison, car si accepter ou rejeter H_0 semblent être des événement contraires, leurs probabilités considérées ne sont pas calculées avec les mêmes lois.

Dans la mesure où la conclusion d'un test d'ajustement dépend de l'observation d'un échantillon aléatoire prélevé dans *P*, celle-ci ne peut être de nature déterministe (péremptoire) comme par exemple : « le caractère *C* suit la loi P_θ ». De plus, il ne faut pas oublier que la loi P_θ est une loi théorique dans un modèle probabiliste postulé pour représenter les variations de *C* dans la population. Il faut bien se garder de confondre modèle et réalité. Les outils de la statistique inférentielle ne donnent que des probabilités pour que des décisions relatives à la population puissent être prises de manière optimale. On ne peut jamais être certain d'avoir raison. C'est pourquoi la notion de risque est essentielle et doit figurer dans toute conclusion d'un test d'ajustement.

On dira par exemple :

« En prenant un risque de me tromper inférieur au seuil de signification α, je choisis de rejeter l'hypothèse H_0 pour conclure que la loi P_θ n'est pas adéquate pour représenter les variations du caractère *C* dans la population *P* »,

ou

« Avec un risque (indéterminé) de me tromper, le test du Khi-deux ne me conduit pas à rejeter l'hypothèse H_0 de l'adéquation de la loi P_θ pour représenter les variations du caractère *C* dans la population *P* ».

Le non rejet de H_0 ne permet cependant pas de conclure que la loi P_θ est adéquate sous réserve du risque de deuxième espèce, car le résultat du test dépend de la manière dont on a effectué la partition de l'intervalle I en sous intervalles I_i.

Simulation d'un test d'ajustement : génération de la loi d'une variable continue[2].

La deuxième partie de l'atelier consistait à simuler la réalisation de deux tests d'ajustement : un ajustement par une loi exponentielle et un ajustement par une loi normale. Dans ce but, il convient de produire des échantillons de valeurs prises par des variables respectivement exponentielles ou normales. Pour ce faire, on a la propriété suivante :

Soit f *la densité d'une loi de probabilité, que l'on suppose continue et strictement positive sur un intervalle* I =]a , b[, *et soit* F *la fonction de répartition de cette loi. Alors, si cette loi est la loi d'une variable aléatoire* X, *la variable* U = F(X) *suit une loi uniforme sur*]0 , 1[.

F est continue strictement croissante sur I, sa restriction à I est donc inversible, notons (abusivement) cette fonction réciproque F^{-1}, elle-même strictement croissante sur]0 , 1[à valeurs dans I. Soit u un réel et F_U la fonction de répartition de la loi de U. Par définition, $F_U(u) = P(U \leq u)$.

F_U vérifie :

- si $u \leq 0$, $F_U(u) = P(F(X) \leq u) = 0$, car F est positive et $P(F(X) = 0) = P(X \leq a) = 0$,
- si $0 < u < 1$, $F_U(u) = P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$,
- si $1 \leq u$, $F_U(u) = 1$ car les valeurs prises par U étant des probabilités, « $U \leq 1$ » est l'événement certain.

F_U est donc la fonction de répartition de la loi uniforme sur]0 , 1[.

On en déduit que si U est uniforme sur]0 , 1[, alors la variable X à valeurs dans I, définie par $X = F^{-1}(U)$, suit la loi dont la fonction de répartition est F. Cette propriété permet, si F^{-1} est calculable, de produire par simulation des échantillons des valeurs d'une variable aléatoire continue de loi quelconque à densité non nulle sur un intervalle I.

1) Simulation d'une loi exponentielle

La loi exponentielle de paramètre λ a pour densité sur IR^+ la fonction $\lambda e^{-\lambda x}$, sa fonction de répartition F est nulle si $x \leq 0$ et $F(x) = 1 - e^{-\lambda x}$ pour $x > 0$. Donc $F^{-1}(u) = -\dfrac{\ln(1-u)}{\lambda}$ sur]0 , 1[. Sur Excel, les valeurs u sont obtenues par la fonction ALEA() et le calcul d'un échantillon de valeurs d'une variable exponentielle X est obtenu par la formule : `x=-(LN(1-ALEA()))/λ`.

2) Simulation d'une loi normale

Si Y suit une loi normale de paramètres (μ , σ), les valeurs de Y se déduisent des valeurs d'une variable normale centrée réduite X par la transformation $Y = \mu + \sigma X$. Mais la loi de X ne peut être simulée par l'utilisation directe de F^{-1}, car la fonction de répartition de la loi normale centrée réduite n'est pas explicitable. Mais sur Excel, la fonction LOI.NORMALE.STANDARD.INVERSE donne ces valeurs.

[2] Pour simplifier les énoncés, on se limite au cas des lois à densité. On trouvera un énoncé plus général page 41 dans *Lois continues, tests d'adéquation, une approche pour non spécialistes*, Groupe Probabilités & statistique de l'IREM de Besançon, Presses universitaires de Franche-Comté, 2005, et une étude complète pages 178 à 180 dans *Statistique au Lycée*, Commission Inter-IREM Statistique et probabilités, brochure APMEP n° 156, octobre 2005.